T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	スポーツ技能獲得における段階的訓練手法に関する研究
Title(English)	A Progressive Training Method for Sports Skill Acquisition
著者(和文)	廖振傑
Author(English)	Chen-Chieh Liao
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第387号, 授与年月日:2025年3月26日, 学位の種別:課程博士, 審査員:小池 英樹,篠田 浩一,三宅 美博,金﨑 朝子,井上 中順
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第387号, Conferred date:2025/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Powered by T2R2 (Science Tokyo Research Repository)

Doctoral Dissertation

A Progressive Training Method for Sports Skill Acquisition

Chen-Chieh Liao

Graduate Major in Computer Science
School of Computing
Institute of Science Tokyo

Supervisor: Hideki Koike

March, 2025

Acknowledgment

I would like to thank my advisor Professor Hideki Koike, my previous deputy supervisor Doctor Dong-Hyun Hwang, and all the students and staff in the Koike Lab for their guidance, inspiration, and feedback. I also give my thanks to my friends and especially my family for their help, encouragement, and support throughout all my Doctor's life here in TokyoTech.

Abstract

Beginners in sports often face significant challenges in improving their skills due to limited prior knowledge and insufficient access to professional coaching. This difficulty is particularly pronounced in complex activities like golf, where mastering the correct swing involves intricate motions and precise timing. Traditional learning methods, such as imitating professional athletes through videos, lack personalized guidance and fail to account for individual differences in physique and skill level. To address these challenges and advance the democratization of skill acquisition, this thesis proposes innovative solutions that leverage deep learning and motion analysis to facilitate effective self-training for beginners.

The central problem tackled is enabling users to intuitively understand the differences between their own motions and those of professional players, providing personalized, actionable guidance for progressive skill improvement. We approached this problem through two key contributions:

Firstly, we developed a method for personalized learning target selection and motion style transformation. By analyzing a comprehensive motion dataset, the system identifies intermediate-level motions that serve as attainable stepping stones toward advanced proficiency. Utilizing a motion style transfer network, we extract skill features from motion data and represent them as styles within a latent space. This allows us to transform professional motions into personalized versions that match the user's appearance and physical characteristics. By visualizing an idealized version of themselves performing attainable movements, users can more effectively imagine the desired state, making skill acquisition more accessible and intuitive.

Secondly, we introduced methods for fine-grained motion comparison and incremental guidance. We developed techniques to synchronize motions with different phases and timings, allowing for accurate alignment and comparison. A motion discrepancy detector then identifies subtle differences in movement patterns, offering clear feedback on specific aspects requiring attention. By interpolating between motions in the latent space, the system generates intermediate steps that guide users incrementally from their current skill level to higher levels. This personalized, step-by-step guidance demystifies the learning process, making skill improvement more achievable for beginners.

Building upon these contributions, we presented two prototype systems: Coach Navi and AI Coach. Coach Navi focuses on navigating intermediate-level motions and utilizes a motion navigator, motion style transformer, and motion visualizer to provide personalized training experiences. AI Coach emphasizes motion discrepancy detection, incorporating modules for motion synchronization, discrepancy detection, and visualization. Both systems were evaluated through comprehensive experiments and user studies.

Our approaches advance the concept of democratizing skill acquisition by providing tools that are both effective and accessible to users without specialized equipment or extensive prior knowledge. By emphasizing personalization and incremental learning, we address key barriers that often hinder beginners in sports skill development. The systems developed enable users to engage in self-training tailored to their individual needs, promoting motivation and enhancing the overall learning experience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"I never teach my pupils; I only attempt to provide the conditions in which they can learn."

— *Albert Einstein*

In sports, it is difficult for beginners to improve their skills without prior knowledge. Therefore, as a conventional method, people attend lessons to meet experts and learn how to play with proper form. However, in most sports, to achieve outstanding results on the field and maintain exceptional physical condition, players spend much time training on their own (Figure 1.1). In such situations, it is important to design and implement both an effective and accurate self-training process.

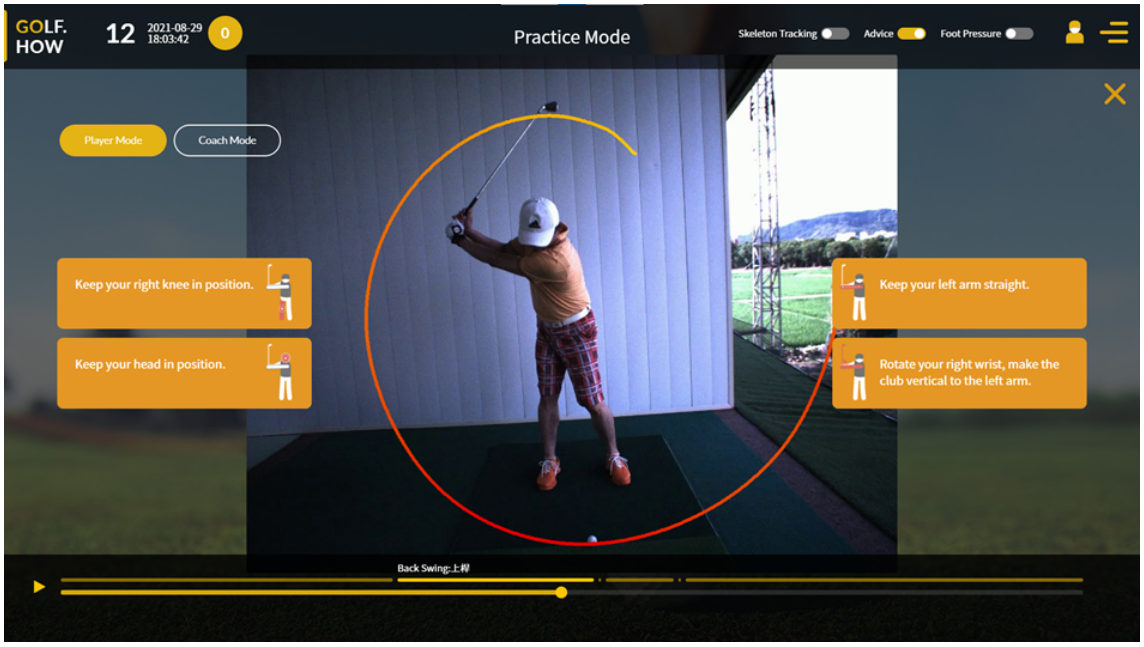

Figure 1.1: Golf self-training analysis system. [4]

A common way to enhance sports skills is to imitate the motions of professional players. People watch the movements of professional athletes through television or the Internet and try to make their bodies move similarly to the professionals. To help accelerate this

process, many systems have recently been developed to assist users in understanding the motions of professionals (Figure 1.2).

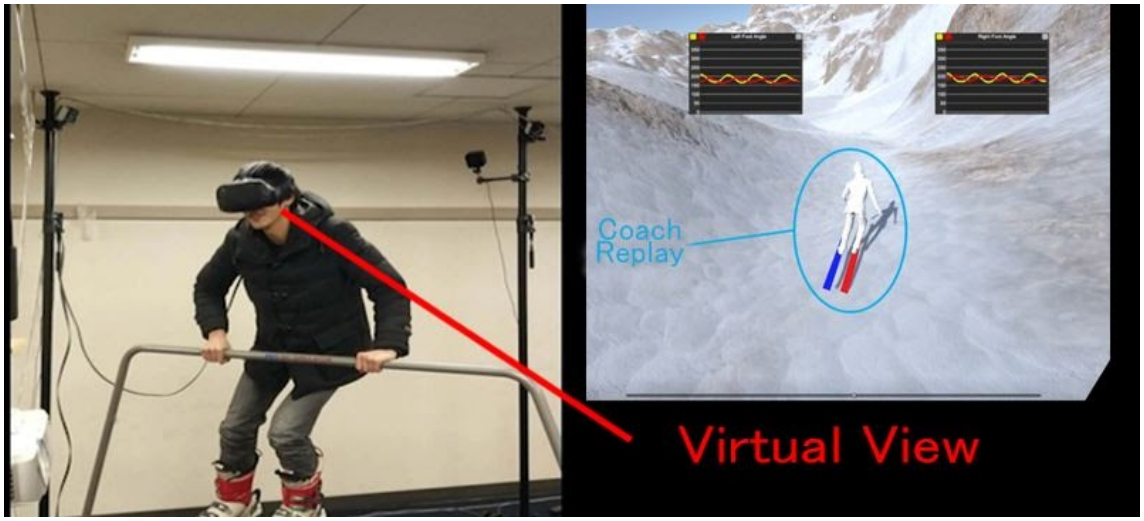

Figure 1.2: VR ski training system where users follow behind a coach. [5]

Deep learning, a subfield of machine learning inspired by the structure of the human brain filled with neurons and synapses, has made significant strides in recent years. Computers can now efficiently simulate brain-like structures with large numbers of nodes and connections. With deep learning technologies, computers can learn to recognize different objects, make predictions, or even anticipate future events (Figure 1.3).

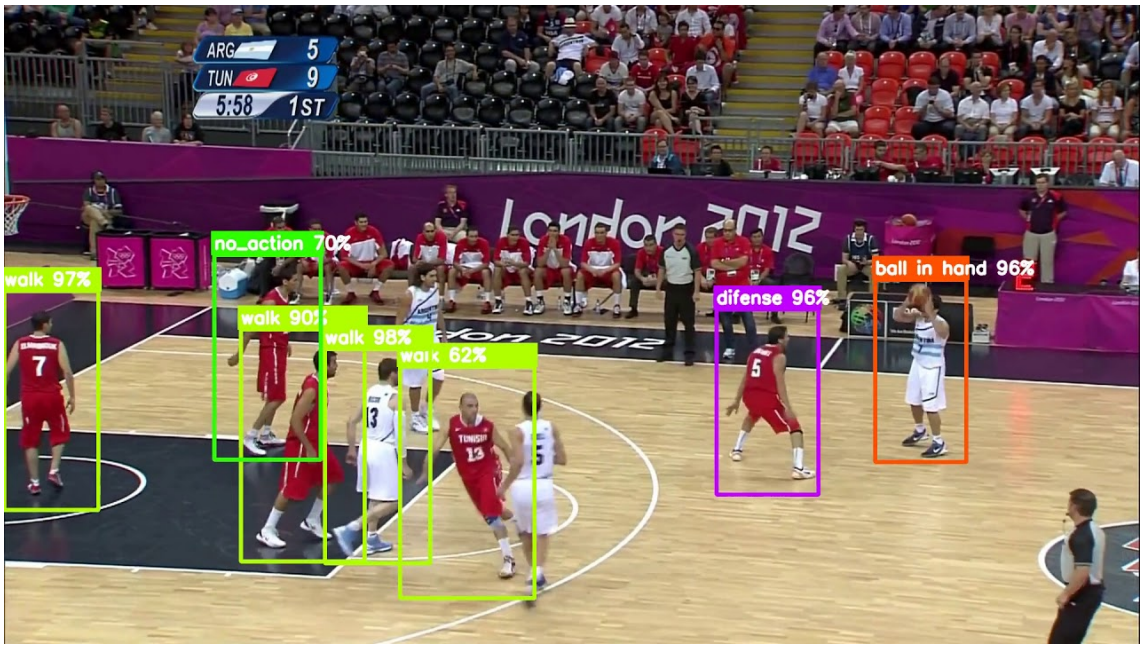

Figure 1.3: Action recognition in a basketball game. [48]

1.1 Advancements in Analyzing Sports Performance

Before the advance of video technology, coaches and athletes relied predominantly on direct observation and experiential knowledge to analyze and improve athletic performance. Coaching methods were grounded in traditions where techniques and strategies were passed down through verbal instruction and physical demonstrations [85]. Athletes depended on their coaches' expertise to identify flaws in technique, while coaches used descriptive language and physical guidance to correct and refine movements. This process was inherently subjective, relying heavily on the coach's personal experience and the athlete's ability to interpret feedback.

To supplement observational coaching, early methods incorporated basic tools such as stopwatches and measuring tapes to quantify aspects of performance like time and distance [16]. For instance, in track and field, coaches timed sprints and measured jumps to assess progress. However, these tools provided limited insight into the biomechanics of complex movements.

The introduction of video technology in the 1960s and 1970s marked a pivotal shift in sports performance analysis [21]. Coaches and athletes began utilizing videotape recordings to capture training sessions and competitions, allowing for replay and slow-motion examination of athletic movements [80]. This innovation provided a more objective and detailed view of performance, enabling the identification of subtle technical errors that were previously difficult to detect. Athletes could see themselves in action, facilitating a better understanding of feedback and fostering self-awareness in technique refinement.

As digital technologies advanced, so did the capabilities of video analysis. Software applications emerged that allowed for frame-by-frame playback, side-by-side comparisons, and overlaying biomechanical models onto video footage [28,51]. These tools enhanced the precision of analysis by enabling measurements of joint angles, velocities, and accelerations directly from the video [119]. Despite these advancements, early video analysis systems were often cumbersome, requiring significant manual input and not readily accessible to all levels of athletes due to cost and technical complexity [89].

The turn of the century saw the rise of more sophisticated digital tools and software that made performance analysis more user-friendly and widely available [18]. The integration of motion capture technology and biomechanical modeling allowed for a deeper understanding of human movement [11, 102]. Inertial measurement units (IMUs) and wearable sensors began to provide real-time data on an athlete's motion, capturing detailed information on body positions, velocities, and accelerations without the need for extensive camera setups [92]. These technologies enabled a more comprehensive analysis of technique, contributing to injury prevention and performance enhancement [116].

1.2 Motion Capture and Computer Graphics Modeling

Motion capture (MoCap) technology (Figure 1.4) has become an indispensable tool in fields such as animation, virtual reality, and sports analysis. Optical motion capture systems, in particular, have gained prominence due to their high accuracy and stability over long-term use compared to inertial (IMU-based) motion capture systems. These systems precisely capture the 3D positions of reflective markers using multiple calibrated cameras, and they can also restore the orientation of rigid bodies or human poses with three or more markers. Furthermore, the fidelity and accuracy of optical motion capture systems can be improved simply by increasing the number of cameras and markers, making them a robust solution for industrial applications such as animation in movies and games, where precision is paramount. However, despite their advantages, optical systems require controlled environments and can be expensive to set up due to the need for multiple cameras and specialized equipment. In contrast, inertial motion capture systems, which use accelerometers and gyroscopes, offer portability but suffer from drift over time and are less accurate for complex movements.

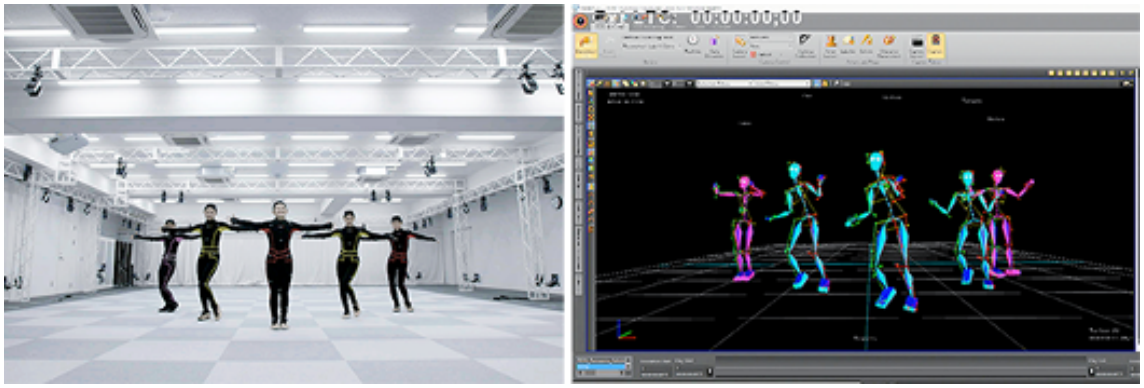

Figure 1.4: Optical motion capture system. [2]

Recently, human motion capture systems without physical markers have become popular because of their convenience and low cost for capturing human motion in any environment. Advances in deep neural networks and hardware accelerators have further revolutionized markerless motion capture. Techniques such as OpenPose and DensePose enable the estimation of 2D human joint positions from a single RGB image without the need for markers. By applying multi-view geometry principles, these 2D detections can be triangulated to obtain 3D joint positions. Recent works have developed neural network-based methods to predict 3D joint positions directly from 2D detections or from raw images and video clips. Consequently, several markerless motion capture systems have been proposed [76, 159], leveraging deep learning to achieve high accuracy in pose estimation even in uncontrolled environments.

After acquiring motion data from a performer, the resulting information can be used to animate content for films, video games, and virtual reality applications (Figure 1.5).

Creating 3D models involves generating polygons and textures in a virtual 3D space. Once the 3D model is completed, the motion data is retargeted to the model to match its geometric attributes (such as bone length) while ensuring natural movement. Traditionally, this process is performed manually by experts who define the desired appearance and adjust the motion to fit the model. Although advances in 3D design software have streamlined certain aspects, manual modeling remains time-consuming and requires significant expertise to achieve realistic and natural results.

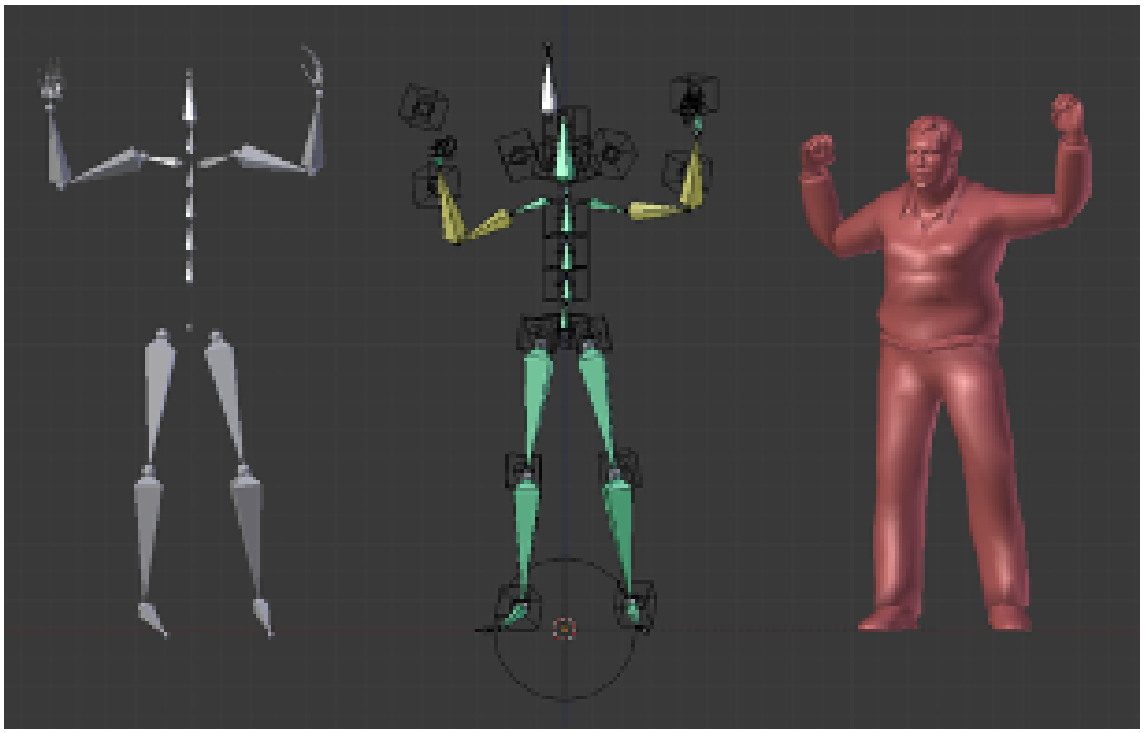

Figure 1.5: 3D character modeling. [1]

An alternative approach involves camera-based 3D scanning methods. This process begins by placing multiple cameras around an object and capturing images from various angles, then applying triangulation to generate a point cloud. The point cloud is further converted into a mesh. While this method accelerates the creation of realistic models from real-world objects, it requires complex hardware setups and substantial computational resources, making it costly and less accessible.

In recent years, research has shown promising results in overcoming the limitations of traditional methods by integrating deep learning into 3D human reconstruction. Some works estimate the human body shape, clothing, and textures from monocular images using neural networks. For example, Habermann et al. [53, 54] proposed methods to capture clothed human models and their motions without the need for multi-camera setups. Additionally, other approaches predict UV texture maps from images [12, 29, 157], thus enabling the creation of photorealistic virtual humans at scale.

Creating realistic 3D virtual humans has been a longstanding goal in computer vision and computer graphics, with applications spanning telecommunications, training systems,

entertainment, online shopping, and medicine. Realistic appearance synthesis is fundamental to achieving photorealism in virtual humans, enhancing the sense of body ownership and, consequently, enriching the experience of training with these virtual contents.

1.3 Democratizing Skill Acquisition and Skill Transfer

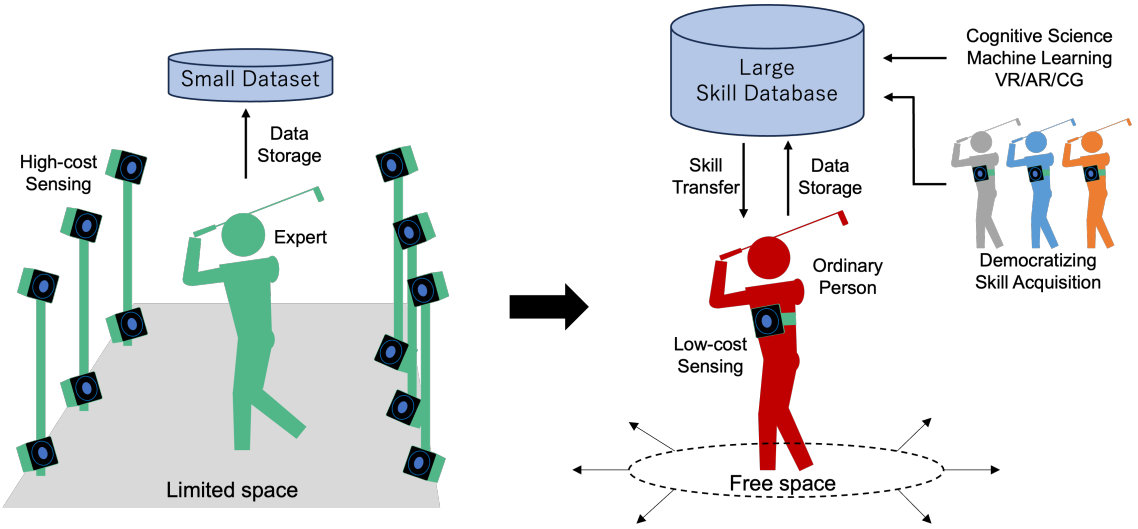

Figure 1.6: A comparison between the current limitations in skill acquisition and the envisioned future of democratized skill transfer. The left side illustrates the challenges faced due to limited access to data and coaching resources, while the right side depicts a future where advanced technologies enable widespread, personalized skill development through accessible devices and cloud-based platforms.

Advancements in computer vision and machine learning have significantly transformed sports analytics since the 2010s [46, 86]. Algorithms capable of detecting and tracking human movement from standard video footage have reduced the reliance on specialized equipment [115]. Deep learning models have been developed to recognize patterns in motion, classify actions, and even predict future movements [148,160]. This shift has enabled automated analysis of complex motions, providing coaches and athletes with detailed insights previously unattainable.

Despite these technological advancements, challenges persist, especially for beginners and amateur athletes. Access to high-end analysis tools and professional coaching is often limited due to cost, resource availability, and geographical constraints [152]. As depicted in Figure 1.6, users often receive limited, non-scientific coaching because they have access to only small datasets and localized expertise. For example, it is challenging for players to receive lessons from a wide range of coaches worldwide to select the best training methods tailored to their needs. Even when using high-end analysis tools like motion capture systems, the data stored locally is limited, restricting comparisons to a small dataset and hindering comprehensive performance evaluations.

Moreover, motion capture systems require substantial financial investment and special-

ized personnel to operate. Common indoor installations result in small capture volumes, further limiting their practicality for end-users and small groups—a phenomenon referred to as the "Curse of Cameras." This term encapsulates the limitations posed by the physical space requirements, equipment costs, and technical expertise needed to effectively utilize such systems. Consequently, the benefits of advanced motion analysis remain largely inaccessible to the broader population seeking to improve their skills.

Recent research has been tackling the problem of the "Curse of Cameras," and technological advancements in hardware and network infrastructures have been significant. The proliferation of affordable depth sensors, wearable devices, and high-resolution cameras has opened new avenues for motion capture and analysis [129]. Concurrently, improvements in cloud computing and data storage technologies enable real-time processing and sharing of large datasets.

We can foresee a future where compact devices capture detailed human motion information and access cloud databases in real-time. As shown in Figure 1.6, we imagine that individuals could simply wear a compact device—such as a chest-mounted camera or a wearable sensor suite—to capture their skill data using computer vision technology. This concept, which we refer to as "Democratizing Skill Acquisition," aims to make advanced skill capture and analysis tools accessible to anyone, regardless of their resources or location.

Users can upload their data to a cloud database or an online platform, similar to how videos are shared on platforms like YouTube. By interacting with a vast skill database that aggregates data from users worldwide, individuals can benefit from a diverse range of techniques and coaching styles. From this large dataset, artificial intelligence technologies can be leveraged to provide users with scientific coaching, offering personalized feedback and training programs. Feedback technologies such as virtual reality (VR) and augmented reality (AR) can enhance this experience by providing immersive and interactive training environments. This phase represents "Skill Transfer," where knowledge and skills are efficiently transmitted from experts to learners through advanced technological means.

The skill acquisition phase, where the system captures and interprets skills based on the dataset, is crucial. Individuals may struggle to identify keyframes or critical moments in their performance that require attention, and interpreting complex data can be daunting without foundational knowledge. This underscores the necessity for accessible, intuitive tools that can provide clear, actionable feedback. By employing user-friendly interfaces and leveraging AI-driven analysis, these tools can bridge the gap between raw data and meaningful insights.

Furthermore, democratizing skill acquisition has the potential to level the playing field by providing equal opportunities for skill development. It enables athletes from underrepresented regions or those lacking access to professional coaching to receive high-quality training resources. This not only fosters individual growth but can also contribute to the overall advancement of sports by uncovering untapped talent pools.

In summary, the integration of advanced computer vision, machine learning, and cloud technologies paves the way for a new era in skill acquisition and transfer. By making sophisticated analysis tools widely accessible, we can empower individuals to take control of their learning journey, receive personalized coaching, and ultimately enhance their performance. The vision depicted in Figure 1.6 represents a paradigm shift towards more inclusive and effective skill development practices, harnessing the full potential of technological innovation.

Chapter 2

Related Work

The proposed system in this work is related to sports skill acquisition, motion capture, style transfer, motion alignment, and discrepancy detection. In this chapter, we briefly discuss these related works.

2.1 Self-training System for Motor Skills

Recently many research works have developed training systems to help beginners develop their skills or to help trainees maintain their performance during training. For motor skills training, recent works [44, 90, 91, 141, 143, 158, 162] utilize multiple sensor feedback as a support for training. Works such as [59, 64, 133, 135, 154] propose multi-modal sports training systems based on sports theories. In their systems, users receive visual, haptic, and audio feedback once they are not moving their bodies or instruments in an ideal way. However, the ideal movement can differ from task to task, causing these methodologies hard to be applied to other sports.

Another way to learn motor skills is by imitating professionals' motions [57,63,83,84,91]. Ikeda et al. [72] propose a golf swing training system using professional golfers' motions. In their system, a user's motion is synchronized with a selected ideal professional's motion, and the two motions are overlaid and projected on the ground during training. Huang [69] demonstrates how motion-tracking technologies like Kinect can transform physical therapy into a more motivating and user-friendly experience. This pilot study underscores Kinect's promise in enhancing rehabilitation for young adults with motor disabilities. On the other hand, Sasaki et al. [122]also report the importance of beginners copying experts' motions and propose a climbing training system using pose prediction. Their system predicts and visualizes a pose of experts from users' current hands and feet positions. These recent works show the effectiveness of multi-modal feedback and the potential of applying neural networks to create AI teachers for sports training. However, due to significant differences in player level and physical ability between beginners and experts, it is hard for users to change their motion forms immediately to match the ideal forms. Thus, building a system that can guide users step by step to improve their sports skills remains challenging.

Recent work like Pose Tutor [40] has shown its ability to correct wrong poses in Yoga, Pilates, and Kungfu. They use a classification-based method for prediction and pose correction and develop an explainable method to find incorrectly formed joints from a human pose. However, they narrowly use a single frame for the network, lacking temporal information, which makes their method valid for evaluating static poses. Since most motor skills include sequential body movements, temporal information is essential when we study the professional's motion. In this work, we explore both temporal and spatial information to determine the crucial timing where the correction needs to be taken care of.

2.2 Machine Machine in Skill Acquisitions

Machine learning has recently gained traction in the field of skill acquisition, particularly in the development of self-training systems that can operate effectively without relying on extensive domain-specific knowledge [147]. This adaptability is advantageous for applications across a wide range of sports, as demonstrated by studies that predict volleyball trajectories from human poses or forecast basketball players' movements for tactical analysis [124, 125]. Another example can be seen in a climbing training framework where pose prediction is employed to guide beginners, using professional poses to suggest how users should place their hands and feet [123]. Despite the success of these methods in providing feedback and identifying pose discrepancies, they often leave users at a loss when it comes to making substantial adjustments to align their bodies with professional standards, especially if there are large gaps in physical capabilities or skill levels. This shortcoming highlights the need for more personalized and actionable guidance in self-training systems, where novices can progress incrementally rather than attempting to replicate expert-level motions immediately.

Within the broader paradigm of skill acquisition, latent space representations have emerged as an influential concept, serving as a bottleneck in neural network architectures that process complex inputs such as audio or motion data. These reduced-dimensional embeddings have been applied to tasks like segmenting different playing phases and evaluating performance quality [106, 117]. Some systems even visualize these embeddings, enabling users to locate their own skill levels relative to others in a database [15, 78]. However, these approaches generally overlook individual differences related to physical attributes or varying levels of expertise, which can significantly affect how users learn and adapt. They also tend to focus more on observing discrepancies rather than providing guidance on how to bridge those gaps in skill. As a result, they offer limited assistance to beginners seeking step-by-step instruction on improving their movements. There remains a pressing need for methods that account for temporal and contextual factors in motion data and deliver personalized, progressive feedback that users can apply to real-world training scenarios.

2.3 Motion Capture and Pose Estimation

Traditional motion capture systems often employ multiple cameras in controlled studio environments. These setups, commonly referred to as marker-based systems, require users to wear specialized suits containing active or passive markers to obtain accurate position data for each body segment. While standalone solutions such as IMU-based motion capture systems can be deployed in outdoor or less-controlled conditions, users must still attach sensors to their bodies. To address these constraints, marker-less motion capture algorithms have been developed [6,24,36,49,67,76,131,132,134,144], enabling the capture of human motion data without specialized suits.

Recent advancements have further enabled systems to operate outdoors using only a minimal number of cameras [13, 26, 43, 113, 118, 120]. These multi-camera setups provide robust and accurate pose estimation by observing subjects from multiple viewpoints. However, the high cost of hardware, the need for synchronization and calibration, and the requirement of professional operators can limit their practicality for consumer-level applications.

In contrast, single-camera human pose estimation has gained considerable attention due to its simplicity and cost-effectiveness. Some methods adopt infrared (IR)-based depth cameras to capture silhouettes for applications such as entertainment systems [17,129,149], but these are prone to outdoor performance issues caused by interference from natural light. Thus, research has increasingly focused on leveraging a single RGB camera to provide flexible and accessible motion capture solutions for everyday users.

Fueled by deep learning advancements, large-scale training datasets have become available for 2D and 3D human pose estimation [14,73,99,131]. In 2D human pose estimation, early neural network approaches directly regressed joint coordinates from images [140]. Although direct regression can be intuitive, subsequent works have shown that heatmap-based methods generally yield superior accuracy and reliability [32,34,35,105,136], making them prevalent in modern pipelines.

For 3D pose estimation, two prominent approaches have emerged:

- **Pipeline Approaches:** These methods first detect 2D joint coordinates, then perform a 3D lifting step [23,96,114,138]. This modular design enables interchangeable 2D detectors without retraining the entire pipeline. However, the accuracy of the 3D stage is heavily dependent on the quality of the initial 2D detection.

- **Direct 3D Joint Regression:** Rather than splitting the process into two stages, these methods predict 3D volumetric heatmaps [95, 111, 112] or location maps [100] directly from input images. By extending 2D heatmaps into the 3D domain, networks can fully exploit spatial information gathered through convolutional layers. This design typically offers a more unified framework for pose estimation.

2.4 Motion Style Transfer

Motion style transfer is an emerging area that aims to modify a given motion sequence to reflect a new style while preserving the underlying motion content. These styles often correspond to specific emotional states, unique character traits, or personalized attributes, enriching virtual avatars or characters with a diverse range of motions. As machine learning techniques mature, neural network-based motion style transfer has become increasingly sophisticated and is inching toward adoption in real-world applications such as animation, gaming, and even sports performance analysis.

Style transfer, initially popularized in the image domain [33,50,70,75,79], has gradually been adapted for motion data [10, 41, 65, 66, 107, 150]. Early image style transfer methods used neural networks to extract both style and content features, and an iterative optimization process generated stylized images by fusing these elements [50]. Subsequent work employed a real-time optimization framework based on perceptual loss [75], significantly improving practical usability. Later research introduced adaptive instance normalization (AdaIN), which helps encode style statistics and reapply them during decoding, leading to more efficient and flexible style transformations.

Inspired by these advances, motion-based techniques initially followed a similar trajectory. The first motion style transfer approaches adopted iterative optimization processes [66], but more recent methods transitioned to training-based paradigms to enable faster inference and reduced memory usage [41,65]. Building on successes in the image domain, Aberman et al. [10] introduced temporally invariant AdaIN to handle style transfers in time-varying sequences, mitigating issues like temporal discontinuities.

Follow-up studies have explored various avenues, including label-free training strategies, more robust spatial-temporal modeling, and mechanisms for generating diverse motion styles. For instance, [151] proposed an autoregressive flow-based architecture that learns style features without explicit style labels. Similarly, [108] emphasized spatial-temporal convolutions and random noise to expand the range of generated styles, while [146] integrated kinematic constraints for more realistic and natural movements. Other frameworks, such as [97], prioritized real-time style transformations to support interactive applications, and Motion Puzzle [74] introduced a method for training without style labels, allowing segment-specific and time-varying style transformations.

Despite these notable strides, current motion style transfer methods still face some key challenges. Many methods work best when transferring styles within narrow, homogeneous motion categories like walking or running, and can struggle with broader motion repertoires that include actions of differing content and complexity. In practical scenarios, such as sports training, motions can be diverse, sequential, and highly context-dependent, making it difficult for current methods to generalize. Achieving style transfer that both preserves the core motion and accommodates a variety of styles without needing extensive retraining for each new motion category remains an open challenge. Addressing these

constraints is particularly important for real-world applications that demand efficiency, scalability, and adaptability.

2.5 Sequential Motion Alignment

An efficient way to evaluate whether a person is performing a motion correctly is to compare their motion with others whose motion is considered correct. However, owing to the various timings and speeds of motion of different individuals, we must align the motions to make them comparable. A conventional method for aligning two temporal sequences is dynamic time warping (DTW), which was introduced by Berndt and Clifford [22]. In this method, every index in a sequence is matched with one or more indices from another sequence, and the mapping of the indices from the first sequence to the other sequences must be monotonically increasing.

The DTW concept has been introduced in several domains. For example, Ikeda et al. [71] proposed a real-time golf swing projecting system that simultaneously visualized professional and user forms with matched timing. To align the two motions in real-time, they measured the DTW only over a short period and penalized the previous cost value at a later time. Halperin et al. [55] utilized the concept of DTW in speech and presented an audio-to-video alignment method for matching speech-to-lip movements. However, the alignments are mainly based on pose error and are without spatial information among joints and temporal relationships among frames. Therefore, while the use of DTW can align motions with different timing, we still need to find out which frame in the motion sequence is important by other metrics.

On the other hand, self-supervised neural networks have recently been developed to tackle video alignment tasks using the latent space representation [42, 101, 127]. In this approach, an embedder is used to compress input videos into a latent space. After the embedding process, a loss is designed to find correspondences through time in the latent space, thus encouraging the network to learn a latent space where similar motions should appear to be close. This method helps synchronize high complexity temporal sequences, such as videos. In addition, previous work such as [88] has shown the usefulness of self-supervised neural networks in aligning sports motion with videos and 3D human poses. In this study, we base the loss of our network on the temporal cycle-consistency loss used by Dwibedi et al. [42], but apply DTW along with the loss for smooth temporal matching.

2.6 Discrepancy Detection

By comparing the two synchronized motions, we can determine the difference between them in terms of human postures. In early studies, abnormal detection referred mainly to finding patterns in data that did not match the expected behavior [7, 126]. Recently, two methods have been proposed for detecting abnormalities In the first approach, abnormal

information is referred to as prior knowledge. For example, Parra-Dominguez et al. [109] trained a binary classifier on annotated data to determine whether abnormal events occurred during a stair descent. In the second approach to abnormal detection, abnormal information is not provided in advance. The research group of Nater et al. [104] proposed an unsupervised learning method for learning normal human behavior. They used a hierarchical representation of the appearance and action level of regular movements to detect abnormal events.

While a network may be trained to detect abnormal events, such as falling to the ground, we focus on whether a neural network can be trained to automatically detect fine-grained differences between two regular motions. We call this detection of the fine-grained difference discrepancy detection. The most relevant of these studies is that of Abati et al. [7]. They designed a deep autoencoder with a parametric estimator that learned a probability distribution from the latent space to detect discrepancies. The encoder effectively remembered standard samples and could distinguish between normal and abnormal images. However, although the network could easily detect surprise samples, fine-grained differences among standard samples were not discussed. In addition to image-based methods, recent studies have focused on systems that use 3D human pose information [20,137]. In this study, we apply discrepancy detection to both videos and 3D human poses and discuss the ability of the system to detect fine-grained differences between two input motions.

Chapter 3

Research Proposal

3.1 Limitation In Previous Works

In the field of sports skill acquisition, replicating professional players' motions is a common approach to improving performance. People watch the movements of professional athletes on television or the Internet and try to emulate them physically. To accelerate this process, many systems have recently been developed to help users understand the movements of professionals [57, 63, 83, 84]. However, as reviwed in the preious chapter, in these previous works and systems, users may struggle to refine their movements due to a lack of guidance. In this section, we raise three main research questions and the motivations behind our research.

3.1.1 Who to Follow?

Firstly, previous systems often use fixed professional template motions as a universal standard. This can be suboptimal, as it disregards individual conditions like biomechanical differences or personal playing styles. Training could be more efficient if users imitate target motions that better fit their unique characteristics.

3.1.2 What to Learn?

Secondly, the appearance of the learning target often differs significantly from the user. Differences in body size, muscle composition, and other physical attributes make it challenging for learners to imagine the desired state in their minds—a key aspect of motor skill acquisition. This disconnect hampers effective imitation-based learning.

3.1.3 How to Improve?

Lastly, users receive little to no guidance on specific aspects of their movements, such as the timing of certain actions or which body parts to focus on. This lack of detailed instructions on which specific timing within the motion to focus on, which parts of the body require attention, and how to adjust movements to align more closely with professionals,

can hinder their ability to refine their movements and align more closely with professional standards.

3.2 Research Approach

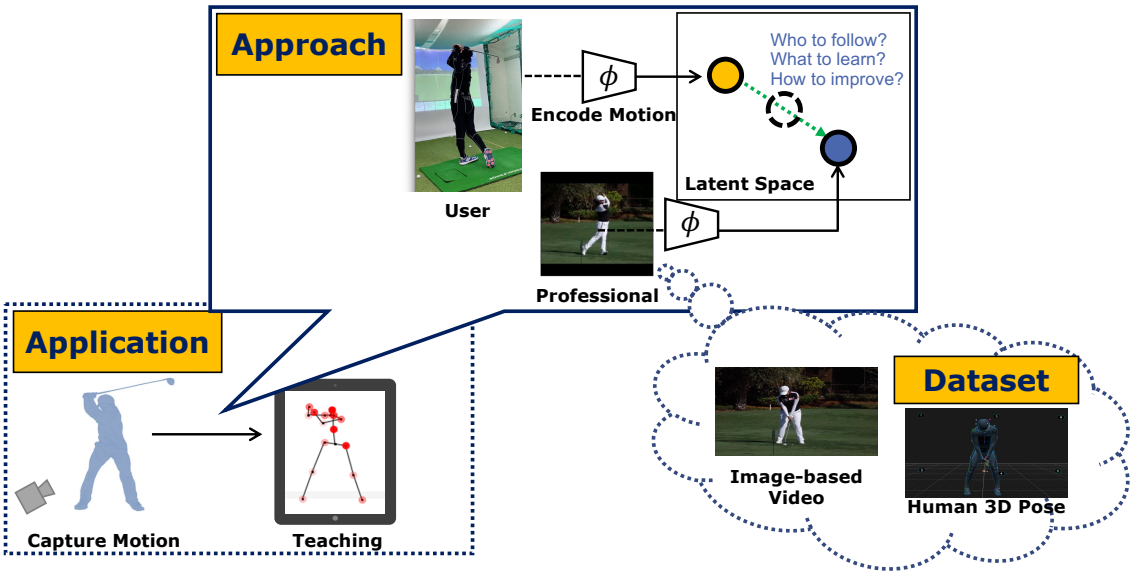

Figure 3.1: The overall scheme of the reseaerch.

To address the gaps found in previous limitation in this field, we urge to propose a framework that can be effective for self-training using recent developed machine learning. Figure 3.1 outlines the objectives of our research. Our goal is to create a system that captures users' motions and provides suggestions to improve their forms by comparing these motions with those in our dataset. To achieve this, we train a neural network to learn effective latent representations from motion data across all skill levels. We then compress the users' motions into this latent space and compare them with professionals' motions within the same space. This approach, which does not require prior knowledge, allows us to apply the system to various skill training processes, making it versatile and scalable.

To achieve *Democratizing Skill Acquisition*, and to answer the three rasied research questions, we propose a training framework using motion datasets to address these gaps, as illustrated in Figure 3.2. This pipeline comprises three main steps:

1. **Personalized Learning Target Selection**: Instead of immediately training from an expert's high-level motion, we suggest an intermediate-level motion more compatible with the user's capabilities. This helps novices tackle skill acquisition in a less intimidating, more progressive manner, aligning with the concept of *Zone of Proximal Development* [145] and *Challenge Point* (Figure 3.3).

2. **Representation with User's Appearance**: In this phase, the target motion is

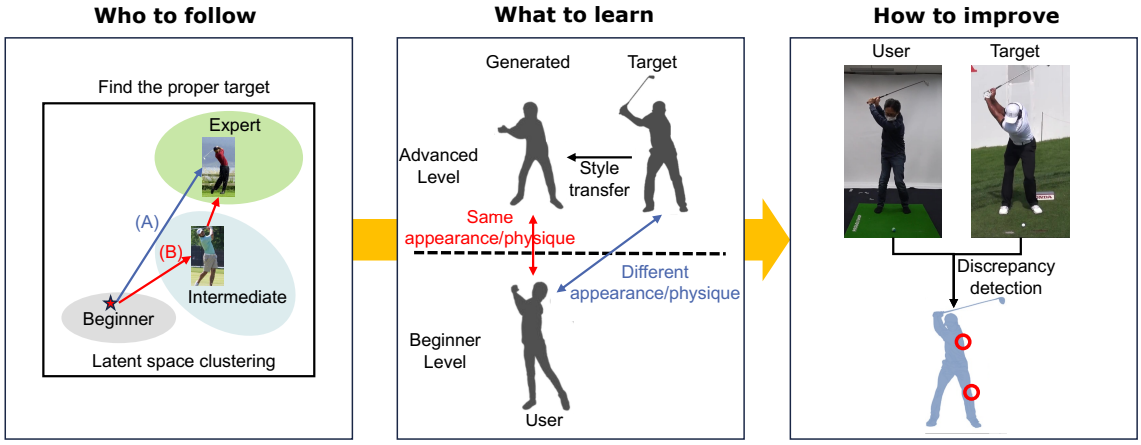

Figure 3.2: A recipe for motor skill imitation.

converted to match the user's own model, effectively creating a personalized higher-level version of the user. By imitating this "ideal me," learners can more easily visualize and engage with their desired state.

3. **Fine-Grained Motion Comparison and Guidance**: The final step focuses on guiding users with interpretable clues that highlight critical timings and specific body parts. This granular feedback is meant to help users spot exactly where and when to adjust their form.

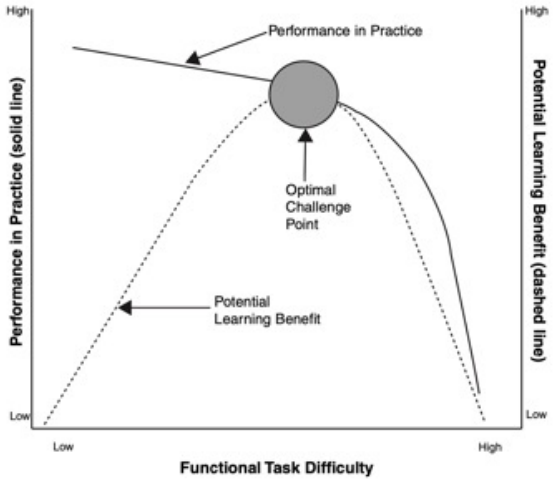

Figure 3.3: The *Challenge Point* [52]

In this work, we realize these three steps through two complementary systems:

- **Coach Navi**: Implements the first two steps—*Personalized Learning Target Selection* and *Representation with User's Appearance*. By leveraging a neural network–based style transformer to capture skill levels and user appearances, Coach Navi helps novices develop a more approachable learning target.

- **AI Coach**: Concentrates on the third step—*Fine-Grained Motion Comparison and Guidance*. Employing discrepancy detection in a latent space, AI Coach offers specific, frame-by-frame or segment-by-segment error cues for deeper refinement.

Together, these systems illustrate how our pipeline adapts to both novice-level concerns (such as intimidation by expert forms) and advanced needs (precise error detection), thereby promoting a broader "democratization" of skill acquisition across multiple levels of expertise.

3.3 Research Positioning

The systems presented in this dissertation—**Coach Navi** and **AI Coach**—offer a comprehensive solution for motor skill learning that stands out from existing methods. While many training systems provide either generic motion templates or isolated feedback mechanisms, our approach integrates multiple innovations to address a wider range of learner needs and skill levels.

Table 3.1 summarizes key features of different form-correction approaches. *Sensor-based Analysis* systems often provide immediate feedback (Real-time ✓) but typically rely on specialized hardware and are sport-specific (✗ in Generalizability). They can capture continuous motion (✓ in Temporal Analysis), yet they detect errors only at a low or partial level (△ in Error Recognition) and rarely adapt to individual traits (✗ in Personalized Adaptation). *Target Motion Imitation* methods frequently offer real-time guidance and can be extended to multiple sports or motions, preserving sequence data but providing only limited error details and no skill-level adaptation. *Classification-based Correction* frameworks partially operate in real time and tend to handle various poses. They are generally single-frame or near-static but do classify correct or incorrect poses effectively. Finally, *Our Proposed Approach* supports partial real-time analysis, works across multiple activities, considers full motion sequences, identifies errors more holistically, and personalizes feedback according to the user's skill and body attributes. This combination addresses many gaps left by the other categories, particularly in adapting feedback to individual users while still providing meaningful temporal analysis of their movements.

Table 3.1: Comparison of major approach categories in form correction/training. Symbols denote full support (✓), partial support (△), or no support (✗) for each feature.

Approach	Real-time	Generalizability	Temporal Analysis	Error Recognition	Personalized Adaptation
Sensor-based Analysis [44, 59, 64, 90, 133, 135, 143, 154]	✓	✗	✓	△	✗
Target Motion Imitation [57, 63, 72, 83, 84, 91]	✓	✓	✓	△	✗
Classification-based Correction [19, 31, 117, 122, 147]	△	✓	△	✓	✗
Ours (Proposed)	△	✓	✓	✓	✓

Furthermore, we highlight the main contribution of this work as the following:

1. **Personalized and Intermediate Learning Targets:** Unlike standard practices that urge beginners to mimic advanced expert movements immediately, Coach Navi tailors each user's progression via intermediate-level targets. This scaffolding strategy reduces intimidation and fosters more consistent improvement. Empirical results from our user studies confirm that participants found intermediate motions more achievable and motivating, aligning with the Zone of Proximal Development principle.

2. **Avatar-Based 3D Visualization with User's Appearance:** Instead of using general skeleton overlays or static 2D video, Coach Navi renders a *personalized "ideal me" avatar*, enhancing self-representation and body awareness. Qualitative feedback and improved performance metrics indicate stronger user engagement, as participants could more intuitively visualize the target posture and motion (reductions in MPJAE).

3. **Fine-Grained Discrepancy Detection and Actionable Feedback:** AI Coach pinpoints discrepancies through a latent-space analysis that goes beyond simple kinematic comparisons, generating precise *error poses* and an easy-to-understand scoring system. Targeted feedback helps learners isolate the critical frames/timings to correct, effectively reducing frustration. Short practice sessions showed that users could swiftly target exact mistakes, thus enhancing the accuracy of subsequent swings.

4. **Comprehensive Pipeline Covering Novice to Advanced Progression:** Existing systems often cater primarily to beginners (e.g., simplified guidance) or to high-level athletes (e.g., advanced biomechanical analytics). In contrast, we present a cohesive pipeline: from *intermediate* and motivating start points (Coach Navi) to *fine-grained* corrections and advanced refinement (AI Coach). User testimonies and performance metrics support the synergy of both approaches. Beginners can start with manageable motions, then transition into deeper, frame-by-frame corrections as they progress, ensuring continued engagement and multi-stage development.

5. **Foundational Principles for Democratized Skill Acquisition:** Our pipeline champions accessibility (eliminating overreliance on professional coaches), personalization (adapting to each learner's level), and adaptability (guiding skill advancement progressively). The user studies not only indicate short-term gains (e.g., fewer errors, improved angles) but also underscore high user motivation and potential for long-term retention given extended practice intervals. This aligns with the ultimate goal of *democratizing skill acquisition* in sports and beyond.

3.3.1 Clarifying the Intermediate Approach and Latent-Space Innovation

Although this dissertation emphasizes generating intermediate-level motions for user training, we do not claim that no prior research has ever attempted adaptive or gradational approaches. Indeed, some existing works provide *adaptive tasks* [8, 130] (e.g., adjusting the difficulty of the environment, such as making a basketball hoop's radius larger for novices [142]), which indirectly aids beginners without forcing them to imitate an expert-level form outright. Furthermore, previous works have investigated the effects of using customized avatars, such as idealistic or realistic version of users, during training [82]. However, there is a gap where methods that explicitly generate an *intermediate-level form* for the learner to imitate remain underexplored. Many training frameworks assume that replicating the expert motion is universally the best standard, or they lack a mechanism to define and synthesize a mid-tier motion. As a result, few existing systems allow users to see and practice a form that is neither too basic nor too advanced but is physically achievable yet still progress-oriented.

One could, in principle, hand-craft intermediate motions by linearly interpolating joint angles or keyframes between beginner and expert data. However, relying on manual or simplistic blends poses inherent limits, such as unclear transitions or a lack of generalization to new users. By contrast, adopting a latent-space approach (e.g., using VAEs) provides:

- **Scalability and Smooth Interpolation**: The model learns a lower-dimensional representation where motions are arranged in a continuous, structured manner. Intermediate motions can thus be generated by traversing this latent space, producing smoother transitions that preserve biomechanical consistency.
- **Generalization**: Once trained, the latent model can adapt to a wide range of user data without rewriting rule-based templates. This yields a more flexible solution for novel sports or additional skill levels.
- **Integration of Multiple Features**: The latent space can incorporate additional constraints (e.g., body shape, skill difficulty, style variance), allowing the system to generate different "middle" forms for each user's unique combination of attributes.

While we leverage latent-space techniques and advanced neural networks, we acknowledge that latent representations are not altogether new in motion generation or style transfer domains. Our core novelty arises from integrating these generative methods into a sports-training pipeline that specifically targets intermediate-level skill forms rather than generic or solely expert styles. This specialized focus on bridging novice and expert, through a skill-labeled latent space, distinguishes our work from prior approaches that might adapt tasks or environment constraints but not the actual form of the motion.

In sum, producing an "intermediate-level motion" is not theoretically impossible without machine learning, but our ML-based pipeline offers scalability, smoother transitions, and adaptability that hand-crafted solutions often lack. This approach helps novices train with a more realistic stepping stone, aligned with the *Zone of Proximal Development* and *Challenge Point* theories, and has the potential to incorporate additional data or constraints for further refinement.

3.4 Thesis Overview

This dissertation is structured into seven chapters, each contributing to the overarching goal of democratizing skill acquisition through innovative training systems. Below is a brief overview of each chapter:

3.4.1 Chapter 1: Introduction

The Introduction chapter sets the stage by outlining the motivation behind democratizing skill acquisition, the significance of motor skill training in sports, and the limitations of existing training systems. It presents the research questions, objectives, and the structure of the dissertation, providing a roadmap for the reader.

3.4.2 Chapter 2: Related Work

In the Related Work chapter, we review existing literature and technologies pertinent to sports skill acquisition, motion capture, style transfer, motion alignment, and discrepancy detection. This chapter identifies key gaps and limitations in current approaches, establishing the foundation for our proposed solutions.

3.4.3 Chapter 3: Research Proposal

The Research Proposal chapter introduces a three-step training pipeline designed to enhance motor skill learning. It details how this pipeline is implemented through two complementary systems—**Coach Navi** and **AI Coach**—and highlights the novel contributions that differentiate this work from existing methods. The section on uniqueness and contributions elucidates the innovative aspects of our approach, emphasizing its comprehensive and scalable nature.

3.4.4 Chapter 4: Personalized Motor Skill Training with Adaptive Learning Targets (Coach Navi)

This chapter presents **Coach Navi**, a motor skill training system tailored to assist users in improving their golf swing through intermediate-level motion targets and personalized 3D avatar visualization. It delves into the system's architecture, including the Variational Autoencoder with Motion Style Transfer (VAE-MST) and the motion style transformer. The chapter also details the methodology and results of a user study comparing Coach Navi with traditional video playback and skeleton visualization methods, demonstrating its effectiveness in enhancing motor skill learning.

3.4.5 Chapter 5: Motor Skill Training System using Motion Discrepancy Detection (AI Coach)

In this chapter, we introduce **AI Coach**, a motor skill training system that leverages neural networks to detect and highlight fine-grained discrepancies between a user's motion and that of professional players. The chapter covers the design and implementation of the motion synchronizer and discrepancy detector, as well as the development of the decoder for intermediate pose restoration. It also discusses the applications for discrepancy visualization and user interaction. Empirical evaluations through multiple user studies are presented, showcasing AI Coach's superior performance in reducing MPJPE, MPJAE, and LSD compared to conventional training methods.

3.4.6 Chapter 6: Discussion

The Discussion chapter synthesizes the findings from the Coach Navi and AI Coach studies, analyzing their contributions to the internal models of motor skill acquisition—the motor and sensory models. It highlights overarching themes such as personalization, 3D visualization benefits, actionable feedback, and alignment between AI and human coaching priorities. The chapter also addresses short-term versus long-term gains and outlines future enhancements to the motor models, including the integration of bioinformatic data and advanced biomechanical modeling. Additionally, it presents a detailed section on the unique contributions of this dissertation, emphasizing how the integrated features of Coach Navi and AI Coach advance the field beyond existing training systems.

3.4.7 Chapter 7: Conclusion

The Conclusion chapter summarizes the key findings and contributions of the dissertation, reiterating how Coach Navi and AI Coach collectively achieve the goal of democratizing skill acquisition. It reflects on the effectiveness of the proposed systems, discusses the implications for future research and practical applications, and outlines potential avenues for further development. The chapter closes by affirming the significance of integrating personalized learning targets with fine-grained discrepancy detection in creating comprehensive, adaptive motor skill training solutions.

Chapter 4

Personalized Motor Skill Training with Adaptive Learning Targets

4.1 Overview

Mastering complex motor skills in sports often presents significant challenges for beginners. Traditional learning methods, such as attending coaching sessions or watching expert demonstrations, may not always be accessible or effective for every individual. Beginners may struggle with directly imitating professional-level movements due to substantial differences in skill level, physical capabilities, and personal learning styles. This can lead to frustration, decreased motivation, and a slower learning process.

In the realm of golf swing training, these challenges are particularly pronounced. The golf swing is a highly intricate motion that requires precise coordination, timing, and technique. Beginners attempting to replicate the swings of professional golfers may find the gap too wide, making it difficult to identify the specific aspects they need to improve. Moreover, existing training systems often rely on fixed templates or generalized feedback, failing to account for individual differences and personalized learning needs.

To address these issues and advance the concept of **democratizing skill acquisition**, we propose **Coach Navi**: a self-training system designed to navigate users through intermediate-level motions tailored to their current skill level. Coach Navi aims to empower users by tailoring the learning experience to their current skill level and individual characteristics, thereby making the process of acquiring new motor skills more intuitive and effective.

The system comprises three main components: a motion navigator, a motion style transformer, and a motion visualizer. The motion navigator analyzes the user's current performance and identifies appropriate learning targets from a motion database. By selecting intermediate motions that are closer to the user's abilities, we provide attainable stepping stones toward advanced proficiency, aligning with the concept of the "Zone of Proximal Development" [145]. The motion style transformer utilizes a motion style trans-

fer network to extract skill features—interpreted as styles—from the motion data. This network transforms the selected target motions to match the user's own appearance and physical characteristics, effectively creating a personalized avatar. By visualizing an idealized version of themselves performing the target movements, users can more effectively imagine the desired state, facilitating more intuitive and effective learning. The motion visualizer presents the personalized target motions to the user, allowing them to observe and imitate the movements in an interactive manner. By focusing on attainable targets that are personalized, users are more likely to remain motivated and engaged in the learning process.

Through these integrated modules, Coach Navi offers a novel approach to self-training that emphasizes personalization and incremental learning. By guiding users through intermediate steps tailored to their abilities, the system facilitates a smoother progression toward advanced skill levels. Users can focus on specific aspects of their performance, receive actionable feedback, and visualize their improvement over time.

To evaluate the effectiveness of Coach Navi, we conducted extensive experiments and user studies. Our motion style transfer network was tested against common VAE implementations, demonstrating superior performance in capturing and transforming golf swing motions across different skill levels. User studies involving beginner golfers revealed that Coach Navi significantly enhances skill acquisition compared to traditional training methods. Participants reported increased motivation, better understanding of the target movements, and a more enjoyable learning experience.

Our main contributions in this work are as follows:

- We introduce **Coach Navi**, an innovative self-training system that personalizes the learning experience by selecting intermediate-level motions tailored to the user's current skill level.
- We develop a motion style transfer network that adapts target motions to match the user's appearance and physical characteristics, facilitating personalized visualization and deeper engagement.
- We design a motion navigator that intelligently recommends appropriate learning targets from a comprehensive motion database, aligning with educational theories on effective learning progression.
- We implement a motion visualizer that presents personalized target motions in an interactive manner, enhancing the user's ability to observe, imitate, and internalize the desired movements.
- We conduct comprehensive experiments and user studies that demonstrate the effectiveness of Coach Navi in improving skill acquisition, user motivation, and overall learning experience compared to traditional methods.

4.2 Method

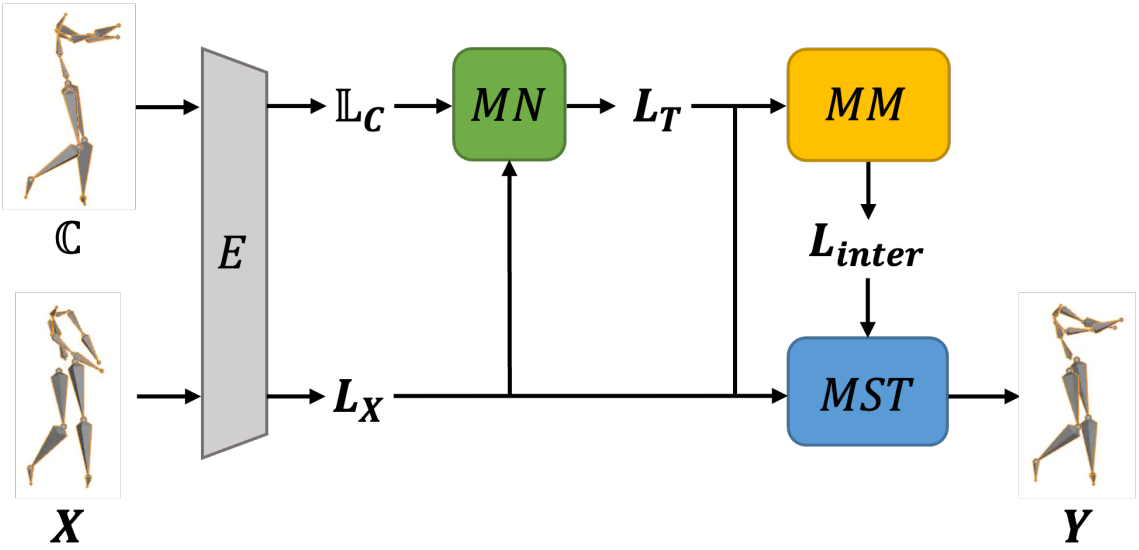

Figure 4.1: System overview: The user's input motion $\boldsymbol{X}$ and candidate motions $\mathbb{C}$ are encoded into latent space by encoder **E**. The motion navigator **MN** selects an optimal learning target. The motion manipulator **MM** combines the user's latent vector with the target to create an intermediate representation, which the motion style transformer **MST** decodes into the output motion $\boldsymbol{Y}$.

This study aims to develop a system that captures a user's motion and provides an appropriate learning target to help improve their form. To achieve this, we propose a method that involves training a neural network using a motion database containing various styles. After training, the system encodes the user's motion into a latent space and compares it with motions from the database within this latent space.

Figure 4.1 illustrates an overview of the proposed system. The workflow is divided into three main components: motion navigation, motion style transfer, and motion manipulation.

Firstly, the system receives the user's motion input, denoted as $\boldsymbol{X}$, and a set of candidate motions from the database, represented as $\mathbb{C} = C_i \mid i = 1, 2, \ldots, n$. The encoder, denoted as $\boldsymbol{E}$, embeds these input motions into the latent space, resulting in latent representations $\boldsymbol{L_X}$ for the user's motion and $\mathbb{L}_{\boldsymbol{C}} = L_{C_i} \mid i = 1, 2, \ldots, n$ for the candidate motions. The encoder is trained to map similar motions to nearby points in the latent space, effectively capturing the underlying features that characterize different motion styles.

Next, utilizing the learned latent space, the motion navigator $\boldsymbol{MN}$ selects an optimal learning target $\boldsymbol{L_T}$ by measuring the Euclidean distances between the user's latent vector $\boldsymbol{L_X}$ and each candidate latent vector in $\mathbb{L}_{\boldsymbol{C}}$. This process identifies the candidate motion that is most suitable for the user to learn from, based on similarity and proximity in the latent space. By selecting a target that is neither too easy nor too difficult, the system ensures that the learning target is appropriate for the user's current skill level.

The motion manipulator $\boldsymbol{MM}$ then integrates the user's latent vector $\boldsymbol{L_X}$ with the

selected target's latent vector $\boldsymbol{L_T}$ to create an intermediate latent vector $\boldsymbol{L}_{\text{inter}}$. This intermediate vector represents a blend of the user's current motion and the target motion, facilitating a gradual progression in skill acquisition. The integration can be achieved through interpolation or other combination techniques that weight the influence of each latent vector.

Finally, the motion style transformer $\boldsymbol{MST}$ decodes the intermediate latent vector $\boldsymbol{L}_{\text{inter}}$ to generate the intermediate motion $\boldsymbol{Y}$. This motion possesses the desired style characteristics of the target motion while retaining the user's own body appearance and unique attributes. By presenting $\boldsymbol{Y}$ to the user, the system provides an attainable learning target that is personalized and aligned with the user's capabilities, helping them to improve their skills effectively and intuitively.

This approach allows users to visualize and practice motions that are tailored to their current level, bridging the gap between their existing abilities and desired performance. By leveraging the latent space representations and neural network transformations, the system offers a flexible and adaptive method for motor skill training without requiring extensive domain-specific information or manual intervention.

4.2.1 Motion Navigator: Exploring Fine-Grained Motion Styles in the Latent Space

To enable effective motion navigation, we aim to design a network that learns a latent space where motion similarity is preserved and skill levels are distinguishable. A common approach to achieving this is by constructing an autoencoder, where the input and output are the same motion data. By utilizing an autoencoder (AE), we can extract meaningful motion features from the input data.

However, while a standard Autoencoder compresses and reconstructs inputs, it does not enforce a smooth or well-structured distribution in the latent space. In contrast, we adopt a *Variational Autoencoder (VAE)*, which imposes a probabilistic framework—typically mapping data points to a Gaussian distribution. This design choice yields two key advantages in our motion context:

1. **Better Generation and Interpolation:** A VAE learns a continuous, regularized latent space where points close together produce motions with similar features. This smoothness is crucial for generating intermediate motions, because linearly moving through the VAE's latent space yields realistic transitions. A plain AE might create a latent representation without any guarantee of continuity, making interpolation less reliable.

2. **Robustness and Generalization:** By introducing a KL-divergence term to align the latent distribution with a known prior (e.g., $\mathcal{N}(0, 1)$), the VAE prevents overfitting and ensures the latent representations remain coherent across varying skill

levels. This helps us model subtle motion variations—important when learning both beginner- and advanced-level forms—while preserving the global structure required to identify skill gaps.

In our work, we therefore employ the concept of a VAE to train a network that learns a latent space representing the essential features of the motions. This framework allows us to differentiate skill levels more effectively, as motions with similar skill-related traits cluster near each other in the latent space, while still preserving enough variation to capture individual user differences. Consequently, the learned latent representation underpins our ability to select intermediate motions well-suited to a given user's current abilities, bridging the gap between novice and advanced forms with smoother, more natural transitions.

Latent Space Navigation

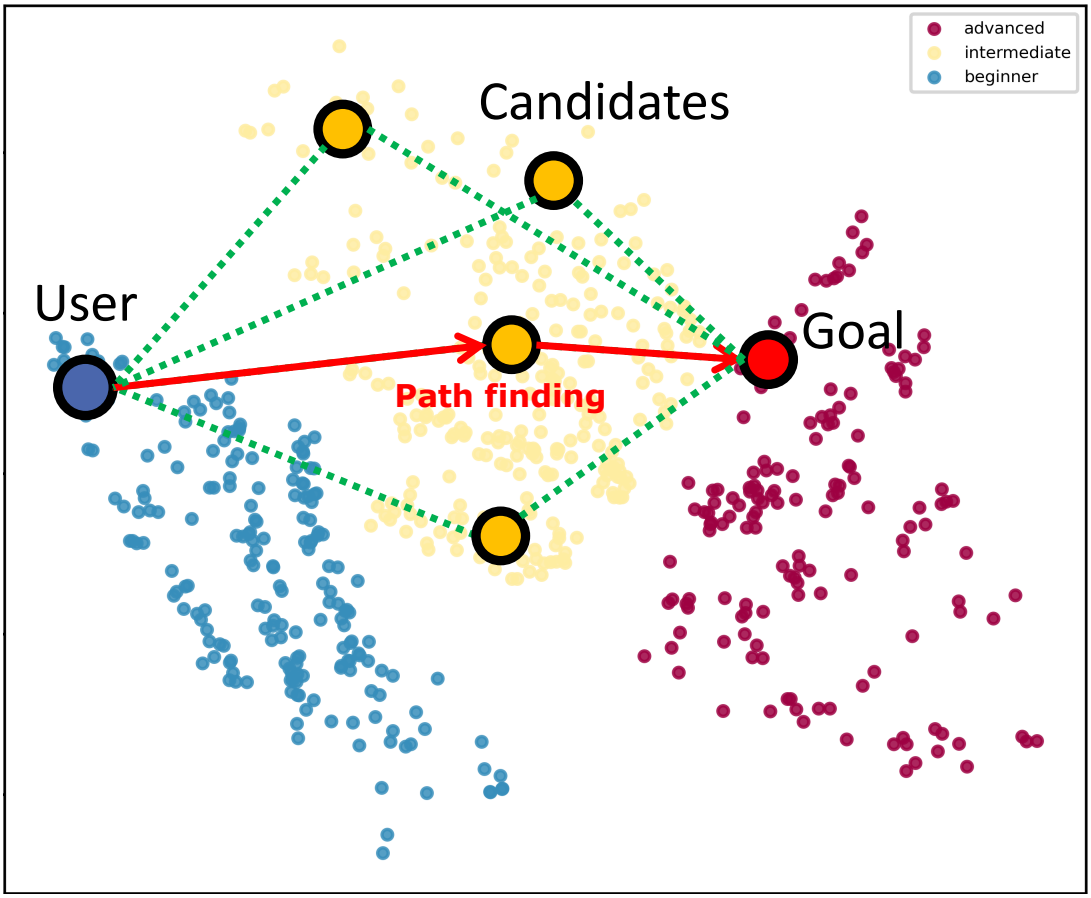

Figure 4.2: Overview of the path finding in the latent space. This 2D latent space visualization is generated by PCA.

Our goal is to learn a network where the latent space reflects the skill levels of motions, such that motions with similar skill levels are located close to each other. As depicted in Figure 4.2, by training on a large dataset containing motions of various skill levels—beginner, intermediate, and advanced—we can form a latent space where skill

progression is represented spatially.

Once we have trained such a network, we implement the motion navigator as follows:

- The encoder $\mathbf{E}$ compresses the user's motion $\boldsymbol{X}$ and the candidate motions from the database $\mathbb{C} = C_i \mid i = 1, 2, \ldots, n$ into latent vectors $\boldsymbol{L_X}$ and $\mathbb{L}_{\boldsymbol{C}} = L_{C_i} \mid i = 1, 2, \ldots, n$, respectively.
- The system selects a motion with the highest skill level as the training goal. This motion can be chosen by the user; if the user does not make a selection, the navigator chooses a target that is closest to the user's motion in the latent space.
- In the latent space, the navigator finds the optimal route from the user's latent vector $\boldsymbol{L_X}$ to the goal latent vector by using Dijkstra's algorithm [39]. It outputs the latent vector $\boldsymbol{L_T}$ of the intermediate learning target, guiding the user toward progressive skill improvement.

Considering computational efficiency during the training process, we constrain the optimization to visit only one node within each skill-level cluster. This means that if the user is currently at the beginner level, the navigator will guide them to the next immediate level, such as the intermediate level, ensuring a manageable progression in skill acquisition.

Network Design

To implement the encoder network, we first propose a motion VAE architecture, as depicted in Figure 4.3. This baseline network consists of a two-layer structure for both the encoder and the decoder. For the input and output formats, we use the rotations of the human body's joints, along with the root position (specifically, the hip position). For representing rotations, we adopt a 6D continuous rotation representation, which can be converted among quaternion and Euler representations. This representation has been shown to be effective for neural network learning in previous works [163].

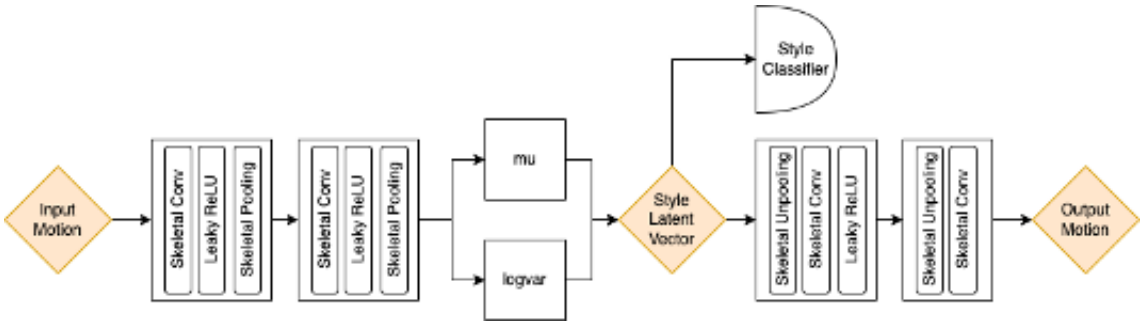

Figure 4.3: Baseline motion VAE network architecture.

To better extract features from the motion capture data, we employ skeletal convolution and pooling methods for encoding and decoding the motions [9]. As illustrated in

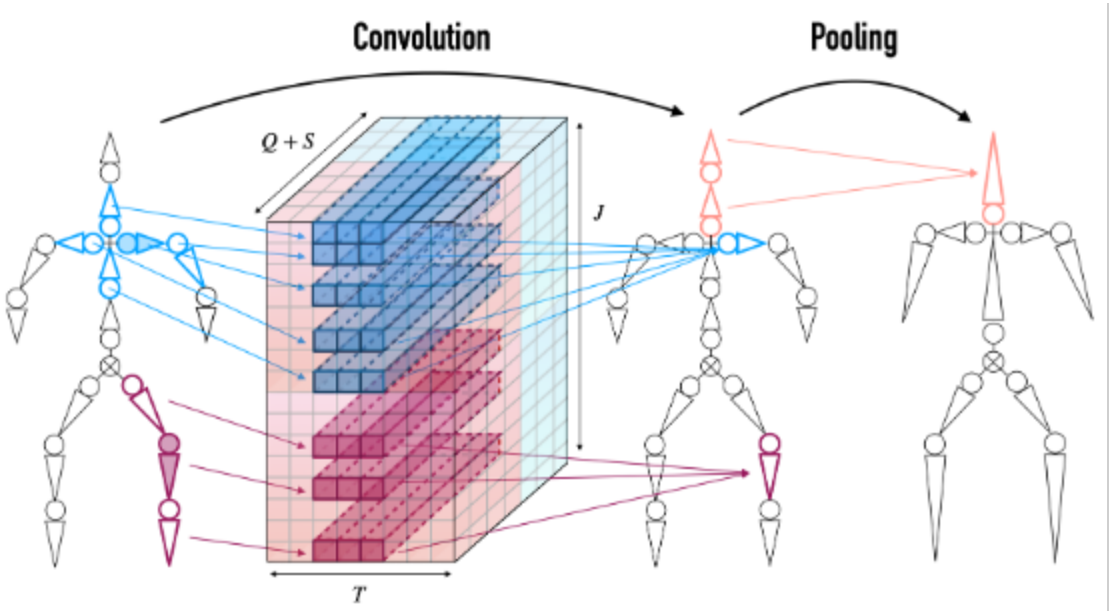

Figure 4.4: Skeletal convolution operation, which processes joints based on the skeletal hierarchy. [9]

Figures 4.4 and 4.5, these skeletal operators consider the connectivity of adjacent joints, allowing the network to learn the hierarchical structure of the human skeleton.

In the encoder, we apply skeletal convolution and pooling layers twice to progressively extract higher-level features. The output is then passed through layers that predict the mean and log variance of the latent distribution, capturing the probabilistic nature of the data in the latent space.

After sampling a latent vector from this distribution, we use the decoder to reconstruct the output motion. The network includes an optional style classifier to encourage motions of similar styles to form distinct clusters in the latent space, enhancing the separation of different skill levels.

Since our motion VAE is trained in a self-supervised manner, we define a reconstruction loss to ensure accurate reconstruction of the input motion. We compute the L1 loss and the global position loss between the input motion and the reconstructed output motion.

The L1 loss $\mathcal{L}_{\text{L1}}$ is defined as:

$$\mathcal{L}_{\text{L1}} = \mathbb{E}\left[|D(E_c(m), E_s(m)) - m|_1\right], \tag{4.1}$$

where m is the input motion, E_c and E_s are the content and style encoders, and D is the decoder.

The global position loss $\mathcal{L}_{\text{fk}}$ is computed using a forward kinematics (FK) layer applied to both the reconstructed motion and the input motion:

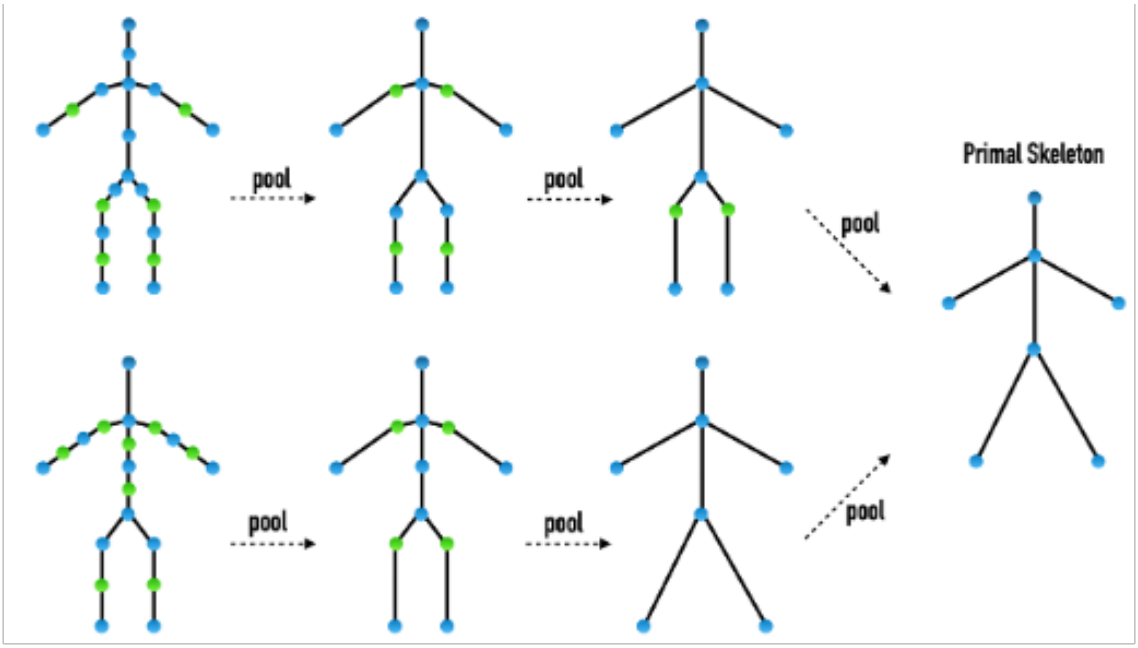

Figure 4.5: Skeletal pooling operation, which reduces the dimensionality while preserving structural information. [9]

$$\mathcal{L}_{\text{fk}} = \mathbb{E}\left[\left|FK\left(D(E_c(m), E_s(m))\right) - FK(m)\right|^2\right], \tag{4.2}$$

where $FK(\cdot)$ computes the global positions of the joints based on the rotations and root positions, ensuring that the reconstructed motion maintains the correct skeletal structure.

To regularize the latent space and encourage the latent features to follow a standard normal distribution, we apply the Kullback-Leibler (KL) divergence loss $\mathcal{L}_{\text{KL}}$:

$$\mathcal{L}_{\text{KL}} = KL\left(E_c(m), |, \mathcal{N}(0,1)\right) + KL\left(E_s(m), |, \mathcal{N}(0,1)\right), \tag{4.3}$$

where $KL(\cdot, |, \cdot)$ denotes the KL divergence between the learned latent distributions and the standard normal distribution $\mathcal{N}(0,1)$.

Finally, the total loss for training the motion VAE network is defined as a weighted sum of the individual loss components:

$$\mathcal{L}_{\text{VAE}} = \lambda_1 \mathcal{L}_{\text{L1}} + \lambda_2 \mathcal{L}_{\text{fk}} + \lambda_3 \mathcal{L}_{\text{KL}}, \tag{4.4}$$

where λ_1, λ_2, and λ_3 are hyperparameters that balance the contributions of each loss component to the overall loss.

By optimizing this loss function, the network learns to accurately reconstruct input motions while structuring the latent space in a way that reflects motion styles and skill

levels. This structured latent space is crucial for the motion navigator to effectively guide users toward appropriate learning targets, facilitating personalized and progressive skill development.

4.2.2 Motion Style Transformer: Reflecting Motion Style onto Content Motion

To enable the network to better extract skill features from motion data, we propose a motion style transfer method designed to handle the complexities of diverse motion contents while preserving essential motion structures. Motion style transfer networks traditionally decompose input motions into content and style components. In this study, we build on this framework by conceptualizing *skill* as part of the motion style, capturing the distinctive ways players execute their movements.

Previous studies have demonstrated the effectiveness of transferring motion styles within the same motion category, such as walking or running [10, 74]. However, these methods encounter significant challenges when dealing with datasets that span diverse motion categories. Preserving content becomes increasingly difficult under such conditions, and this limitation restricts the generalizability of motion style transfer in real-world applications. Training separate networks for each motion category is inefficient and impractical, particularly in sports contexts, where performances often involve sequences of varied motions with distinct content.

Moreover, networks constrained to specific motion categories cannot capitalize on styles absent from their training datasets or those originating from other motion categories. This lack of adaptability diminishes their utility in broader contexts. To address these challenges, we propose a motion style transfer method capable of transferring styles across diverse motion contents while preserving the structural integrity of the content motion. The network incorporates a content-conditioned module that dynamically adapts the style transfer process to the content motion. By leveraging Variational Autoencoders (VAEs), the design facilitates balance between content and style constraints, enabling meaningful interpolation within the latent space and enhancing the system's adaptability.

One concern is whether our approach might conflate differences in *skill* with variations in body shape or skeleton structure, inadvertently creating an "intermediate" motion purely due to anatomical similarity. However, we treat skill level as a facet of the style dimension, whereas content represents the user's core motion patterns and body-specific factors. Thus, our latent-space representation does not assign "middle" motions simply based on a user's physical proportions; rather, we rely on explicit skill labels (e.g., beginner, intermediate, advanced) or skill-related features. We preserve biomechanical consistency (the content) while allowing the style transformation to adapt the movement's skillful characteristics to each user's motion constraints. In other words, the skill-based style is what the network modifies, ensuring that what changes is the user's perceived proficiency,

not merely their physique.

Network Design

The motion style transfer network, illustrated in Figure 4.6, is based on Variational Autoencoders (VAEs) and processes two motion inputs: a content motion and a style motion. Both inputs are represented using the robust 6D rotation representation [163]. To encode these inputs, the network employs a two-layer structure with skeletal convolution and pooling [9], capturing the essential features of both content and style.

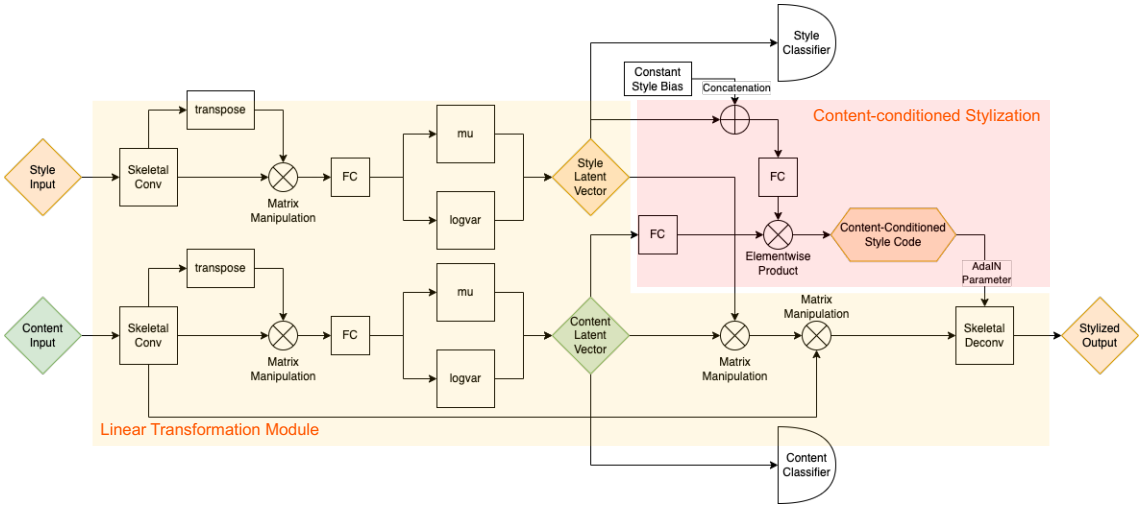

Figure 4.6: Content-preserving motion stylization network architecture.

The network design includes two distinct latent spaces: one for content and one for style. This separation ensures that the structural attributes of the motion (content) and its stylistic elements (style) can be independently processed and manipulated. However, transferring style between arbitrary motion pairs, particularly those from distinct categories, introduces challenges in maintaining content integrity. To address these challenges, the network incorporates mechanisms that enforce constraints on both content and style during training, ensuring that the stylized output retains the structure of the content motion while reflecting the desired stylistic attributes.

In practical applications, particularly in sports training, it is often desirable to allow users to control the intensity of the style or combine multiple styles to create novel motions. This flexibility is achieved by leveraging the latent spaces for direct manipulation, enabling applications such as personalized training regimens and the expansion of motion databases.

Linear Transformation Module

A critical challenge in motion style transfer is ensuring that the latent space represents skill progression in a structured and interpretable manner. To address this, we designed the Linear Transformation Module, as shown in Figure 4.7, to project nonlinear styles into a linear latent space. This transformation, which is inspired by a preivous image-to-image

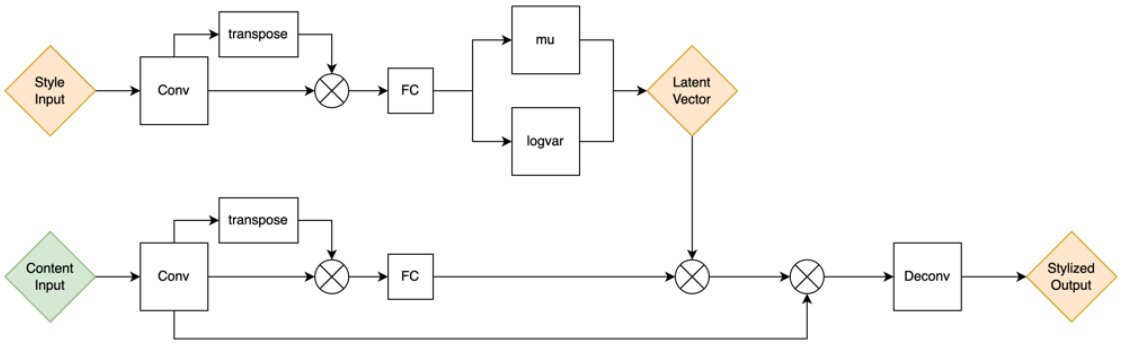

Figure 4.7: Linear transformation module.

work [93], facilitates smooth interpolation and ensures that transitions between different skill levels remain intuitive and consistent.

By introducing this module, the latent space becomes more interpretable, aligning skill progression with linear trajectories. For example, the transition from novice to intermediate or advanced skill levels follows a smooth, predictable path in the latent space. This property is particularly valuable in real-world applications, where gradual refinement of skill is a fundamental aspect of learning. The linear transformation also enhances usability by enabling controlled manipulation of styles, making it easier to fine-tune motion attributes according to specific training objectives.

Furthermore, the Linear Transformation Module ensures that transitions between styles maintain continuity, avoiding abrupt or unnatural changes that could disrupt the learning process. This smoothness within the latent space supports practical applications requiring incremental adjustments to motion characteristics, such as sports training scenarios where learners benefit from gradual improvements.

Content-Conditioned Style Encoder

Preserving the structural integrity of content motion during style transfer is critical, particularly when applying stylistic transformations across diverse motion categories. To achieve this, we introduce the Content-Conditioned Style Encoder, a module designed to adapt the style encoding process dynamically based on the content motion. This design, inspired by prior work in image-to-image translation [121], ensures that the stylized motion retains its original content structure while adopting the desired stylistic features.

The Content-Conditioned Style Encoder generates a style code informed by both the style and content inputs, as shown in Figure 4.8. By conditioning the style encoding on content features, the module dynamically adapts its output to align with the structural requirements of the content motion. This ensures that the resulting stylized motion preserves essential characteristics of the content, such as key postures and temporal dynamics, while seamlessly incorporating stylistic attributes.

This module addresses a common limitation of traditional style transfer methods,

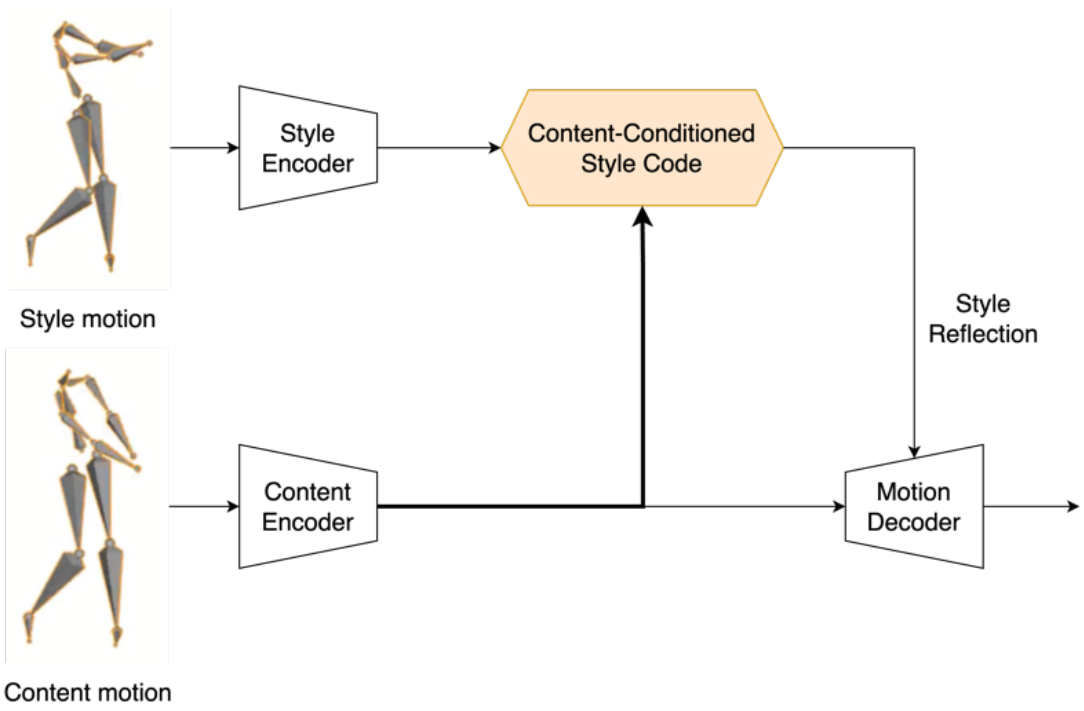

Figure 4.8: Content-conditioned style encoding.

which often treat content and style independently, leading to distortions or inconsistencies when transferring styles across motion categories. By integrating content features into the style encoding process, the Content-Conditioned Style Encoder ensures that stylistic transformations respect the structural constraints imposed by the content motion.

Additionally, this module enables finer control over the degree of stylization by balancing the contributions of content and style during the transfer process. This adaptability supports applications where precise adjustments to stylistic intensity are required, such as personalized training or experimental motion synthesis.

Network Losses

The network's losses include reconstruction loss and latent space loss, similar to those in the VAE network. However, we introduce an additional term, the triplet loss, to enhance the separation of styles and contents in the latent spaces. The triplet loss encourages the network to position similar styles or contents closer together while pushing dissimilar ones apart. Additionally, we use the Kullback-Leibler (KL) divergence loss to regularize the latent spaces, ensuring that the distributions approximate a standard normal distribution.

The triplet loss $\mathcal{L}_{\text{tri}}$ is defined as:

$$\mathcal{L}_{\text{tri}} = \mathbb{E}\left[\left|E_s(m) - E_s(m')\right| - \left|E_s(m) - E_s(m'')\right| + \delta\right], \tag{4.5}$$

where $E_s(\cdot)$ is the style encoder, m is the anchor motion, m' is a motion with a similar

style (positive example), m'' is a motion with a dissimilar style (negative example), and δ is a margin parameter. This loss encourages the style representations of similar motions to be closer than those of dissimilar motions by at least the margin δ.

Controlling the reflection of style in the output motion is crucial. We impose latent space constraints on the stylized output motion. After generating the stylized motion, we encode it back into the latent space and enforce its content and style representations to be similar to those of the input content and style motions, respectively. We define the latent space constraint loss $\mathcal{L}_{\text{lsc}}$ as:

$$\mathcal{L}_{\text{lsc}} = KL\left[E_c\left(D\left(E_c(m_c), E_s(m_s)\right)\right), |, E_c(m_c)\right] + KL\left[E_s\left(D\left(E_c(m_c), E_s(m_s)\right)\right), |, E_s(m_s)\right], \tag{4.6}$$

where $E_c(\cdot)$ is the content encoder, $D(\cdot,\cdot)$ is the decoder, m_c is the content motion, and m_s is the style motion.

Additionally, we introduce a cycle consistency loss to ensure the network maintains content and style information through transformations. After obtaining the stylized output, we use it as an input (either as content or style) and generate a new stylized output. This second output should reconstruct the original input motions. We compute the L1 loss and forward kinematics loss to enforce cycle consistency. The cycle consistency loss $\mathcal{L}_{\text{cc}}$ is defined as:

$$\mathcal{L}_{\text{cc}} = \mathbb{E}\left[|D\left(E_c\left(D\left(E_c(m_c), E_s(m_s)\right)\right), E_s(m_c)\right) - m_c| + |D\left(E_c(m_s), E_s\left(D\left(E_c(m_c), E_s(m_s)\right)\right)\right) - m_s|\right], \tag{4.7}$$

where the first term ensures that decoding the encoded stylized motion with the original content style recovers the original content motion, and the second term does the same for the style motion.

4.2.3 Motion Manipulator: Generating Intermediate Motions Between Styles

During self-training, beginners often find it challenging to imitate an ideal motion form that significantly differs from their current capabilities. To facilitate a smoother learning progression, we propose a motion manipulator that generates intermediate motions between the user's current motion and the target motion. By providing these transitional motions, we aim to bridge the gap and make the learning process more attainable.

As previously discussed, our network learns a latent space that reflects motion style similarity, where each point in this high-dimensional space represents a specific motion style. We utilize a VAE-based network structure to project input motions into this latent space, ensuring that the latent distribution is smooth and linear. This property allows us to

blend different styles through linear interpolation, effectively generating new intermediate motions that are not present in the original dataset.

The core idea is to create an intermediate motion from an interpolated latent vector. Specifically, we compute a new latent vector $\boldsymbol{L}_{\text{inter}}$ by linearly interpolating between the user's latent vector $\boldsymbol{L_X}$ and the learning target's latent vector $\boldsymbol{L_T}$:

$$\boldsymbol{L}_{\text{inter}} = (1 - \alpha) \times \boldsymbol{L_X} + \alpha \times \boldsymbol{L_T}, \tag{4.8}$$

where $\boldsymbol{L_X}$ is the latent representation of the user's current motion, $\boldsymbol{L_T}$ is the latent representation of the target motion selected by the motion navigator, and $\alpha \in [0, 1]$ is the interpolation parameter that controls the influence of each latent vector. By adjusting the value of α, we can control the degree to which the intermediate motion resembles the target motion. For example, setting $\alpha = 0$ yields the user's original motion, while $\alpha = 1$ yields the target motion.

By increasing the value of α, the intermediate latent vector $\boldsymbol{L}_{\text{inter}}$ moves closer to $\boldsymbol{L_T}$ in the latent space. The motion style transformer then decodes $\boldsymbol{L}_{\text{inter}}$ to generate the intermediate motion $\boldsymbol{Y}$. Ideally, as α increases, the reconstructed motion $\boldsymbol{Y}$ progressively incorporates more features of the target motion, allowing the user to practice motions that are incrementally more advanced.

This approach provides users with personalized intermediate motions that are tailored to their current skill level. By practicing these intermediate motions, users can gradually adjust their movements, making the learning process less overwhelming and more effective. The ability to generate motions that smoothly transition between the user's current form and the desired form enhances motivation by offering achievable steps toward improvement.

In summary, the motion manipulator leverages the properties of the latent space learned by the VAE-based network to create intermediate motions through linear interpolation. This method empowers users to engage in a progressive learning experience, facilitating skill acquisition by providing attainable motion targets that bridge the gap between their current abilities and their goals.

4.3 Evaluation

4.3.1 Experimental Setup

To assess the accuracy and effectiveness of the three modules introduced in the previous section—motion navigator, motion style transformer, and motion manipulator—we conducted experiments using precise golf swing motion data collected in a controlled laboratory environment. We implemented two models of the network and performed statistical analyses under four different training conditions:

- VAE without skill-level labels
- VAE-MST without skill-level labels
- VAE with skill-level supervision
- VAE-VST with skill-level supervision

To give supervision to the newtowk, we replace the style classifier shown in Figure 4.6 with a simple regressor predicting the level of the swing motion. Note that for the condition of VAE without skill-level supervision, we still utilize the labels of individuals for the triplet loss. The following sections detail the data collection process, the defination of skill level, and the evaluation metrics employed in this study.

4.3.2 Dataset

Previous works suggest that using 3D human pose data instead of videos can enhance the precision of motion analysis and facilitate practical implementations. However, relying on skeleton data estimated from videos has several drawbacks:

- The accuracy of synchronization is heavily dependent on the precision of human pose estimation algorithms.
- High-precision human pose estimation typically requires significant computational resources and longer inference times.

These limitations can lead to unstable application performance and hinder real-time capabilities. To address these challenges, we utilized a high-quality motion capture system to provide precise human poses for fine-grained motion analysis.

Off-the-shelf optical motion capture systems, equipped with high-frame-rate cameras, can record human motions with high accuracy even during rapid movements like golf swings (see Figure 4.11). By employing such a system, we obtained highly accurate human pose data with minimal processing delays, enabling the possibility of real-time functionality in our system since the proposed networks are computationally lightweight.

For this study, we created a new dataset of diverse golf swing motions, captured using an OptiTrack motion capture system. Recognizing the impact of equipment in sports performance, we also recorded the motion data of the golf club during swings. Additionally, to facilitate future development and evaluation, we used a golf simulator to record ball trajectories as outcomes of the swings. The following sections elaborate on the data collection process.

Environment Setup

Our new golf swing dataset includes three types of data:

- **Swing Motion**: The detailed movements of the golfer's body during the swing.
- **Golf Club Motion**: The precise trajectory and orientation of the golf club throughout the swing.
- **Swing Outcome**: The resulting ball trajectory and performance metrics captured by a simulator.

To collect the swing and golf club motions, we used an OptiTrack optical motion capture system. As shown in Figure 4.9, we set up a motion capture studio equipped with 12 high-speed cameras to ensure comprehensive coverage and accuracy.

Figure 4.9: Motion capture studio setup with 12 high-speed cameras.

For tracking human motion, we employed the "Conventional" marker set provided by OptiTrack, which generally follows the Helen Hayes marker placement style (Figure 4.10). We carefully calibrated the marker setup for each participant, as individual body proportions vary. This meticulous calibration allowed the motion capture system to accurately map the attached markers to a virtual skeletal model representing the participant's pose (Figure 4.11).

Data Collection

To standardize the dataset and minimize individual biases, we instructed all participants to use the same golf club—a 7-iron provided by the research team. Additionally, we

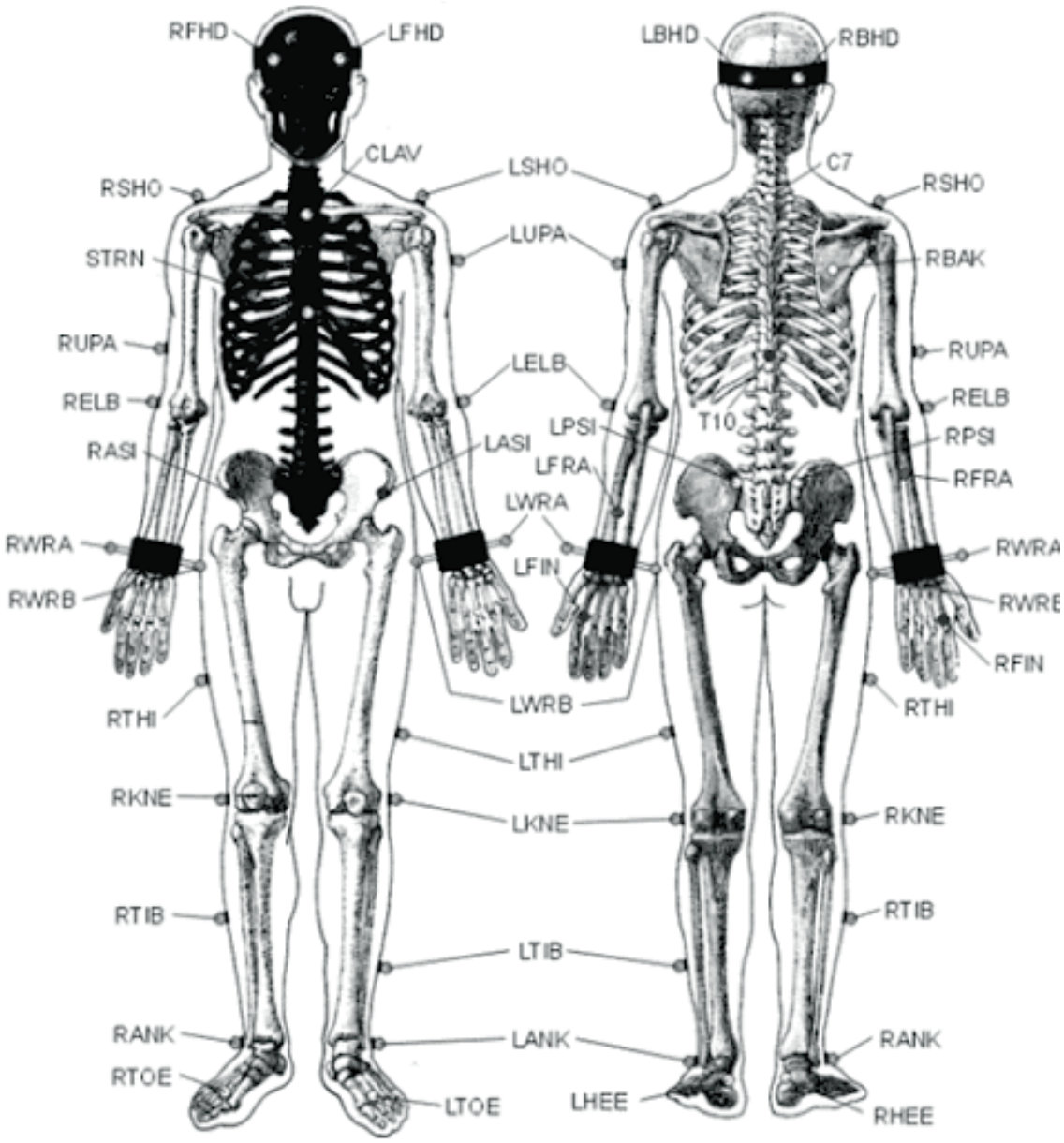

Figure 4.10: Helen Hayes style marker placement for motion capture. [3]

maintained consistent conditions throughout the data collection process:

- A golf ball dispenser was used to place the ball at a fixed position for each swing.
- A yellow tape was affixed to the ground to indicate the target line, ensuring consistent aiming direction.
- Participants were asked to stand at a designated spot, positioning their bodies perpendicular to the target line.

During the data collection sessions, each participant performed 50 swings. To prevent fatigue and maintain performance consistency, participants took at least a 5-minute break after every 10 swings.

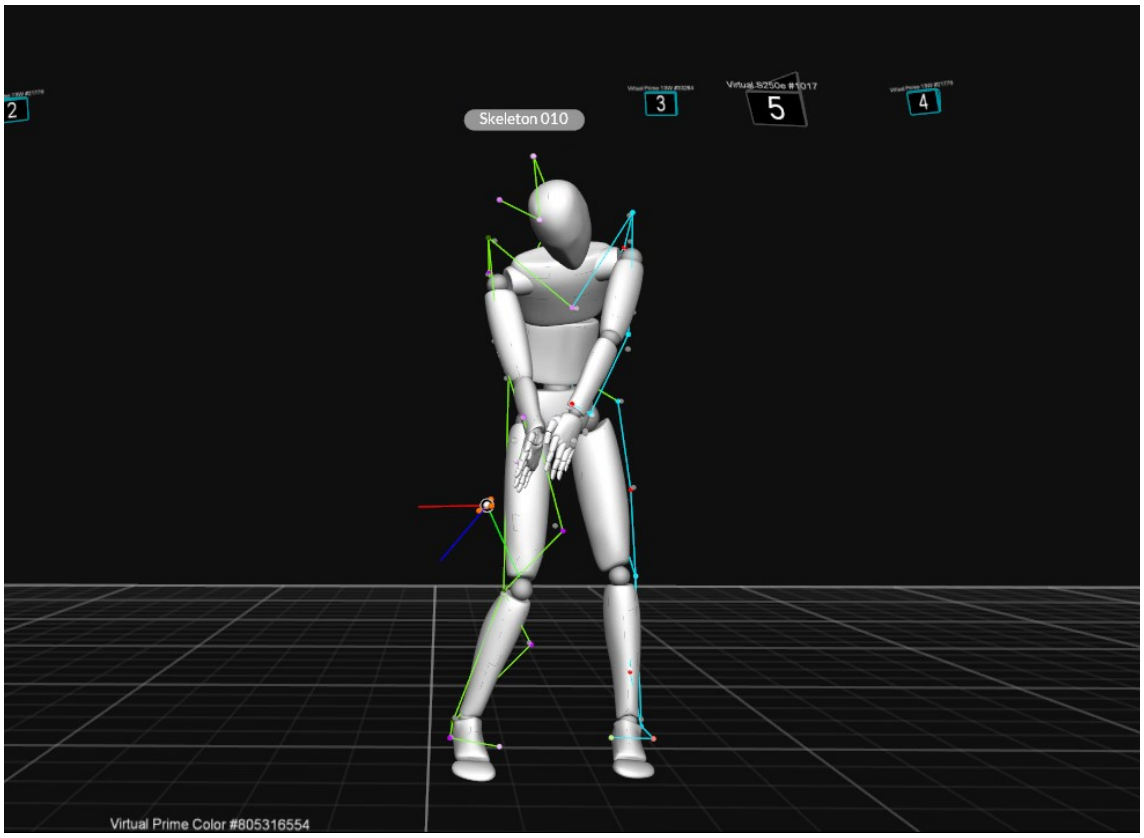

Figure 4.11: High-quality human motion capture using an optical motion capture system.

GolfMDB: Golf Swing Motion Capture Dataset

For the purposes of this study, particularly in clustering different motion styles corresponding to various skill levels, we recruited amateur golfers across three levels: beginner, intermediate, and advanced. A common metric for identifying an amateur golfer's skill level is their average round score—the total strokes taken across all holes in a standard course.

We defined the skill levels as follows:

- **Beginners**: Individuals with no experience playing on a golf course and thus no recorded scores.
- **Intermediate Players**: Golfers whose average round scores are over 100 strokes.
- **Advanced Players**: Golfers whose average round scores are under 100 strokes.

According to data from the United States Golf Association (USGA), the average score for recreational golfers on a par-72 course is approximately 91 strokes, with about 50% of golfers typically breaking 100. We used a score of 100 as a benchmark to distinguish intermediate players from advanced players, as it signifies significant progress and improvement in one's golf game.

In the data collection phase, we gathered motion data from 10 individuals:

- 3 advanced golfers
- 4 intermediate players
- 3 beginners

Notably, there were no participants with scores between 90 and 100, providing clear distinctions between the skill levels. Table 4.1 summarizes the dataset characteristics.

Table 4.1: GolfMDB: Golf Swing Motion Capture Dataset Summary

Category	**Advanced**	**Intermediate**	**Beginners**
Number of Participants	3	4	3
Average Round Score	$70 \sim 90$	$100 \sim 120$	N/A
Swings per Person	50		
Frames per Second (FPS)	240		

Using the 240 FPS motion capture system, we collected over 500 frames per swing, resulting in a total of more than 250,000 motion data frames ($500 \times 50 \times 10$). We meticulously cleaned and annotated the raw data to ensure its quality. Our intention is to make this dataset publicly available to promote reproducibility and encourage future research in this area.

4.3.3 Evaluation Metrics

Our objective was to determine whether the network effectively learned a latent space where similar motion styles are clustered closely together, particularly reflecting different skill levels. We aimed to encourage the network to capture skill-level features and form distinct clusters for beginners, intermediate players, and advanced players in the latent space.

To assess this, we trained both the VAE and VAE-MST networks with and without incorporating skill-level labels from the dataset. We then calculated the relationships among the skill levels in the latent space using the following evaluation metrics:

L2 Distance

The L2 distance d_{ij} measures the average Euclidean distance between the mean latent vectors of two groups:

$$d_{ij} = \left| \bar{\mathbb{L}}_i - \bar{\mathbb{L}}_j \right|, \tag{4.9}$$

where $\bar{\mathbb{L}}_i$ and $\bar{\mathbb{L}}_j$ are the mean latent vectors for groups i and j, respectively. This metric quantifies how far apart different skill-level clusters are in the latent space.

Cosine Similarity

We employed cosine similarity to quantify the alignment of vectors representing transitions between skill levels in the latent space. The cosine similarity between two vectors is calculated as:

$$\text{Cos}(\vec{IJ}, \vec{IK}) = \frac{\boldsymbol{L}ij \cdot \boldsymbol{L}ik}{|\boldsymbol{L}ij|\,|\boldsymbol{L}ik|}, \tag{4.10}$$

where:

$$\boldsymbol{L}_{ij} = \bar{\mathbb{L}}_i - \bar{\mathbb{L}}_j. \tag{4.11}$$

Specifically, we calculated cosine similarities between the vectors representing:

- Transition from beginners to advanced players ($\vec{BA}$)
- Transition from beginners to intermediate players ($\vec{BI}$)
- Transition from intermediate to advanced players ($\vec{IA}$)

High cosine similarity between these vectors indicates that they are aligned closely in the latent space, suggesting a consistent direction of skill progression. Such alignment is desirable for our recommendation system, as it implies that intermediate-level motions are positioned appropriately between beginner and advanced motions, facilitating structured and progressive skill development.

Pearson's Correlation

In statistics, the Pearson correlation coefficient ρ is a measure of linear correlation between two sets of data. In our experiment, we compare the distance in the latent space with the mean per joint point error (MPJPE), and the definition for ρ is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{4.12}$$

where:

- cov is the covariance
- X is the distance in the latent space
- Y is the MPJPE
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

4.3.4 Results

Quantitative Results

Table 4.2 presents the cosine similarity, L2 distance, and Pearson's correlation coefficients for the four models: VAE and VAE-MST, each trained with and without skill-level labels.

Table 4.2: Cosine Similarity, L2 Distance, and Pearson's Correlation Between Skill Levels and MPJPE in Latent Space

Model	**Cosine Similarity**		**L2 Distance**			**Pearson's Correlation**
	$\text{Cos}(\vec{BA}, \vec{BI})$	$\text{Cos}(\vec{BA}, \vec{IA})$	d_{BI}	d_{IA}	d_{BA}	
VAE (without skill-level supervision)	0.609	0.647	1.658	1.724	2.126	0.522
VAE (with skill-level supervision)	0.578	0.628	2.464	2.584	3.048	0.641
VAE-MST (without skill-level supervision)	0.329	0.879	2.099	4.157	4.345	0.484
VAE-MST (with skill-level supervision)	**0.885**	**0.906**	2.334	2.562	4.390	**0.857**

The experimental results reveal distinct differences among the four models:

VAE (without skill-level supervision):

- Achieved moderate cosine similarities of 0.609 and 0.647.
- Indicates some basic alignment in the latent space but lacks a strong progression between skill levels.
- L2 distances are relatively small, suggesting clusters are not well separated.
- Pearson's correlation of 0.522 indicates a moderate positive relationship between latent space distances and MPJPE.

VAE (with skill-level supervision):

- Cosine similarities slightly decreased to 0.578 and 0.628.
- Adding skill-level labels did not significantly enhance the representation of skill transitions.
- L2 distances increased, but the model still lacks a structured progression suitable for recommendations.
- Pearson's correlation improved to 0.641, showing a stronger relationship between latent space distances and MPJPE compared to the unsupervised VAE.

VAE-MST (without skill-level supervision):

- Cosine similarity between $\vec{BA}$ and $\vec{BI}$ dropped to 0.329, indicating poor alignment.
- However, the similarity between $\vec{BA}$ and $\vec{IA}$ increased significantly to 0.879.

- Suggests that the model effectively captures the extremes (beginner and advanced) but struggles with the intermediate level without supervision.
- L2 distances are larger, especially between intermediate and advanced, indicating better separation.
- Pearson's correlation of 0.484 indicates a weak positive relationship, reflecting inconsistent alignment between latent space distances and MPJPE.

VAE-MST (with skill-level supervision):

- Achieved the highest cosine similarities of 0.885 and 0.906.
- Demonstrates a well-aligned latent space that reflects a consistent, linear progression between skill levels.
- L2 distances indicate that clusters are appropriately spaced, with intermediate motions positioned meaningfully between beginner and advanced.
- Pearson's correlation of 0.857 signifies a very strong positive relationship between latent space distances and MPJPE, underscoring the model's effectiveness in accurately capturing skill discrepancies.

PCA-based Latent Space Visualization

To further analyze how each model arranges the motion data in its latent space, we conducted Principal Component Analysis (PCA) on the learned embeddings for all four models. For each model, we project the latent vectors onto two principal components and plot two versions of the scatter diagram:

1. **Color-coded by participant**, which highlights individual differences (e.g., personal style).
2. **Color-coded by skill level**, which reveals whether the model organizes the space in a way that reflects a progression from beginner to intermediate to advanced.

Figures 4.12–4.15 present these PCA plots, and we discuss their implications below.

VAE (without skill-level supervision). Figure 4.12 illustrates how the standard VAE (trained without skill-level labels) distributes motions in the latent space.

- **Participant-based coloring**: Points from the same participant show some clustering but also partial overlap with other participants. This suggests that personal nuances are partially captured, but the clusters are not sharply differentiated.

- **Skill-level coloring**: We see moderate grouping by skill (beginner, intermediate, advanced), but there is considerable overlap, especially between beginner and intermediate. This aligns with the moderate metrics (e.g., cosine similarities around 0.60 and Pearson's correlation 0.52), indicating that the model partially encodes skill differences but does not form a clear progression.

VAE (with skill-level supervision). As shown in Figure 4.13, adding skill-level labels during training yields a latent space that is somewhat more structured, yet still not strongly segregated by skill.

- **Participant-based coloring**: Similar to the unsupervised VAE, points from the same participant group together to a certain degree, but the separation is still moderate.

- **Skill-level coloring**: Compared to the unsupervised version, there are slightly more defined clusters corresponding to each skill category. However, some intermediate points blend into either beginner or advanced clusters, suggesting the skill dimension is not fully learned. These observations match the improved but still modest quantitative scores (e.g., Pearson's correlation of 0.64).

VAE-MST (without skill-level supervision). Figure 4.14 shows the unsupervised version of VAE-MST.

- **Participant-based coloring**: The model distinctly separates certain participants, reflecting that it captures individual differences to a strong degree. However, some participants' points are scattered, indicating inconsistent clustering.

- **Skill-level coloring**: We notice good separation for advanced and beginner motions, but intermediate motions often cluster unevenly or overlap with the extremes. This phenomenon is consistent with the quantitative results, where one skill-pair shows high cosine similarity while the other is quite low, leading to an overall weaker correlation.

In essence, the unsupervised VAE-MST excels at pushing extremes apart (beginner vs. advanced) but does not systematically position intermediate motions.

VAE-MST (with skill-level supervision). Finally, Figure 4.15 portrays the latent space for our best-performing model, VAE-MST trained with skill-level labels.

- **Participant-based coloring**: Individual participants are reasonably grouped, and the arrangement shows fewer overlaps than with the other models. This indicates that the architecture preserves personal styles while providing a coherent global structure.

- **Skill-level coloring**: Beginners, intermediates, and advanced motions each form relatively distinct clusters, with intermediate clusters bridging the gap between beginner and advanced. This clear progression mirrors the strong cosine similarities (0.885 and 0.906) and high Pearson's correlation (0.857), underscoring the model's ability to encode skill transitions in a smoothly graduated manner.

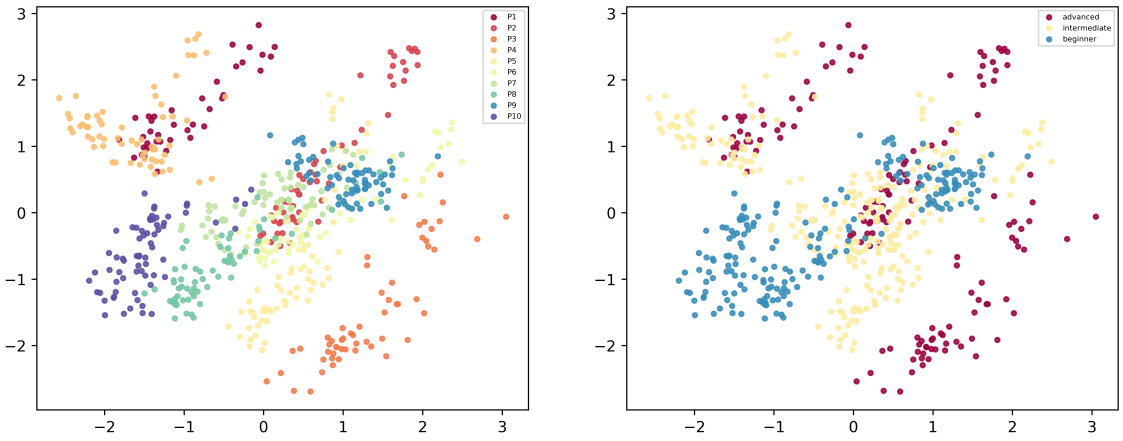

Figure 4.12: PCA of the latent space for **VAE (without skill-level supervision)**. Left: color-coded by participant. Right: color-coded by skill level. Note the partial overlap in skill clusters.

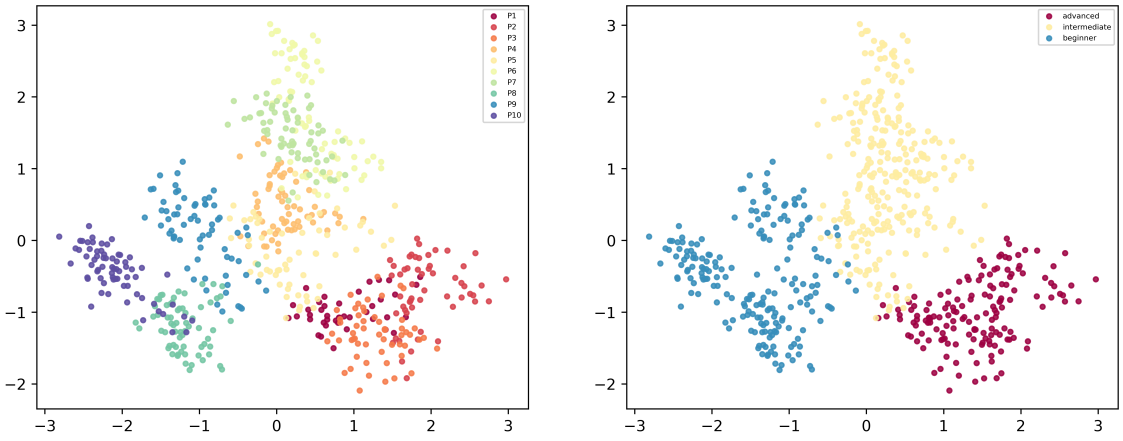

Figure 4.13: PCA of the latent space for **VAE (with skill-level supervision)**. Left: color-coded by participant. Right: color-coded by skill level. Clusters show slightly better skill separation than in the unsupervised case, but still overlap significantly.

Across the four models, the PCA visualizations offer a qualitative lens to corroborate the quantitative metrics (cosine similarity, L2 distance, Pearson's correlation) reported in Table 4.2. The standard VAE, whether unsupervised or labeled, exhibits moderate grouping by skill but lacks a clear progression. Meanwhile, the unsupervised VAE-MST effectively pushes extremes apart but does not anchor intermediate skills in a structured continuum. In contrast, the VAE-MST model trained with labels forms distinct and smoothly transitioning clusters for each skill category. These findings reinforce our con-

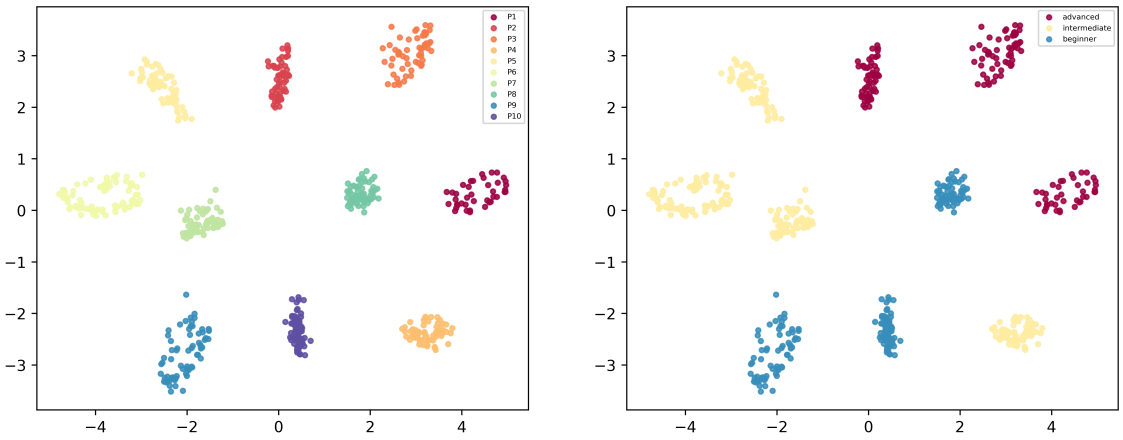

Figure 4.14: PCA of the latent space for **VAE-MST (without skill-level supervision)**. Left: color-coded by participant. Right: color-coded by skill level. The model separates extremes of skill but struggles with the intermediate level.

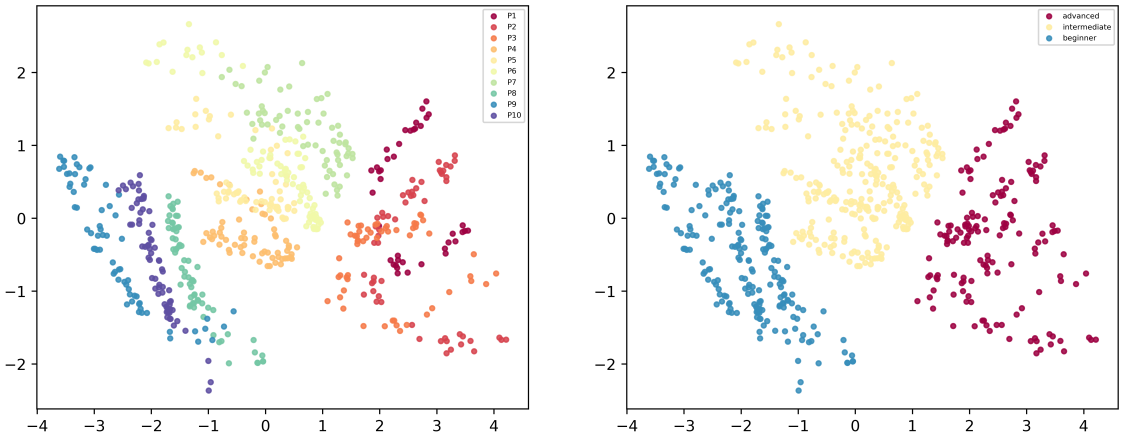

Figure 4.15: PCA of the latent space for **VAE-MST (with skill-level supervision)**. Left: color-coded by participant. Right: color-coded by skill level. This model yields distinct and smoothly transitioning clusters, aligning with the high cosine similarities and strong Pearson's correlation.

clusion that *both the MST architecture and skill-level supervision are crucial* for a well-aligned, progression-focused latent space.

Comparison Among Different Modules (Ablation Study)

To further investigate the contributions of different components in our model, we conducted an ablation study. In this study, we evaluated the performance of our model by removing key modules: the Linear Transformation Module (LTM) and the Content-Conditioned Stylization Module (COCO). The results of this comparison are shown in Table 4.3, where we report the cosine similarity and Pearson's correlation for each model variant.

The model trained without both LTM and COCO achieved cosine similarities of 0.840 and 0.864 for $\text{Cos}(\vec{BA}, \vec{BI})$ and $\text{Cos}(\vec{BA}, \vec{IA})$, respectively, along with a Pearson's correla-

Table 4.3: Ablation study. LTM stands for linear transformation module. COCO stands for content-conditioned stylization module.

Model (Type)	**Cosine Similarity**		**Pearson's Correlation**
	$\text{Cos}(\vec{BA}, \vec{BI})$	$\text{Cos}(\vec{BA}, \vec{IA})$	
Ours w/o LTM & COCO	0.840	0.864	0.826
Ours w/o LTM	0.819	0.860	0.826
Ours w/o COCO	**0.885**	**0.909**	0.840
Ours	**0.885**	0.906	**0.857**

tion of 0.826. These results suggest that while the model captures basic alignment between skill levels, the latent space lacks the structure necessary for a clear skill progression.

When the LTM module is removed, the model's performance drops further, with cosine similarities of 0.819 and 0.860, and the same Pearson's correlation of 0.826 as the model without both modules. This indicates that the LTM plays a significant role in organizing the latent space to better represent the relationships between skill levels, but its absence does not drastically hinder the model's overall alignment.

On the other hand, when the COCO module is removed, the model performs significantly better, achieving cosine similarities of 0.885 and 0.909, as well as a Pearson's correlation of 0.840. This indicates that the LTM contributes more to the overall alignment of the latent space, particularly in the separation of skill levels, which enhances the model's ability to represent the skill progression accurately.

The full model, with both LTM and COCO modules, achieves the highest overall performance, with cosine similarities of 0.885 and 0.906, and a Pearson's correlation of 0.857. These results clearly demonstrate that the combination of both modules produces the best performance, as it aligns the latent space effectively while capturing the smooth progression between skill levels.

Comparison Among Latent Dimensionalities

We conducted additional experiments to assess how varying the dimensionality of the latent space impacts the model's ability to represent skill-level progression. Table 4.4 summarizes the results for four configurations: $1/4x$, $1/2x$, $1x$, and $2x$ of the baseline dimensionality (where $1x$ corresponds to the original setting from previous work, i.e., 132 dimensions). The cosine similarities and Pearson's correlation reflect how effectively each latent dimension setting captures the relative positions of beginner, intermediate, and advanced skill levels.

As the table shows, using $1x$ dimensionality yields the strongest overall performance, with cosine similarities of 0.885 and 0.906, and a Pearson's correlation of 0.857. This indicates that our baseline dimensionality is well-suited for separating beginner, intermediate, and advanced motions in a manner that aligns with motion discrepancies. Reducing the latent space to $1/4x$ or $1/2x$ still produces decent results (cosine similarities above 0.88),

Table 4.4: Ablation study on latent dimensionalities. The factor $1x$ indicates the baseline dimension of 132, while $1/4x$ and $1/2x$ reduce this dimensionality, and $2x$ doubles it.

Dimension Setting	**Cosine Similarity**		**Pearson's Correlation**
	$\text{Cos}(\vec{BA}, \vec{BI})$	$\text{Cos}(\vec{BA}, \vec{IA})$	
$1/4x$	0.883	0.898	0.822
$1/2x$	0.881	0.902	0.840
$1x$	0.885	**0.906**	**0.857**
$2x$	**0.887**	0.896	0.770

but we observe lower correlation values (0.822 and 0.840, respectively), suggesting that fewer dimensions make it harder for the model to preserve the fine-grained distinctions necessary to correlate distance in latent space with actual motion differences. Conversely, increasing the dimensionality to $2x$ slightly improves one cosine similarity measure (0.887) but noticeably reduces the Pearson's correlation (0.770), implying that an excessively large latent space may introduce noise or redundancy that weakens the overall skill-level alignment.

These findings demonstrate that our baseline dimensionality strikes a favorable balance between capturing subtle skill variations and maintaining robust correlation with motion discrepancies. Although smaller or larger latent spaces do not drastically degrade the model's performance, the $1x$ setting provides the most consistent and well-structured representation of skill-level progression in the latent space.

4.3.5 Discussion

The quantitative results (Table 4.2) confirm that the VAE-MST model trained with skill-level labels surpasses the other models in capturing a clear, structured progression between skill levels. The high cosine similarities of 0.885 and 0.906, together with a Pearson's correlation of 0.857, indicate a well-aligned latent space that methodically positions beginners, intermediates, and advanced learners in an orderly continuum. This property is crucial for our recommendation system, which seeks to guide novice users step by step toward advanced proficiency.

Examining the extended PCA visualizations for each model offers further insight into these quantitative findings. In the case of the standard VAE (unsupervised), the PCA plots display partial clustering by skill, but there is noticeable overlap between beginner and intermediate categories. Although personal nuances are evident, the model's latent space does not present a clear progression path, aligning with moderate cosine similarities (around 0.60) and relatively low Pearson's correlation values. When skill-level labels are introduced to the standard VAE, slight improvements emerge in its PCA plots, yet many intermediate points still blur with beginner or advanced motions. This outcome is reflected in the somewhat higher (but still limited) correlation of 0.64, indicating that the standard VAE's representation of intermediate skill remains incomplete even under supervision.

In contrast, the unsupervised VAE-MST exhibits a stronger separation of extremes, with beginner and advanced clusters clearly distinguishable in PCA space, although intermediate motions do not consistently occupy a well-defined zone between these extremes. This situation is reflected in the model's mixed quantitative scores, where one skill pair shows high alignment while another pair remains poorly aligned. Without explicit labels, the unsupervised VAE-MST effectively isolates skill extremes but struggles to place intermediate motions in a meaningful continuum. The most compelling evidence for a well-organized progression appears in the PCA plots for the VAE-MST trained with labels, which display neatly separated clusters for beginner, intermediate, and advanced levels, along with smoother transitions between them. This arrangement is perfectly in keeping with the model's high cosine similarities (0.885 and 0.906) and strong Pearson's correlation (0.857), suggesting that a linear skill progression is successfully embedded in the latent space.

Notably, the PCA visualizations also reveal interesting per-participant differences, demonstrating how each user retains some unique traits while still aligning their skill levels within the broader distribution. This balance between individual variation and skill-based structure enhances the capacity for personalized coaching, as it indicates that the VAE-MST (with labels) can simultaneously reflect individual styles and neatly position intermediate skill levels. Such an arrangement provides practical "stepping stones" that facilitate incremental improvements, especially when integrating pathfinding algorithms such as Dijkstra's to recommend viable skill-development pathways.

Overall, both the qualitative PCA analyses and the quantitative metrics converge on the same conclusion: while the standard VAE architecture fails to capture intermediate skills robustly, and the unsupervised VAE-MST excels mostly at isolating extremes, the VAE-MST model with skill-level labels successfully forms a cohesive, progression-oriented latent space. This model not only separates beginners, intermediates, and advanced learners into coherent clusters, but also places intermediate motions between the other two levels in a manner conducive to user-friendly and data-driven recommendations.

Based on these findings, we selected the VAE-MST model with skill-level supervision for our recommendation system. Its ability to accurately represent the progression between skill levels in the latent space makes it an ideal choice for guiding beginners through incremental learning steps toward advanced proficiency. By leveraging this structured latent space, we can effectively apply algorithms like Dijkstra's algorithm to find optimal paths for skill improvement, providing users with personalized and progressive training experiences.

4.4 Coach Navi

To verify the effectiveness of the proposed system in helping users learn a golf swing in real-world scenarios, we developed a graphical user interface (GUI) application called **Coach Navi** and conducted a user study with beginner golfers. Coach Navi comprises three main modules: the motion navigator, the motion style transformer, and the motion visualizer.

First, the system receives the user's motion input $\boldsymbol{X}$ and a set of candidate motions from the database, denoted as $\mathbb{C} = C_i \mid i = 1, 2, \ldots, n$. The encoder $\boldsymbol{E}$ embeds these input motions into the latent space, resulting in latent representations $\boldsymbol{L_X}$ for the user's motion and $\mathbb{L}\boldsymbol{C} = LC_i \mid i = 1, 2, \ldots, n$ for the candidate motions. The encoder is trained to learn a latent space where similar motions are positioned close to each other, effectively capturing the underlying motion similarities.

Next, utilizing the learned latent space, the motion navigator $\boldsymbol{MN}$ selects an optimal learning target $\boldsymbol{L_T}$ by measuring the Euclidean distances between the user's latent vector $\boldsymbol{L_X}$ and the candidate latent vectors $\mathbb{L}_{\boldsymbol{C}}$. This process identifies the most suitable motion for the user to learn from, based on proximity in the latent space. The system then employs the motion style transformer $\boldsymbol{MST}$ to decode $\boldsymbol{L_T}$ along with $\boldsymbol{L_X}$, generating a new motion $\boldsymbol{Y}$ that possesses the desired motion style. This motion $\boldsymbol{Y}$ serves as a personalized advanced skill demonstration for the user, rendered with their own body appearance.

Finally, using the outputs from the motion navigator and motion style transformer, the motion visualizer renders the stylized motion $\boldsymbol{Y}$ using a 3D textured avatar that reflects the user's appearance. This personalized visualization aims to enhance the user's ability to imagine and replicate the desired movements during training.

4.4.1 Motion Navigator and Motion Style Transformer

For the motion navigator and motion style transformer modules, we reused the pipeline introduced in Section 4. To determine the most effective configuration for Coach Navi, we analyzed the accuracy and effectiveness of these modules.

Our experimental results demonstrated that the VAE-MST (Variational Autoencoder with Motion Style Transfer) model trained with skill-level labels is the most effective for structuring skill progression within the latent space. The high cosine similarities achieved by this model indicate a well-organized latent space where different skill levels are aligned in a progressive and linear fashion. This alignment makes it ideal for our recommendation system, as it allows us to guide beginner users through intermediate learning targets, gradually advancing toward more advanced skill levels.

By selecting the VAE-MST with labels, our system can provide personalized, step-by-step recommendations that facilitate skill development. This approach aligns with our overarching goal of supporting and enhancing the learning experience for beginner golfers.

As a result, we chose the VAE-MST with labels as the final version of the network for our application implementation.

4.4.2 Motion Visualizer

Building upon the results of the previous modules, we implemented the motion visualizer to provide users with visual feedback that reflects their own appearance. We recognized that it can be challenging for beginners to perceive the differences between their own form and that of advanced golfers, especially when the skill gap is significant. Therefore, the motion visualizer displays a learning target motion that is at a skill level closer to the user's current ability, making it more attainable and easier to imitate.

Moreover, beginners may find it difficult to imagine the desired motion state if the learning target differs greatly in terms of body shape and appearance. To address this issue, the motion visualizer generates a 3D model that shares the same appearance as the user, aiming to help users envision the "ideal me" during training.

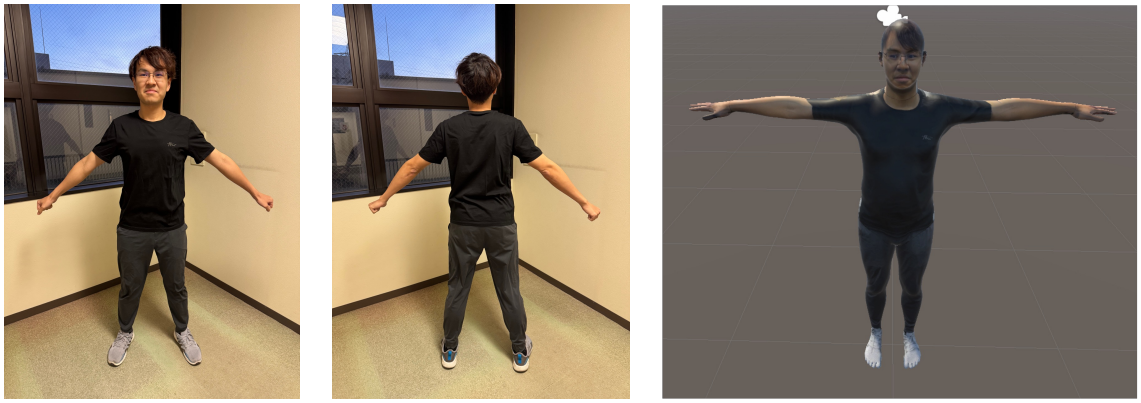

Figure 4.16: Creation of the 3D SMPL avatar with the user's appearance.

To create a 3D avatar of the user, we utilized the widely used SMPL (Skinned Multi-Person Linear) model [94], which is a parametric human body model that can be easily animated using motion capture data. We first employed OpenPos [27] to detect 2D keypoints from the user's image. These 2D detections were then used in conjunction with the method proposed by Pavlakos et al. [110] (commonly referred to as SMPLify-X) to estimate the SMPL parameters from a single image. By iteratively refining both shape and pose parameters, as well as the camera parameters, SMPLify-X aligns the projected 3D joints of the SMPL mesh with the 2D keypoints from OpenPose. Using the parameters thus obtained, we created a controllable 3D model that closely matches the user's body shape.

Next, to enhance the realism of the avatar and help users identify with it, we prepared a texture map of the user to be applied to the 3D SMPL model. To create a 2D UV texture map, we captured two images of the user—one from the front and one from the back (see Figure 4.16). We then projected the generated SMPL mesh onto these two images to

extract color information. Specifically, we relied on the standard SMPL "template OBJ," which encodes per-vertex UV coordinates (u_i, v_i) in $[0, 1]^2$. Because these UV coordinates are consistent across all SMPL models, any newly fitted shape can be mapped onto a standardized 2D texture space. To populate this space with color, we performed reverse rasterization:

1. **Triangle Identification**: We begin with a blank texture of size $N \times N$ (e.g., 1000×1000) aligned to the UV layout. For each pixel (X, Y), we identify which SMPL triangle in UV space covers that pixel.

2. **Barycentric Coordinates**: We compute barycentric coordinates for (X, Y) within the identified triangle.

3. **Projection to 2D Image**: We apply those coordinates to the *projected* triangle in the user's 2D photograph, obtaining a corresponding pixel (x, y).

4. **Color Sampling**: We read the color at (x, y) in the photograph and store it in UVTexture$[Y, X]$.

Figure 4.17: Creation of UV colored texture.

By iterating over every pixel in the UV texture, we effectively "pull" color from each 2D photo into the SMPL UV layout (Figure 4.17) Repeating the process for both front and back images allows us to cover more of the user's surface and merge occluded or partially visible regions from multiple viewpoints. Nevertheless, some areas may remain unfilled if they are off-camera or obscured in all views. To address these gaps, we apply a *Breadth-First Search* (BFS) color diffusion strategy:

1. **Boundary Pixels**: We maintain a queue of boundary pixels that have valid color but neighbor at least one unfilled pixel.

2. **Neighbor Check**: For each boundary pixel (x, y) removed from the queue, we examine adjacent pixels (nx, ny). If (nx, ny) is unfilled but in the same connected region, we enqueue (nx, ny) and accumulate color from (x, y) into `texture_sum`$[nx, ny]$, incrementing `texture_count`$[nx, ny]$.

3. **Averaging**: Once (nx, ny) has enough contributions, we assign

$$\texttt{texture}[nx, ny] \;=\; \frac{\texttt{texture\_sum}[nx, ny]}{\texttt{texture\_count}[nx, ny]},$$

thus filling that pixel with an averaged color from its filled neighbors.

Through this BFS-based inpainting, small holes or missing regions in the texture are smoothed over, ensuring a continuous, visually coherent surface. By inversely mapping the textures from the images to the standard SMPL UV space, we ultimately obtain a 3D model that closely resembles the user's appearance. The final result of the textured SMPL avatar is shown in Figure 4.16.

4.4.3 Application

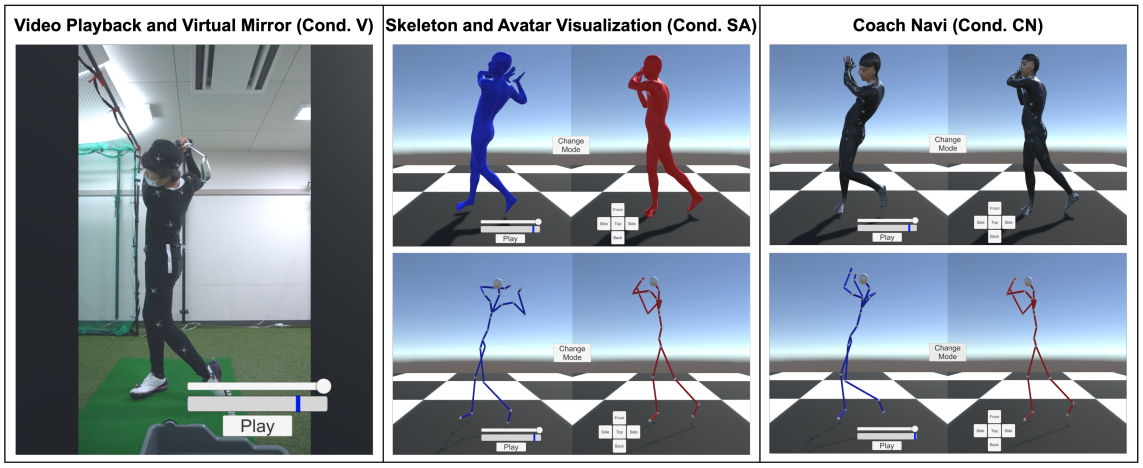

Figure 4.18: Overview of the three different systems used in the user study. A webcam records the user's movement in the video condition, while the motion capture system captures the user's 3D motion in the skeleton/avatar condition and the Coach Navi condition.

As shown in Figure 4.18, we implemented several functions in the application to support users during training. The following sections detail these functionalities.

Coach Motion Playback

The application displays the coach's motion (i.e., the personalized learning target) to the user. The interface provides a timeline slider that allows users to control the timing of the playback, as well as buttons to play, pause, or loop the motion. This feature enables users to observe the motion in detail and at their own pace, facilitating better understanding and imitation.

Real-Time User Motion

In many motor skill training contexts, trainees benefit from real-time feedback by observing themselves in mirrors. Similarly, our application provides real-time visual feedback to the

user by rendering their 3D poses using virtual cameras. This allows users to see their own movements from various angles and compare them with the coach's motion.

Multi-View Manipulation

The system offers multiple viewing angles—including front, back, left side, right side, and top views—to enable users to examine their swing forms and the coach's poses from different perspectives. This flexibility helps users to focus on specific aspects of their movements and understand the spatial relationships involved in the swing.

3D Mesh and Skeleton Representations

The application provides two representations of the 3D human models: the textured SMPL avatar and a 3D skeletal model. Users can select between these two representations for both the coach's motion and their own motion, resulting in a total of four combinations (coach and user each with either avatar or skeleton representation). This customization allows users to choose the visualization that they find most helpful for their learning process.

4.5 User Study

We implemented Coach Navi using the motion style transfer network and the motion visualizer based on our previous results. The coach's motion was selected from the advanced golfers in our GolfMDB dataset to serve as the learning target.

4.5.1 Hypotheses

In this user study, we investigated the practical effectiveness of the proposed system by formulating the following hypotheses:

***H*1** A 3D visualization is more acceptable for training than a 2D presentation.

***H*2** Coach Navi can help users improve their skills better than conventional methods.

***H*3** Coach Navi's recommendation of an intermediate-level target is more acceptable for beginners to imitate.

***H*4** The 3D model with the user's appearance allows users to identify with the avatars and enhances the training experience.

4.5.2 Participants

Following approval from the Institutional Review Board, we recruited participants from a local university population, including students and faculty members. No incentives were provided for participation or performance. To counterbalance three conditions we examine

in this study, we invited a total of six subjects (two females and four males, aged 20 to 30) to participate in the study. All participants were inexperienced in golf. Two of the participants were left-handed, and the remaining four were right-handed.

4.5.3 Conditions

To validate our hypotheses, we compared our system to a conventional training method using video replay and a real-time skeleton visualization method, serving as baseline skeleton-based visual feedback.

Video Playback and Virtual Mirror (*Cond. V*):

As a conventional baseline, we designed a system that provides video playback of a coach's swing motion. A camera was used to provide a mirrored view of the user, simulating a virtual mirror setup commonly found in practice environments.

Skeleton and Avatar Visualization (*Cond. SA*):

In this condition, the system displays either the 3D skeleton or 3D avatar of the participant in real-time alongside the coach's playback motions. Additionally, timeline controls and view-changing functions are implemented in the same way as in the proposed Coach Navi system.

Coach Navi (*Cond. CN*):

In this condition, the system displays the 3D skeleton or 3D avatar of the participant in real-time alongside the learning target's playback motions. The learning target is optimized by the Coach Navi system to provide an intermediate-level motion tailored to the user's current skill level.

4.5.4 Hardware Setup

The setup for the user study is depicted in Figure 4.18. Participants' motions were captured using a motion capture system comprising 12 OptiTrack Prime 13W cameras. We used Unity as the control interface, with the application receiving tracking data from the motion capture system via the OptiTrack Unity plugin.

In each condition, the visual feedback was projected onto a front projection screen using a HITACHI LP-WU6500 projector. For user interaction, participants were provided with a portable monitor (ASUS MB16AMT) featuring a touch screen to manipulate the application's functions. An iron 7 golf club (Callaway DCB 55R) was used throughout the study.

4.5.5 Procedure

The study procedure was as follows:

1. **Instruction and Practice:** Participants received instruction on handling a golf club and performing a swing, followed by a 5-minute free practice session.

2. **Pre-training Recording:** Participants performed five golf swings, which were recorded as the pre-training baseline performance.

3. **Training Sessions:** Participants experienced the three training conditions (*Cond. V*, *Cond. SA*, and *Cond. CN*) in a counterbalanced order. Each training session lasted 10 minutes.

4. **Post-training Performance:** After each training session, participants performed five golf swings, which were recorded as the post-training performance.

5. **Questionnaires and Interviews:** Participants completed a questionnaire after each training condition and a post-study survey, followed by an interview at the end of the study.

Since all participants were inexperienced in golf, we provided a brief lecture on handling a golf club and performing a swing before the training sessions. After the instruction, participants warmed up and practiced swinging for 5 minutes. Following the free practice, participants performed five golf swings, which were recorded as the pre-training baseline.

Participants then practiced with one of the three conditions for 10 minutes and answered a post-training questionnaire. Each participant experienced all three conditions, with the order counterbalanced to mitigate order effects. During the training sessions, participants were free to rest at any time and were not required to handle the golf club continuously when practicing the swinging motion. After each training session, participants performed five golf swings that were recorded as the post-training performance.

4.6 Results

4.6.1 Quantitative Results

To evaluate the performance improvement of the participants, we used two metrics: Mean Per Joint Angle Error (MPJAE) and Mean Per Joint Position Error (MPJPE). These metrics indicate the differences between the averages of the five golf swings before and after each training condition. Improvement in a metric reflects how well participants improved their swing forms to resemble the coach's motion compared to their performance before training. A greater improvement indicates a better learning effect.

We conducted statistical analyses using a one-way ANOVA to compare the effect of the three different training conditions on the improvement of the golf swing. If a significant effect was identified, we performed Tukey's HSD post-hoc test for pairwise comparisons between conditions.

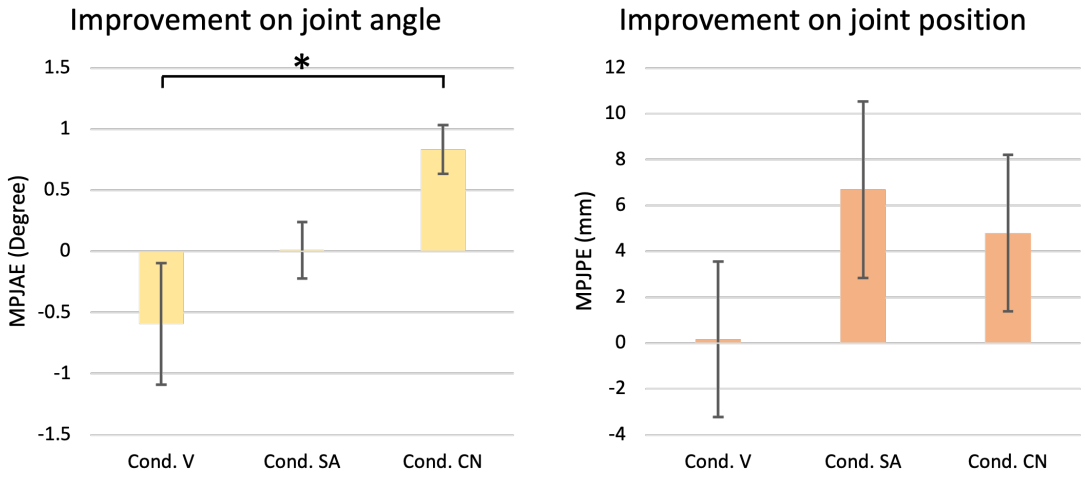

Figure 4.19: Average improvement after each training condition. MPJAE: Mean Per Joint Angle Error. MPJPE: Mean Per Joint Position Error. Brackets indicate significant pairwise differences ($p < 0.05$). A greater improvement indicates a better learning effect.

Improvement in MPJAE

Mean Per Joint Angle Error (MPJAE) measures the average angular difference between the participant's joints and the coach's joints in aligned motions. This metric reflects how accurately participants rotated their bodies to replicate the coach's movements.

The average MPJAE before training was 7.911°. After training, the average MPJAE was:

- *Cond. V*: 8.101°
- *Cond. SA*: 7.444°
- *Cond. CN*: 7.191°

As shown in Figure 4.19, the improvement in MPJAE was:

- *Cond. V*: Decreased by 0.593° (a decline of 7.50%)
- *Cond. SA*: Improved by 0.010° (an increase of 0.13%)
- *Cond. CN*: Improved by 0.833° (an increase of 10.54%)

An ANOVA test revealed a significant effect of the training condition on MPJAE improvement ($F(2, 10) = 3.778$, $p = 0.046$). Tukey's HSD post-hoc test found that the improvement in MPJAE was significantly different between *Cond. V* and *Cond. CN* ($p = 0.038$). There were no statistically significant differences between *Cond. V* and *Cond. SA* ($p = 0.494$) or between *Cond. SA* and *Cond. CN* ($p = 0.284$).

Improvement in MPJPE

Mean Per Joint Position Error (MPJPE) measures the average positional difference between the participant's joints and the coach's joints in aligned motions. This metric indicates how well participants matched their postures to the coach's posture.

The average MPJPE before training was 87.270 mm. After training, the average MPJPE was:

- *Cond. V*: 76.694 mm
- *Cond. SA*: 72.264 mm
- *Cond. CN*: 75.154 mm

As shown in Figure 4.19, the improvement in MPJPE was:

- *Cond. V*: Improved by 0.167 mm (an increase of 0.19%)
- *Cond. SA*: Improved by 6.693 mm (an increase of 7.66%)
- *Cond. CN*: Improved by 4.796 mm (an increase of 5.49%)

However, the ANOVA indicated that the improvement in MPJPE was not significantly different among the three conditions ($F(2, 10) = 0.741$, $p = 0.492$).

4.6.2 Usability and Workload Assessments

To measure the usability of the systems, we used the System Usability Scale (SUS) [25]. The SUS scores were: *Cond. V*: 79.583, *Cond. SA*: 78.750, *Cond. CN*: 85.833.

These scores suggest that all systems were above average in usability, with *Cond. CN* receiving the highest score.

We conducted a more detailed analysis using the NASA Task Load Index (NASA-TLX) [58] to evaluate participants' perceived workload in each condition. Statistical

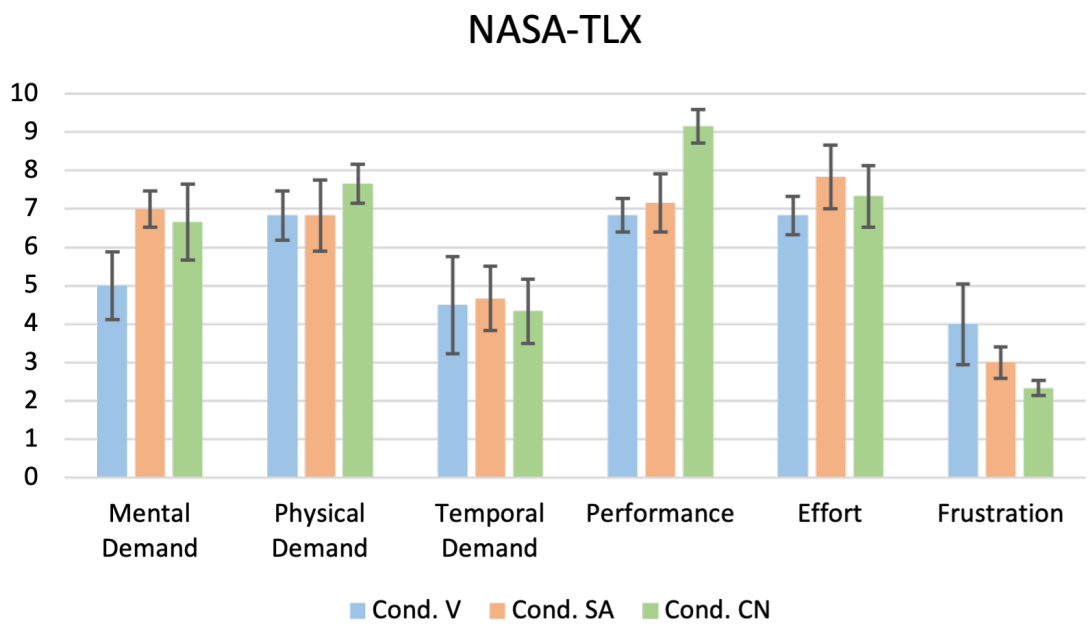

Figure 4.20: Average scores of NASA-TLX for each condition. Error bars represent standard error.

Table 4.5: UEQ-S scores for the three conditions. 'P' stands for pragmatic quality, 'H' stands for hedonic quality, and 'C2B' indicates comparison to benchmark [62].

	Cond. V			***Cond. SA***			***Cond. CN***		
	M	*SD*	*C2B*	*M*	*SD*	*C2B*	*M*	*SD*	*C2B*
P	0.958	0.858	Below Average	1.417	0.719	Above Average	**1.792**	0.697	**Excellent**
H	−1.833	0.701	Bad	0.208	1.269	Bad	**1.458**	1.134	**Good**
Overall	−0.438	0.641	Bad	0.813	0.887	Below Average	**1.625**	0.884	**Excellent**

analysis using the Friedman one-way repeated measures test revealed no statistically significant differences among the three conditions for the following factors: Mental demand ($\chi^2(2) = 5.158$, $p = 0.076$), Physical demand ($\chi^2(2) = 0.857$, $p = 0.651$), Temporal demand ($\chi^2(2) = 0.364$, $p = 0.834$), Performance ($\chi^2(2) = 5.478$, $p = 0.065$), Effort ($\chi^2(2) = 0.381$, $p = 0.827$), and Frustration ($\chi^2(2) = 1.238$, $p = 0.538$).

This suggests that participants experienced similar workload levels across all conditions.

4.6.3 User Experience Questionnaire Short Version (UEQ-S)

We assessed the user experience of the three systems using the User Experience Questionnaire Short Version (UEQ-S) [62], focusing on pragmatic quality, hedonic quality, and overall scores. The results are presented in Table4.5.

The results indicate that *Cond. CN* received the highest scores in all categories, suggesting a superior user experience compared to the other conditions.

Table 4.6: Questions from the post-study survey.

Abbreviation	Question
Acceptability	I found this training method acceptable for learning golf swings.
Imitation	I found it easy to imitate the target movements shown in this training method.
Motivation	This training method motivated me to engage more in the training.
Self-Representation	I felt that the visual representation allowed me to see myself accurately performing the movements.
Proper Level	I felt that the difficulty level of the movements was appropriate for my skill level.
Target Preference	I preferred the level of the learning target provided in this training method.
Body Awareness	I was more aware of my body movements with this training method.
Avatar Identification	I was able to identify with the avatar (or virtual human) used in this training method.
Ideal Imagination	I was able to imagine an ideal version of myself.
Error Correction	I was able to detect and correct my errors using this training method.

4.6.4 Post-Study Survey

We designed a customized questionnaire to gather participants' perceptions of the training systems and their features. The questions are listed in Table 4.6.

Participants responded using a seven-point Likert scale. Figure 4.21 illustrates the results of the post-study survey.

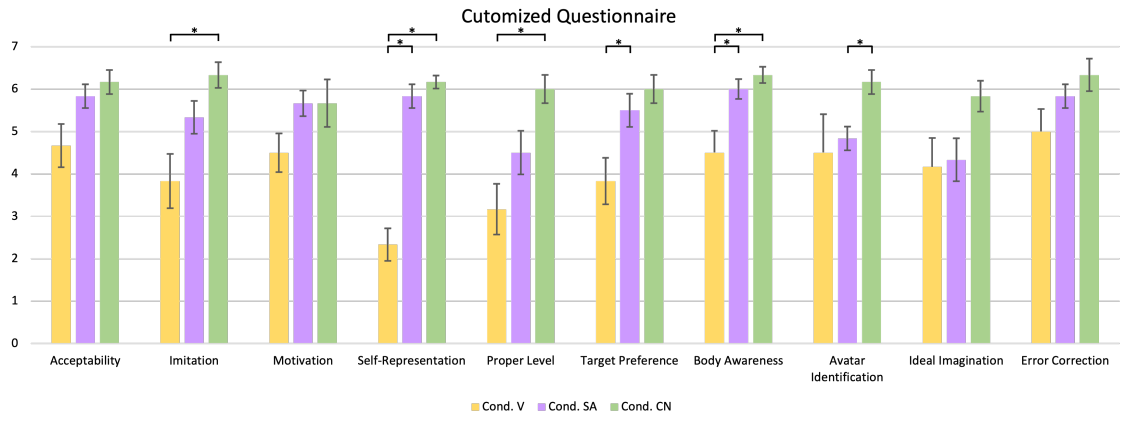

Figure 4.21: Average scores from the post-study survey for each condition. Error bars represent standard error. Brackets indicate significant pairwise differences ($p < 0.05$).

We analyzed the responses using the Wilcoxon signed-rank test. Significant differences were found between the conditions for several questions:

- **Imitation:** Participants found it easier to imitate the target movements in *Cond. CN* compared to *Cond. V* ($Z = 2.041$, $p = 0.041$).

- **Self-Representation:** Participants felt the visual representation allowed them to see themselves more accurately in both *Cond. SA* and *Cond. CN* compared to *Cond. V* (*Cond. V* vs. *Cond. SA*: $Z = 2.232$, $p = 0.026$; *Cond. V* vs. *Cond. CN*: $Z = 2.214$, $p = 0.027$).
- **Proper Level:** Participants felt the difficulty level was more appropriate in *Cond. CN* compared to *Cond. V* ($Z = 2.232$, $p = 0.042$).
- **Target Preference:** Participants preferred the level of the learning target in *Cond. SA* over *Cond. V* ($Z = 2.271$, $p = 0.023$).
- **Body Awareness:** Participants were more aware of their body movements in both *Cond. SA* and *Cond. CN* compared to *Cond. V* (*Cond. V* vs. *Cond. SA*: $Z = 2.060$, $p = 0.039$; *Cond. V* vs. *Cond. CN*: $Z = 2.232$, $p = 0.026$).
- **Avatar Identification:** Participants were better able to identify with the avatar in *Cond. CN* compared to *Cond. SA* ($Z = 2.060$, $p = 0.039$).

These results support our hypotheses that 3D visualization, particularly when personalized with the user's appearance, enhances the training experience and is more acceptable for beginners learning new motor skills.

4.7 Discussion and Future Work

This study investigated the effectiveness of three different training conditions for beginner golfers: Video Playback and Virtual Mirror (*Cond. V*), Skeleton and Avatar Visualization (*Cond. SA*), and Coach Navi (*Cond. CN*). Our aim was to determine which method most effectively enhances skill acquisition, user engagement, and overall training experience. The results provide valuable insights into how beginners perceive and benefit from various training modalities.

4.7.1 Video Training

The Video Playback and Virtual Mirror condition represented the traditional approach to golf training, where learners observe an expert's swing alongside their own reflection. Despite its widespread use, our findings suggest that this method may not be the most effective for novices attempting to grasp complex motor skills like golf swings.

Participants showed minimal improvement in their ability to replicate the coach's movements after training with *Cond. V*. The lack of significant enhancement in movement accuracy indicates that simply watching an expert and trying to mimic their actions may not suffice for beginners. This could be due to the cognitive challenge of translating two-dimensional video representations into three-dimensional bodily movements, which can be particularly demanding for those without prior experience.

Subjectively, while the System Usability Scale (SUS) score for *Cond. V* was acceptable, it was lower than that of Coach Navi. The User Experience Questionnaire Short (UEQ-S) further revealed that *Cond. V* scored the lowest among the three systems in pragmatic quality, hedonic quality, and overall experience. Participants reported difficulties in imitating the movements and accurately perceiving their own body positions. One participant commented on the limitations of this method, stating, "The video has a fixed view, which makes the system not very useful." This suggests that the static nature of video playback, without interactive or adjustable perspectives, may hinder learners from fully understanding and replicating the movements.

Additionally, some participants noted the benefit of seeing the golf club in the video condition, which was absent in the other methods. As one user mentioned, "The good thing in the video condition is I can see the golf club, while the others do not contain the club's motion." This highlights the importance of including equipment representations in training systems to provide a more comprehensive understanding of the movements.

4.7.2 Real-Time Skeleton and Avatar Visualization

The Skeleton and Avatar Visualization condition provided participants with real-time feedback through 3D representations of their movements and the coach's motions. This method aimed to enhance learning by offering a more immersive and interactive experience compared to traditional video playback.

Participants exhibited modest improvements in replicating the coach's movements under *Cond. SA*. While the quantitative measures indicated better performance than *Cond. V*, the differences were not statistically significant. However, the trend suggests that real-time 3D visualization may offer advantages over 2D video by providing a more detailed and engaging representation of movements.

Subjectively, the SUS score for *Cond. SA* was slightly lower than that of *Cond. V*, indicating comparable usability. However, the UEQ-S results showed that *Cond. SA* outperformed *Cond. V* in pragmatic quality and overall user experience, suggesting that participants found this method more effective in achieving their training goals.

Participants expressed a preference for the avatar over the skeleton visualization. One participant stated, "If the avatar is available, I will use the avatar instead of the skeleton because I can realize more the rotation when seeing the 3D avatar. Skeleton is hard for me to recognize the rotation." This indicates that avatars, with their more detailed and lifelike representations, can aid in understanding complex movements by providing visual cues that skeletal models lack.

Despite these advantages, some participants faced challenges due to the dissimilarity between their own bodies and the generic avatars used. A participant noted, "When using the SA condition, the target's skeleton body is significantly different from mine, so it is impossible for me to compare with the skeleton visualization. I used the 3D avatar

instead." This suggests that while 3D visualization is beneficial, personalization is key to maximizing its effectiveness.

4.7.3 Coach Navi

Coach Navi, our proposed system, combined real-time 3D visualization with personalized avatars and adaptive learning targets. By presenting an intermediate-level motion that matched the user's skill level and using an avatar with the user's appearance, Coach Navi aimed to create an "ideal self" for learners to emulate.

Participants demonstrated significant improvement in their ability to replicate the coach's movements using *Cond. CN*. Statistical analysis confirmed that this improvement was significantly greater than that observed with *Cond. V*. Although the positional accuracy did not show a statistically significant difference, the overall trend indicated better performance with Coach Navi.

Subjectively, Coach Navi received the highest SUS score among the three conditions, reflecting excellent usability. The UEQ-S results further supported this, with *Cond. CN* achieving excellent ratings in pragmatic quality and good ratings in hedonic quality. Participants reported that they found it easier to imitate the movements, felt more motivated, and experienced better self-representation with Coach Navi.

The personalized avatars were a standout feature of *Cond. CN*. Participants expressed that seeing an avatar resembling themselves enhanced their engagement and made the training more enjoyable. One participant remarked, "The avatar with my appearance is very fun," while another stated, "The avatar with my texture makes me engage more in the training." This personalization facilitated a deeper connection with the training material, making it easier for learners to identify with the movements and visualize their own improvement.

The adaptive learning targets also played a crucial role. Participants appreciated the intermediate-level targets provided by Coach Navi, finding them more achievable and less intimidating than the advanced-level motions in the other conditions. A participant commented, "The intermediate level is more acceptable to learn, as the rotation of the body in the advanced motion is unreachable for me." This suggests that tailoring the difficulty of the learning targets to the user's current ability can enhance learning effectiveness by maintaining an optimal level of challenge.

4.7.4 "Ideal Me" Despite User Body Adaptation

A recurring question is why we still refer to the generated motion as an "ideal me" if the user's body shape and physical traits inevitably alter the expert's original motion. In other words, once we adapt an advanced-level swing to suit an individual's anthropometrics, do we risk diluting its "idealness"?

In our user study, learning these intermediate-level motions (i.e., user-specific ver-

sions of the advanced motion) indeed helped learners move closer to the advanced-level target more effectively than traditional methods. Participants showed greater improvements in relevant metrics (e.g., MPJAE) when training with their personalized "ideal me." This indicates the potential benefits of customizing advanced motions to each user's skeletal structure while still retaining skillful features of the original expert reference. In future evaluations, we plan to explore additional metrics, such as golf score or other domain-specific performance indices, to further validate the efficacy of these personalized intermediate targets.

Although our current MST approach focuses on joint angles and body dimensions, future work could incorporate more sophisticated biomechanical data—for instance, force-production or muscle activation profiles—to ensure that the "ideal" state remains biomechanically optimal and not just visually similar. By blending advanced-level motion's critical principles with each user's physical constraints, we continue to refine the definition of "ideal me," ensuring it remains both skillfully robust and practically achievable.

4.7.5 Enhanced Self-Representation and Body Awareness

A recurring theme across the study was the importance of self-representation and body awareness in motor skill learning. Participants in *Cond. CN* reported the highest levels of self-representation, indicating that the personalized avatars helped them perceive their movements more accurately. This enhanced perception likely contributed to better error detection and correction, as participants could more easily identify discrepancies between their own movements and the target motions.

The significant differences observed in the self-representation and body awareness questions between *Cond. CN* and the other conditions underscore the value of personalized visual feedback. By seeing an idealized version of themselves performing the correct movements, learners may develop a more precise internal model of the skill, which is essential for effective motor learning.

4.7.6 Motivation and Engagement

Motivation is a key factor in the learning process, influencing both the quantity and quality of practice. Participants reported higher levels of motivation and engagement with Coach Navi compared to the other conditions. The combination of personalized avatars and attainable learning targets appeared to make the training more enjoyable and rewarding.

Participants' comments reflect this increased motivation. One user mentioned, "The proposed system is very interesting," while another stated, "The target's avatar with my appearance helps me imagine how my motion looks like once I improve my skill." These remarks suggest that the personalized and adaptive nature of Coach Navi not only enhances the learning experience but also encourages learners to invest more effort in their practice.

4.7.7 Summary

We can summarize the results by confirming the hypotheses raised before the user study:

- **A 3D visualization is more acceptable for training than a 2D presentation ($H1$).**

  Participants rated the 3D visualization conditions higher in acceptability and reported better self-representation and body awareness compared to the 2D video condition. The enhanced visual feedback provided by the 3D models appears to facilitate a more effective learning experience.

- **Coach Navi can help users improve their skills better than conventional methods ($H2$).**

  Coach Navi led to significant improvements in participants' ability to replicate the target movements, both quantitatively and subjectively. Users found it easier to imitate the motions, felt more motivated, and experienced better self-representation, confirming that Coach Navi is more effective than traditional training methods.

- **Coach Navi's recommendation of intermediate-level targets is more acceptable for beginners to imitate ($H3$).**

  Participants preferred the intermediate-level targets provided by Coach Navi, finding them more appropriate for their skill level. This preference likely contributed to the improved performance and increased motivation observed in *Cond. CN*.

- **The 3D model with the user's appearance allows users to identify with the avatars and improve the training experience ($H4$).**

  The personalized avatars in Coach Navi enhanced users' identification with the learning target, leading to a better training experience. Participants reported higher levels of self-representation and body awareness, supporting the hypothesis that personalized visualizations can improve motor skill learning.

4.7.8 Limitations

While the study provides valuable insights into the effectiveness of different training methods for beginner golfers, several limitations should be considered. Firstly, the sample size was limited, which may affect the generalizability of the findings. Future studies with larger and more diverse participant groups are necessary to validate these results and explore their applicability across different populations.

Additionally, all participants were golf beginners, and the training methods were evaluated solely within this context. The effectiveness of these methods for intermediate or advanced golfers remains unknown. More experienced players may have different needs or respond differently to the training conditions.

Another limitation pertains to the representation of the golf club in the training systems. Participants noted that the absence of the golf club in the 3D visualizations made it challenging to fully understand the swing mechanics. Including equipment representations could enhance the realism of the training and provide a more comprehensive understanding of the movements.

Furthermore, the fixed camera angle in the video condition was identified as a drawback. A static perspective may not capture all aspects of the golf swing, making it difficult for learners to grasp the nuances of the motion. Future systems could incorporate multiple camera angles or interactive controls to allow users to view the movements from different perspectives.

Lastly, while the personalized avatars were generally well-received, not all participants focused on the avatar's appearance. One user mentioned, "I did not try to imagine the ideal version of me because I only focus on correcting my pose and body rotation. Therefore, I did not really look into the texture of the avatar." This suggests that personalization may not be equally important for all learners, and training systems should consider providing options to customize the level of personalization according to user preference.

4.7.9 Future Applications

The findings of this study have important implications for the design of training systems in golf and other sports requiring complex motor skills. Incorporating personalized avatars and adaptive learning targets can enhance user engagement, motivation, and skill acquisition.

Future applications could explore the integration of equipment visualizations to provide a more complete representation of the movements. For example, including the golf club in the 3D models could help learners understand how to coordinate their body movements with the equipment, leading to more effective training.

Advancements in technology, such as virtual reality (VR) and augmented reality (AR), offer opportunities to create more immersive and interactive training environments. Implementing the training systems within VR or AR platforms could further enhance the user's sense of presence and engagement, potentially improving learning outcomes.

Moreover, the adaptive approach used in Coach Navi could be extended to other domains where skill levels vary widely among learners. By tailoring the difficulty of the learning targets and providing personalized feedback, training systems can accommodate a broader range of users and facilitate more effective learning.

Additionally, future research could investigate the long-term effects of using personalized and adaptive training systems. Assessing how these methods influence skill retention, transfer to real-world performance, and sustained motivation over time would provide a more comprehensive understanding of their effectiveness.

4.8 Conclusion

In this work, we proposed a golf swing analysis tool that leverages neural networks to help users intuitively understand the differences between their own swings and those of professional players. Our approach is divided into three main components: motion navigation, style transformation, and manipulation.

Firstly, the motion style transformer network extracts skill features—interpreted as styles—from the input data. Experiments demonstrated that our implementation of the motion style transformer network achieves better performance than common Variational Autoencoder (VAE) implementations, particularly in capturing the nuances of golf swing motions across different skill levels.

Secondly, utilizing the proposed networks, we developed a motion navigator to identify appropriate learning targets within the latent space. Through comparative analysis, we concluded that the motion navigator effectively provides users with intermediate targets that are more accessible for them to begin with, facilitating a smoother progression in skill development.

Thirdly, based on synchronization and discrepancy detection results, we introduced a module called the motion manipulator to reconstruct motion data from the latent space. By interpolating latent vectors between different motions, we can generate additional intermediate steps that guide users from their current skill level to the next, offering a personalized and gradual improvement pathway.

Building upon these components, we presented Coach Navi, the first self-training system that navigates intermediate-level motions for motor skills, and implemented a prototype specifically for golf swing training. The proposed system comprises three modules: the motion navigator, motion style transformer, and motion visualizer.

For the prototype application, we selected the motion style transformer network with indoor motion capture inputs based on our quantitative results, which showed superior performance in terms of cosine similarity compared to other methods. A user study was then conducted to examine the effects of the proposed system. Quantitative metrics and qualitative feedback were collected, leading to the conclusion that training methods incorporating 3D visualization, personalized avatars, and adaptive learning targets can significantly enhance skill acquisition and user experience for beginner golfers. Coach Navi, in particular, outperformed traditional video training by providing a more engaging and effective learning environment.

By enabling learners to visualize an idealized version of themselves performing attainable movements, Coach Navi facilitated better identification with the learning target, increased motivation, and improved performance. These findings support the integration of personalization and adaptability in training systems to meet the needs of individual learners.

Future work should address the limitations identified in this study and explore the

application of these principles in other contexts. By leveraging technological advancements and focusing on user-centered design, training systems can continue to evolve and offer more effective solutions for skill development across various domains.

Chapter 5

Motor Skill Training System using Motion Discrepancy Detection

5.1 Overview

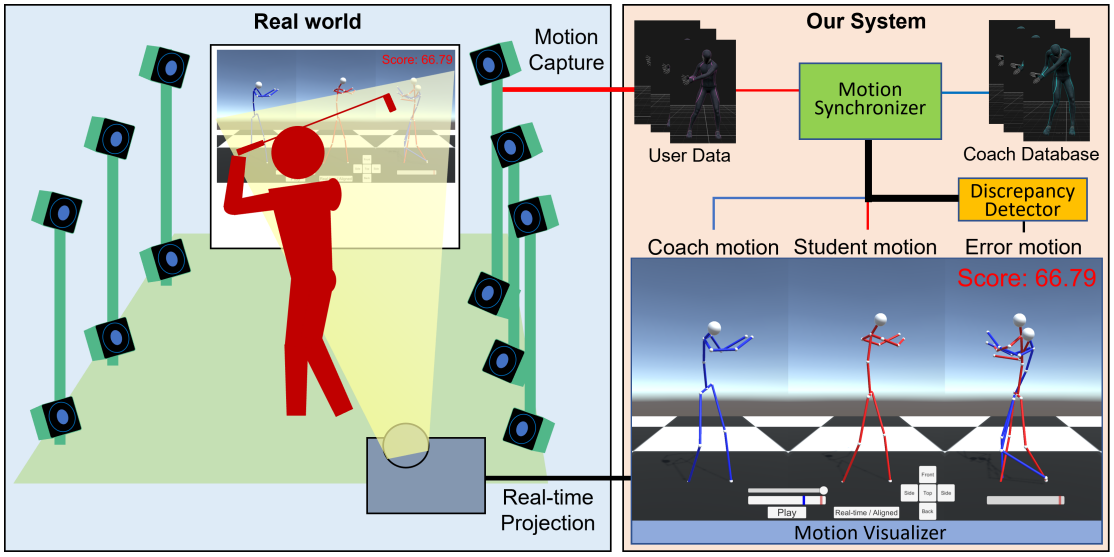

Figure 5.1: System overview of AI Coach. The system captures the user's motion and compares it with a coach's motion from our database. AI Coach visualizes synchronized motions of the two motions and generates a pair of error poses as a recommendation for users to understand the difference between them and correct their forms.

Learning a motor skill such as sports can be divided into several stages [47]: a cognitive phase, an associative phase, and an autonomous phase. Among them, the cognitive phase includes the most basic but essential phase, where a learner starts to understand the skill and mimic a motion of an expert (or someone equivalent) where movements are inconsistent and controlled consciously. Thanks to the development of video technology and network, there are many lecture resources for beginners to obtain this knowledge. However, if the initial recognition of the expert's motion is incorrect, errors in the motion are introduced and will further affect the following learning stage. Furthermore, it

is difficult for learners to observe themselves objectively; thus, currently, many feedback systems [56, 57, 63, 83, 84, 87] have been developed to support the training process. Nevertheless, most of these current feedback systems only focus on specific motion errors instead of the sequence of motion which causes the error.

For example, the critical timestamp of a golf swing is when the golf club collides with the ball, which determines the final ball trajectory. However, it is less effective to tell a learner the error of the collision moment since it may be caused by much previous motion, such as the initial downswing form. Therefore, in those feedback systems, users may still struggle to refine their movement with no idea of which timing of the whole motion, which parts of the body they should focus on, and how they can change their body movements to get their form closer to an expert sequentially.

Nowadays, with the significant advance in machine learning technologies, many systems are built to recognize different actions, estimate precise postures, or even predict the future motions [27, 30, 68, 161, 164], and researchers have been focused on producing self-training systems with neural networks [103, 155, 156]. Recent work [40] introduces an explainable system for pose correction using deep neural networks. Instead of simply showing the joint with the most significant error against a coach, they use a classification-based explanation model to find out which human joints maximally caused a wrong pose in some posture-based skills such as Yoga, Pilates, and Kungfu. The advantage of incorporating machine learning is that the system can learn without prior knowledge; therefore, no expert or coach is required to build such systems, and it can be generalized to many other tasks. In addition, mapping motions of different skill levels to a latent space can create a connection closer to human growth and is thus more explainable than simply using a position error or angle error of human joints.

In this paper, we use golf swing as our learning target to provide users with interpretable clues so they can intuitively understand the difference between themselves and professional players. Since golf is an individual but sequential sport, and the player's standing position is fixed when performing a swing, it is a good starting point for our method. Recent works [77, 81] show the ability of deep learning to retrieve fine-grained information necessary for golf swing analysis. However, after retrieving the essential factor from human motion, the system designer must know which part of the body is vital in golf and determine which parts must be processed for the analysis. Furthermore, the calculation is manually designed for golf swing; thus, it is hard to generalize the proposed methods to other sports analyses without prior knowledge and the help of experts.

In this work, we propose AI Coach: a motor skill training system using neural networks to help users recognize the difference between the user's swing and an expert's motion. Furthermore, we address the previous issues using an unsupervised manner to encourage the network to learn the standard features from the professional players without adding domain-specific information. Consequently, the proposed network can be simply applied to other sports and skill training processes. AI Coach consists of three modules: a motion

synchronizer, a discrepancy detector, and a motion visualizer. The motion synchronizer matches and aligns two input motions with different timing and speed. The motion discrepancy detector can recognize the difference between the two motions and find in which frame the difference is significant. Finally, the motion manipulator is designed to produce intermediate motions between the two motions to provide more intuitive instructions for users to learn.

To evaluate the accuracy and effectiveness of the proposed modules, we collect golf swing data from existing databases and generate a pseudo database containing raw video data, 3D pose data, and labels for the phases during the swing. Next, we examine the accuracy of the motion synchronizer and the capability of motion discrepancy using three types of inputs (raw video, video without background, and 3D human pose). Finally, we discuss the accuracy of the results and explain the correlation among the learned latent space, human motion, and other features. On the other hand, qualitative results are shown to evaluate the ability of the motion manipulator to reproduce motions from the latent space and create new motions unseen in the database. Furthermore, we discuss the proposed analytical tool and its possible applications in future research. The proposed application visualizes the image frames and human motions where the discrepancy between the expert and the user's swing is large and helps the user quickly recognize the motion that needs to be corrected.

To verify the proposed system's effectiveness in helping users learn a golf swing in the real world, we conduct three studies (Study 1, Study 2, and Study 3) to examine the system's performance and usability. The user studys include two comprehensive comparitive studies of the proposed system against video training and skeleton-based visualization methods, and one subjective evaluation study incorporating with a professional golf coach. We discuss in the user studys the quantitative and qualitative results and conclude that the proposed system is effective and valid for motor skills training. In summary, the prototype AI Coach application that combines these modules not only highlights significant errors but also offers actionable frames for self-correction, a feature that our user studies confirmed to be both more helpful and more motivating than traditional video-based training methods.

Our main contributions can be summarized as follows:

- We present the first motion discrepancy detection-based self-training system for motor skills and implement a prototype system for golf swing training.
- The proposed golf training system distinguishes the difference between two sequential motions and visualizes the key point to users to help them understand where and how they make mistakes.
- The proposed method provides intermediate poses that are acceptable during the early learning phase of sports.

- Crucial factors that can influence the accuracy of sports analysis are discussed.
- Three comprehensive user study are conducted with both quantitative and qualitative evaluations, which suggests the proposed system has a better training effect than conventional methods.

5.2 Method

This study aims to create a system that captures user motions and provides fine-grained feedback to improve users' forms by comparing their motions with those of professionals. To achieve the goal of building such an application, the method proposed in this study is to first train a neural network with professionals' motion data. After training the network, the system compresses the user motions through the network into a latent space and compares their motions with those of professionals in the latent space. Figure 5.2 shows an overview of the proposed system. The workflow of the approach is divided into three parts: motion synchronization, motion discrepancy detection, and motion manipulation. The system first receives two motion inputs $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ and uses encoder $\boldsymbol{E}$ to embed the input motions into the latent space, where the two motions are represented as $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$. The encoder is trained to learn a latent space in which similar motions appear to be close.

Next, using the learned latent space, the motion synchronizer $\boldsymbol{MS}$ matches the timing of the two motions in the latent space by measuring the Euclidean distance between $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$. The motion discrepancy detector then captures the two synchronized latent vectors $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$, measures the difference between them, and passes a distance vector $\boldsymbol{D_{12}}$ to the motion manipulator $\boldsymbol{MM}$.

Finally, the $\boldsymbol{MM}$ integrates $\boldsymbol{V_1}$, $\boldsymbol{V_2}$, and $\boldsymbol{D_{12}}$ to specify the key frames where large differences occur and create an intermediate latent vector $\boldsymbol{V_{inter}}$. The system uses decoder $\boldsymbol{D}$ to restore the intermediate human poses $\boldsymbol{Y}$ from $\boldsymbol{V_{inter}}$ for users to gradually improve their motion forms.

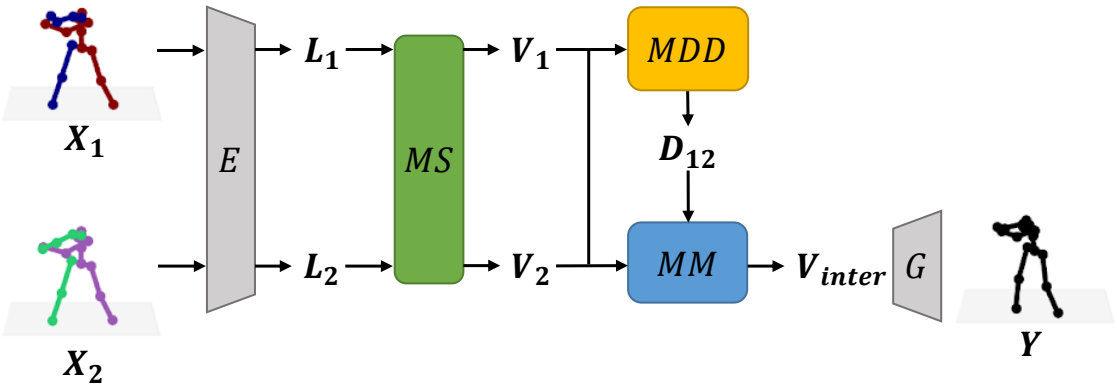

Figure 5.2: System overview. $\boldsymbol{X}$ is the input motion sequence and $\boldsymbol{Y}$ is the output human poses restored from the latent space.

5.2.1 Motion Synchronizer: Aligning Motion Sequences with Different Timing

For motion synchronization, we aim to design a network that learns a latent space that shows motion similarity. A common way to achieve this is by constructing an autoencoder and decoder, whose input and output are the same motions. On the other hand, previous studies have shown that cycle-consistency methods are useful for aligning video inputs with different phasing and timing. Our method is inspired by the temporal cycle consistency (TCC) learning method proposed by Dwibedi et al. [42]. The TCC network is designed to allow the network to learn not only the similarity of motion, but also the temporal order of the entire motion. As shown in Figure 5.3, we implement the TCC algorithm as follows:

- The encoder $\boldsymbol{E}$ compresses two motion sequences to latent vectors $\boldsymbol{L_1}$, $\boldsymbol{L_2}$.
- For each node of $\boldsymbol{L_1}$, find the nearest node of $\boldsymbol{L_2}$.
- For each identified node of $\boldsymbol{L_2}$, we find the nearest node of $\boldsymbol{L_1}$ (cycle back).
- If the node is cycling back to itself, there is no loss; otherwise, TCC loss is calculated.

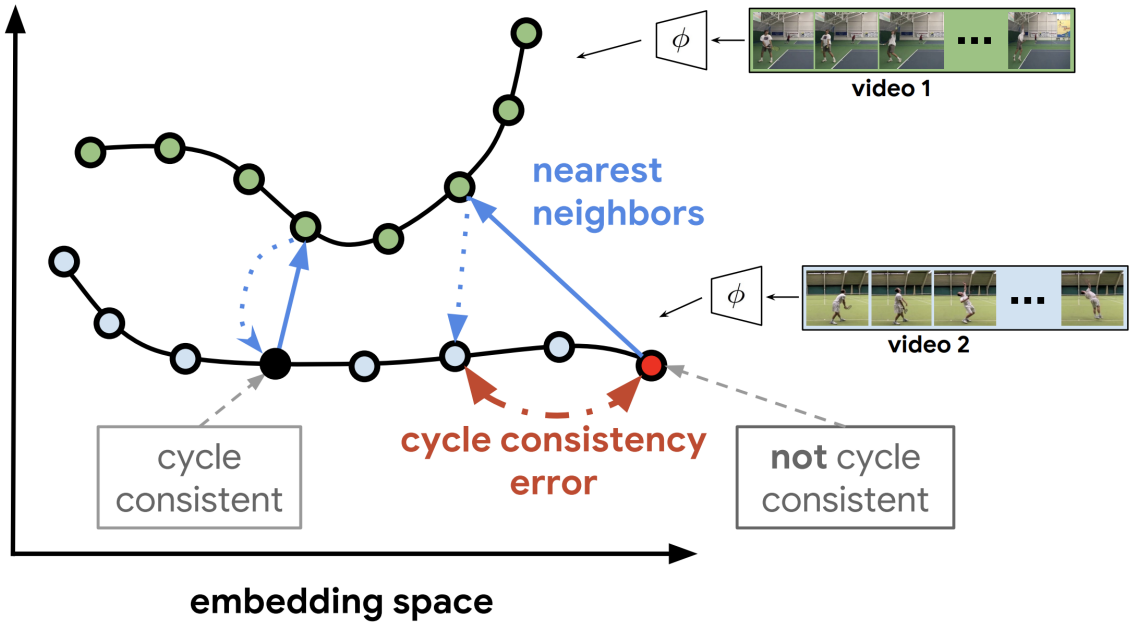

Figure 5.3: The Temporal Cycle-Consistency (TCC) loss. [42]

Cycle-Consistency

For more solid mathematics, we introduce how the TCC loss is calculated in this section.

As shown in Figure 5.4, we assume that we are given two sequences $S = \{s_1, s_2, ..., s_N\}$ and $T = \{t_1, t_2, ..., t_M\}$, with lengths N and M, respectively. s_i and t_i are the i-th frame in the sequence of S and T. The embeddings of the two input sequences are computed

as $U = \{u_1, u_2, ..., u_N\}$ and $V = \{v_1, v_2, ..., v_M\}$ such that $u_i = \phi(s_i; \theta)$ and $v_i = \phi(t_i; \theta)$, where ϕ is the neural network encoder parameterized by ϕ. To check whether a point $u_i \in U$ is cycle consistent, we define its nearest neighbor in V as:

$$\tilde{v} = \sum_{j}^{M} \alpha_j v_j \tag{5.1}$$

$$\alpha_j = \frac{e^{-||u_i - v_j||^2}}{\sum k^M e^{-||u_i - v_k||^2}} \tag{5.2}$$

where:

- α is the similarity distribution that signifies the proximity between u_i and each $v_i \in V$.

Next, we calculate the similarity vector β as:

$$\beta_k = \frac{e^{-||\tilde{v} - u_k||^2}}{\sum j^N e^{-||\tilde{v} - u_j||^2}} \tag{5.3}$$

Same as α, the similarity vector β plays the role of the proximity between $\tilde{v}$ and each $u_k \in U$. At this point, we expect it to have a peak value around the i^{th} index which is the original point that we are checking for the cycle consistency. Note that β is a discrete distribution of similarities over time; therefore, we impose a Gaussian prior on β by minimizing the normalized squared distance:

$$L_{nsd} = \frac{|i - \mu|^2}{\sigma^2} \tag{5.4}$$

$$\mu = \sum_{k}^{N} \beta_k \times k \tag{5.5}$$

$$\sigma^2 = \sum_{k}^{N} \beta \times (k - \mu)^2 \tag{5.6}$$

Furthermore, we enforce β to be peakier around i by applying additional variance regularization. Finally, the final objective is defined as:

$$L = L_{nsd} + \lambda log(\sigma) \tag{5.7}$$

where:

- λ is the regularization weight.

Note that we minimize the log of variance as using just the variance is more prone to numerical instabilities. All these formulations are differentiable and can simply be optimized with conventional back-propagation.

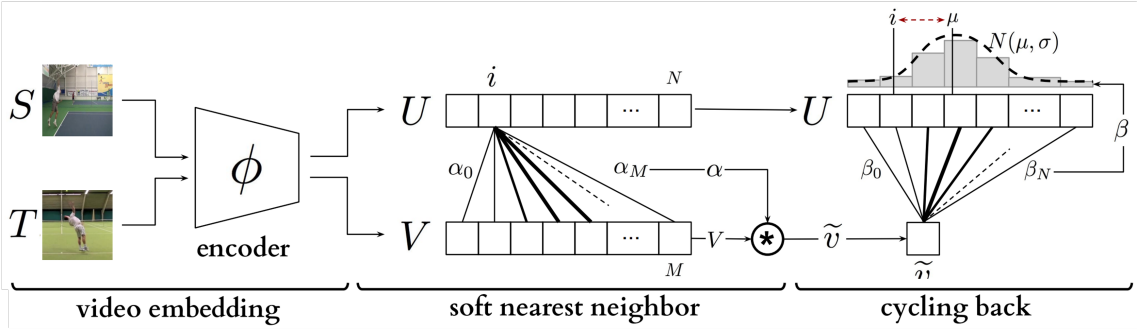

Figure 5.4: Cycle back regression. [42]

Network Design

For the encoder network, we implement a video-based network and two skeleton-based networks. Each network consists of a base network and an embedder network. While the base network is designed to extract features from a given video or skeleton sequence, the embedder network uses the output of the base network and embeds it into the latent space.

For the embedder network, we first store the features of any given frame together with its context frames along the time dimension. Next, we apply 3D convolutions to aggregate the temporal information and reduce the dimensionality using 3D max-pooling. Finally, we use two fully connected layers and a linear projection to obtain a 128-dimensional embedding for each frame.

Three types of networks are implemented in the base networks: video TCC (V-TCC), skeleton TCC (S-TCC), and skeleton-attention TCC (SA-TCC). V-TCC is a network that uses videos as its input. The original TCC implementation is followed to construct the V-TCC. We use the ResNet-50 [61] architecture pre-trained with ImageNet [38] to extract features from the output of the Conv4c layer (Figure 5.5). All frames in a given video sequence are resized to 224×224, and the extracted convolutional features are $14 \times 14 \times 1024$. The convolutional features produced are then fed into the embedder network.

S-TCC is a straightforward implementation of our baseline method that uses skeletons (human poses) as its input. The S-TCC consists of only fully-connected layers. All frames in a given skeleton sequence have a size of 3×16 (joints), and we expand the skeleton input to a single 1×48 vector and feed it to the embedder network.

Because the plain implementation of S-TCC may not be able to learn the relationship among the 3D joints, SA-TCC is another skeleton input version of our TCC network that uses the concept of the self-attention mechanism. As shown in Figure 5.6, in the SA-TCC network, we first expand the 3×16 skeleton input to a single 1×48 vector $\boldsymbol{x}$ and then turn it to query $\boldsymbol{q}(\boldsymbol{x})$, key $\boldsymbol{k}(\boldsymbol{x})$, and value $\boldsymbol{v}(\boldsymbol{x})$. The output $\boldsymbol{y}$ of the attention block is fed to the embedder network, and the transformation in the attention block is formulated

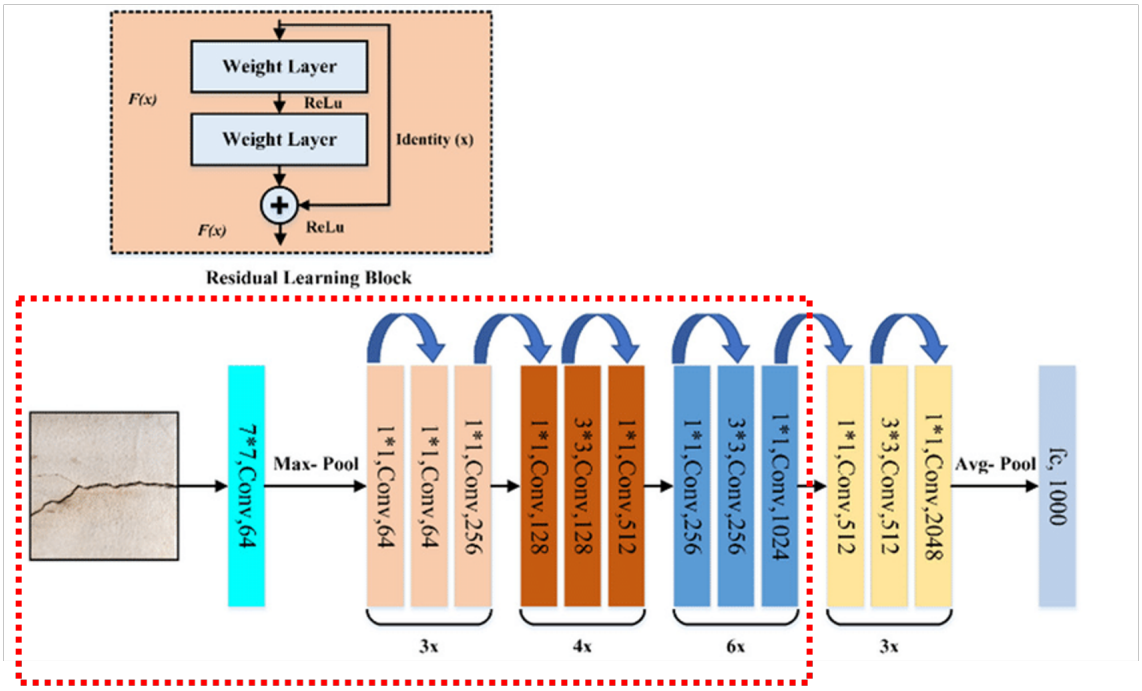

Figure 5.5: ResNet-50 until the 4th stage of convolutional layers. [61]

as follows:

$$\beta_{ij} = \frac{exp(s_{ij})}{\sum_i exp(s_{ij})}, \text{where } \boldsymbol{s} = \boldsymbol{q}(\boldsymbol{x})^T \boldsymbol{k}(\boldsymbol{x}) \tag{5.8}$$

$$\boldsymbol{y_i} = \boldsymbol{f}(\sum_j \beta_{ij} \boldsymbol{v}(\boldsymbol{x})_j) \tag{5.9}$$

In the above transformation, the weights to be learned for $\boldsymbol{q}(\boldsymbol{x})$, $\boldsymbol{k}(\boldsymbol{x})$, and $\boldsymbol{v}(\boldsymbol{x})$ are implemented as 1×1 convolutions. We aim to enable the network to recognize the relationships between different skeleton joints by learning the attention matrix inside the attention block.

5.2.2 Motion Discrepancy Detector: Finding Fine-Grained Motion Differences

After training the network, similar motions must be close together in the latent space. In this section, we focus on the distance between two motions in the latent space to detect and retrieve fine-grained discrepancies and compare two different swing forms, particularly for the differences between beginners and experts. As previously mentioned, the TCC network synchronizes the input sequences by calculating the Euclidean distance between latent vectors. At this point, similar motions appear close to each other in the latent space. As shown in Figure 5.7, because we assume that the network is trained using the golf swings of advanced golfers, a small difference is computed when the input motion is performed similarly to advanced golfers. On the other hand, if the poses between the two input motions are dissimilar in a specific frame, causing a large distance between the aligned

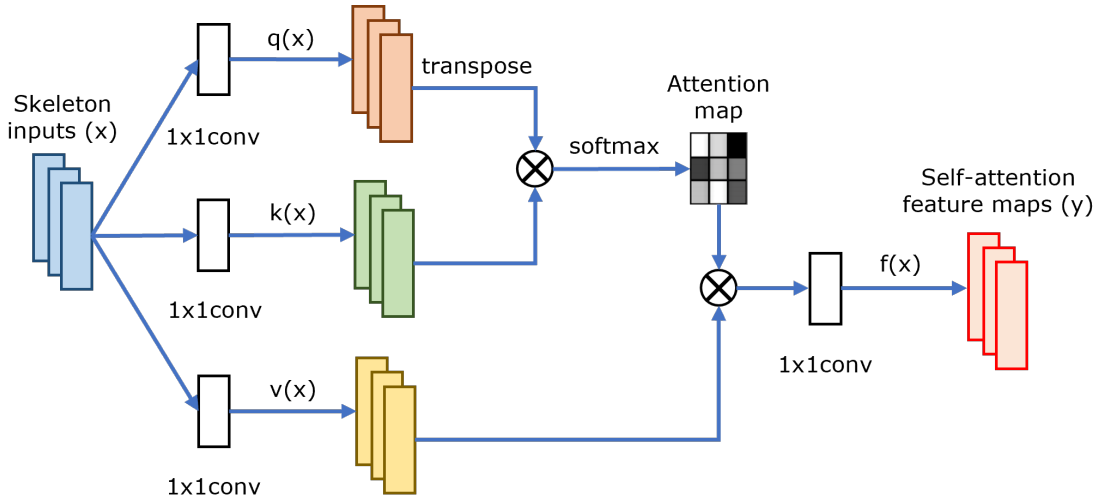

Figure 5.6: Self-attention block. x is the skeleton input, and y is the output. $q(x)$, $k(x)$, and $v(x)$ is the production of the query, key, value respectively. $\otimes$ is the matrix multiplication.

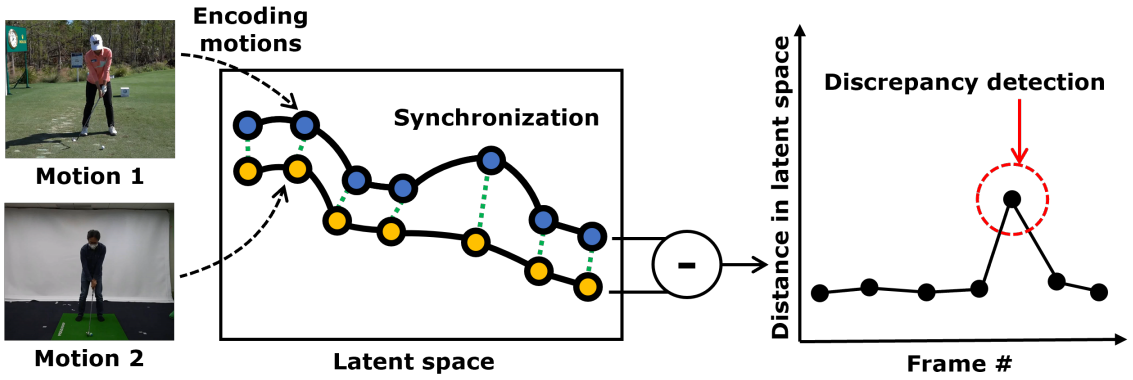

Figure 5.7: Discrepancy detection. The proposed network is encouraged to find a latent space where similar motions appear to be close. After synchronization, frames with large distances in the latent space are considered keyframes where large motion differences occur.

latent vectors, we may find a significant motion difference in that frame. Therefore, we take the latent vectors of the input motions $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$, which are timing-matched by the motion synchronizer, and calculate the frame-by-frame distance vector $\boldsymbol{D}$, which indicates the degree of difference between the motions:

$$\boldsymbol{D_i} = \sum_i ||\boldsymbol{V_{1i}} - \boldsymbol{V_{2i}}||^2 \tag{5.10}$$

5.2.3 Motion Manipulator: Discovering Intermediate Motion between Human Poses

During self-training, it is not always simple for beginners to imitate an ideal motion form, which is very different from their current form. In this study, we propose a motion decoder to generate an intermediate motion. As mentioned previously, the TCC network is trained

to learn a latent space that shows motion similarity. Therefore, we understand that a high-dimensional data point in latent space can be representative of a human pose. To retrieve the intermediate motion between two points in the latent space, we train a decoder using latent vectors to predict human poses that are the same as the inputs. In particular, we first input the training set data into the trained TCC network to obtain the outputs of the latent vectors. Next, using the latent vectors as inputs, we trained a simple decoder consisting of a single fully-connected layer to produce the outputs of the human poses. For the loss function L_{MSE}, we take the mean square error (MSE) between the output and input poses:

$$L_{MSE} = \sum_{i} ||\boldsymbol{Y_i} - \boldsymbol{X_i}||^2 \tag{5.11}$$

where $\boldsymbol{X_i}$ is the i^{th} joint of the input human pose $\boldsymbol{X}$ and $\boldsymbol{Y_i}$ is the i^{th} joint of the output human pose $\boldsymbol{Y}$.

After training the motion decoder, we retrieve a new latent vector $\boldsymbol{V_{inter}}$ between the two timing-matched latent vectors using linear interpolation:

$$\boldsymbol{V_{inter}} = (1 - \alpha) \times \boldsymbol{V_1} + \alpha \times \boldsymbol{V_2} \tag{5.12}$$

where $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$ are the two latent vectors synchronized by the motion synchronizer, and $\alpha \in 0.0, 1.0$ is the magnitude parameter. In the above formulation, by increasing the value of α, we can obtain a human pose whose latent vector is closer to $\boldsymbol{V_2}$, and the restored human pose should ideally be more similar to the human poses generated from $\boldsymbol{V_2}$.

5.3 Evaluation

5.3.1 Experimental setup

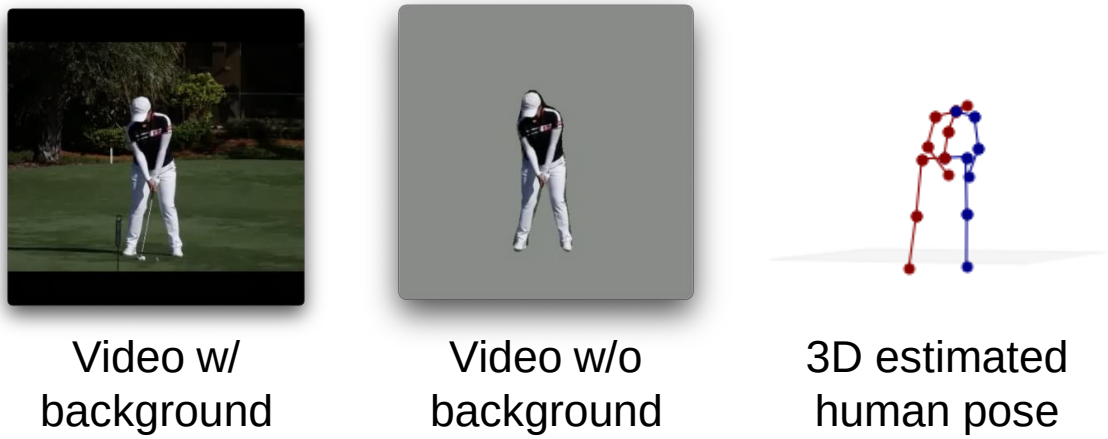

Figure 5.8: Three types of datasets.

To evaluate the accuracy and effectiveness of the three modules introduced in the previous section, we collected golf swing data via the Internet and created a pseudo database consisting of raw video data, without-background-video data, and 3D pose data (Figure 5.8). Next, we implemented three models of the network (V-TCC, S-TCC, and SA-TCC) utilizing TCC loss and conducted statistical analysis under four different conditions:

- V-TCC using video inputs with backgrounds
- V-TCC using video inputs without backgrounds
- S-TCC using 3D human pose inputs
- SA-TCC using 3D human pose inputs

The following sections introduce the data collection process and the evaluation metrics used in this study.

5.3.2 Dataset

Video dataset

GolfDB [98] is a video dataset collection for all types of golf iron swing and driver swing, consisting of 1400 high-quality golf swing videos of male and female professional golfers. While the GolfDB provides preprocessed video clips for a frame size of 160 x 160, we want to rebuild our dataset with high-resolution videos. As the GolfDB contains YouTube video ID and frame numbers of starting and ending of the golf swings, we are able to redownload the original videos from the Internet and trim the videos using this annotation. Note that the YouTube videos are sampled at 30 fps; therefore, in some cases that the original high-quality videos are 60 fps, we downsample the video so that we can follow the annotation of the GolfDB.

This video dataset is used for human pose extraction and input for our neural network, which will be explained in later sections. In order to fit the video input to the various tasks, we make a script to automatically crop and resize the downloaded videos.

Besides the clean videos, we also create a video dataset without background information. This is because the information of the background, for example, the human shadow, may influence the alignment of the network. Moreover, the motion of props such as the golf club may also be learned by the network. We use Mask R-CNN [60], which is a generic object detection and segmentation network, to detect the human body in a single image frame. Then, we remove the background pixels and leave only the part of the human body (Figure 5.9).

3D pose dataset

To conduct a more precise analysis of only human poses, we created a new pseudo dataset consisting of 3D point data of human body poses. As reviewed in Chapter 2, HRNet [136]

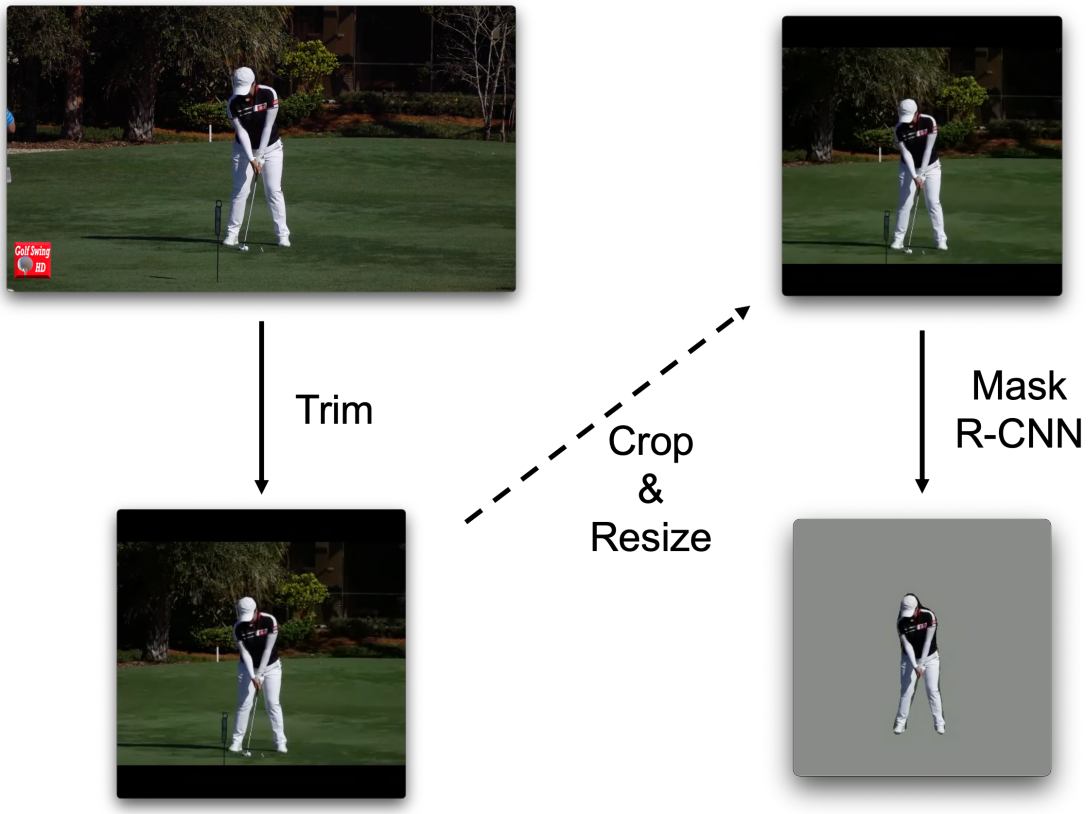

Figure 5.9: The result of background subtraction using MaskRCNN.

has a structure to retrieve a higher resolution of human poses. In this dataset, we first used HRNet to retrieve the time series of 2D human poses from golf-swing videos. The result of the human pose predicted by HRNet is depicted in Figure 5.10. The video is resized to 384 x 288 to fit the HRNet input. While the time series of 2D poses could roughly represent human motion, 2D poses could vary significantly owing to camera poses, and it was difficult to address the normalization problem in 2D space. Therefore, human 3D poses were produced using the simple linear network structure proposed in [96]. The estimated 2D poses from the HRNet were fed to a linear network to retrieve the 3D human poses (Figure 5.8).

The 3D joint point J_i used in this work are defined as follows:

J_0 Hip

J_1, J_2, J_3 Right Hip, Right Keen, Right Foot

J_4, J_5, J_6 Left Hip, Left Keen, Left Foot

J_7 Spine

J_8 Thorax

J_9 Neck

Figure 5.10: The result of HRNet.

J_{10} Head

J_{11}, J_{12}, J_{13} Left Shoulder, Left Elbow, Left Wrist

J_{14}, J_{15}, J_{16} Right Shoulder, Right Elbow, Right Wrist

Normalization

As we are using the ResNet-50 architecture as our video base network, which is pre-trained with the ImageNet, we use the following mean and standard deviation to normalize all the video inputs:

$$X_i^{'} = \frac{(X_i - M)}{S} \tag{5.13}$$

$$M = \begin{pmatrix} 0.485 \\ 0.456 \\ 0.406 \end{pmatrix} \quad S = \begin{pmatrix} 0.229 \\ 0.224 \\ 0.225 \end{pmatrix} \tag{5.14}$$

where:

- X_i is the i-th pixel of an image frame X.
- X_i is the normalized image frame.
- M is the mean.
- S is the standard deviation.

For the skeleton inputs, we want to focus only on the human motion; therefore, to get rid of the impact of the scale of humans, we calculate unit vectors between every joint pair

Figure 5.11: Key event and phase. The impact moment and the top moment are labeled as key events. Frames between them are labeled as swinging down phases.

(Figure 5.12):

$$V_m^n = \frac{J_n - J_m}{|J_n - J_m|} \tag{5.15}$$

where:

- J_i is the i^{th} joint of the skeleton.
- V_m^n the unit vector with the starting joint J_m and the terminal joint J_n.

The 16 unit vectors and the corresponding joints are defined as follows:

V_0^1 Unit vector from hip (J_0) to right hip (J_1)

V_1^2 Unit vector from right hip (J_1) to right keen (J_1)

V_2^3 Unit vector from right keen (J_2) to right foot (J_2)

V_0^4 Unit vector from hip (J_0) to left hip (J_3)

V_4^5 Unit vector from left Hip (J_4) to left keen (J_4)

V_5^6 Unit vector from left keen (J_5) to left foot (J_5)

V_0^7 Unit vector from hip (J_0) to spine (J_6)

V_7^8 Unit vector from spine (J_7) to thorax (J_7)

V_8^9 Unit vector from thorax (J_8) to neck (J_8)

V_9^{10} Unit vector from neck (J_9) to head (J_9)

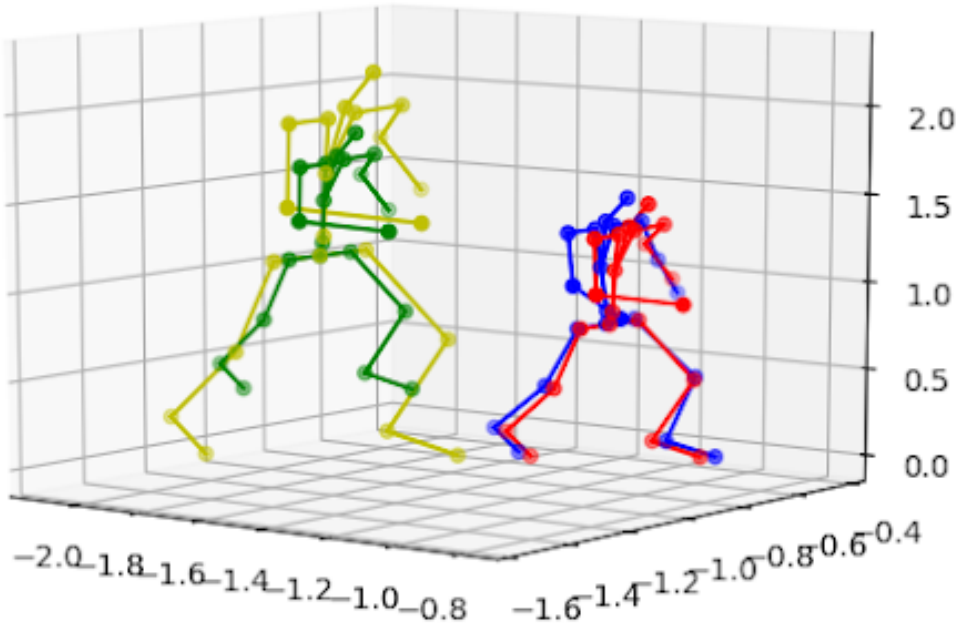

Figure 5.12: 3D normalization of 2 skeletons.

V_8^{11} Unit vector from thorax (J_8) to left shoulder (J_{10})

V_{11}^{12} Unit vector from left shoulder (J_{11}) to left elbow (J_{12})

V_{12}^{13} Unit vector from left elbow (J_{12}) to left wrist (J_{13})

V_8^{14} Unit vector from thorax (J_8) to right shoulder (J_{14})

V_{14}^{15} Unit vector from right shoulder (J_{14}) to right elbow (J_{15})

V_{15}^{16} Unit vector from right elbow (J_{15}) to right wrist (J_{16})

5.3.3 Evaluation Metrics

Because we used a self-supervised learning method, we trained the network until the TCC loss converged. To evaluate how well the network was trained, we applied an accuracy metric showing the precision of the alignment using two label types: key events and phases. A key event is a single frame showing a particular moment, and the phase is a time series between two key events. For example, as shown in Figure 5.11, a key event in golf may be the moment when the golf club hits the ball (impact), and the motion before the golf club hits the ball can be considered as the phase of the golf club approaching the ball (swinging down). Note that all the frames in the period between two key events have the same phase label. Following the key events annotation of the GolfDB (Figure 5.13), the key events are as follows:

1 Address (A). The moment just before the takeaway begins, i.e., the frame before movement in the backswing is noticeable.

2 Toe-up (TU). Shaft parallel with the ground during the backswing.

3 Mid-backswing (MB). Arm parallel with the ground during the backswing.

4 Top (T). The moment the clubhead touches the golf club changes directions at the transition from backswing to downswing.

5 Mid-downswing (MD). Arm parallel with the ground during the downswing.

6 Impact (I). The moment the clubhead touches the golf ball.

7 Mid-follow-through (MFT). Shaft parallel with the ground during the follow-through.

8 Finish (F). The moment just before the golfer's final pose is relaxed.

Figure 5.13: The 8 key events annotation of the GolfDB. [98]

Phases were then labeled between every two key events, for a total of seven phases:

1 A-TU. The movement between address and toe-up.

2 TU-MB. The movement between toe-up and mid-backswing.

3 MB-T. The movement between mid-backswing and top.

4 T-MD. The movement between top and mid-downswing.

5 MD-I. The movement between mid-downswing and impact.

6 I-MFT. The movement between impact and mid-follow-through.

7 MFT-F The movement between mid-follow-through and finish.

Phase Classification

The phase classification accuracy was per frame phase classification. To calculate the accuracy, we first used the encoders of our TCC networks to extract latent vectors. We then trained a simple classifier on the latent vectors to predict the labeled phases. The classifier was trained under several conditions by changing the percentage of the given labeled data. After the classifier was trained, we used all labeled data to calculate the phase classification accuracy. In general, the larger the size of the given labeled data, the higher the accuracy of the classifier.

Procrustes Analysis

To explore more specifically the effectiveness of the network in detecting discrepant motion, we investigated whether the distance in the latent space could represent the discrepancy between two input sequences by computing Pearson's correlation coefficient. Because 3D poses vary from different camera views, we cannot compare two human poses directly using the results from the 3d pose estimator. Thus, we apply Procrustes analysis to align two human poses to make sure they are facing the same direction (or are seen from the same camera direction). The Procrustes analysis is a method of similarity transformation that takes several point groups and aligns the groups to let the points be superimposed (Figure 5.14). This is done by optimally translating, rotating, and uniformly scaling the objects, which are the skeletons in our study. In this thesis, superimposing two skeletons A and B, the optimization equation is defined as:

$$\min_{R,T} \sum_{i}^{N} ||A_i - RB_i + T||^2 \tag{5.16}$$

where:

- $A_i \in R^3$ is the i^{th} joint of skeleton A
- $B_i \in R^3$ is the i^{th} joint of skeleton B
- R is the rotation matrix
- T is the translation matrix.

A brief explanation of this equation is finding an optimal rotation that creates the minimum mean per joint position error (MPJPE). Together with the MPJPE, the Procrustes analysis allows us to compare the motion difference with the latent vectors and apply further analysis in the following sections. The result of the Procrustes analysis generated from two human poses is depicted in Figure 5.15.

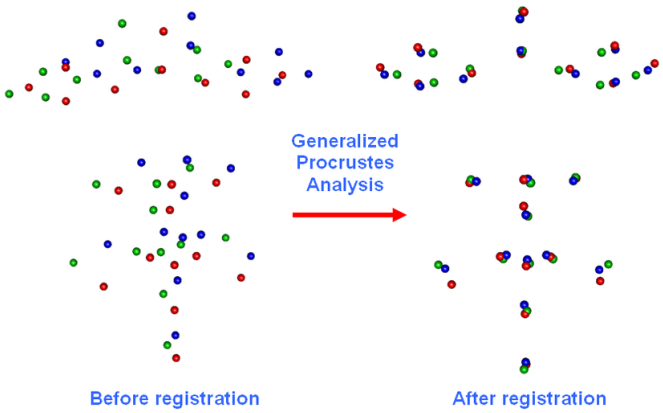

Figure 5.14: Procrustes Analysis. This illustrates an example for registering (aligning) three sets of 21 facial landmarks (displayed in red, green, and blue) obtain for three individuals. [139]

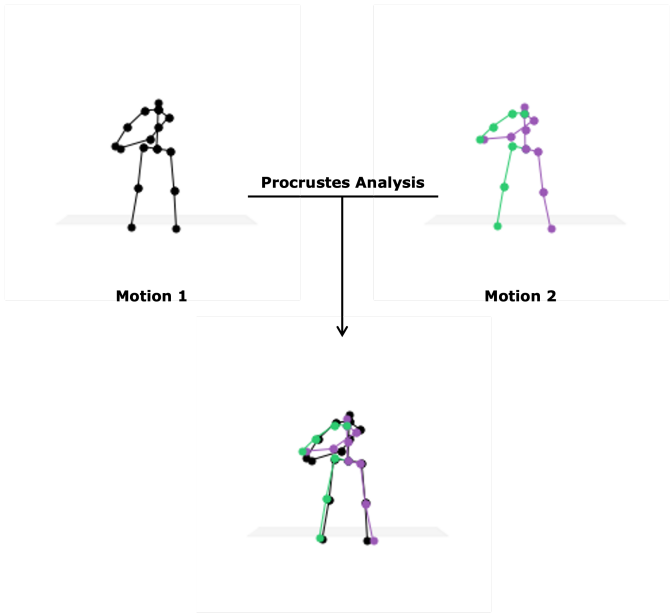

Figure 5.15: The result after the Procrustes analysis.

Pearson's Correlation

In statistics, the Pearson correlation coefficient ρ is a measure of linear correlation between two sets of data. In our experiment, we compare the distance in the latent space with the

mean per joint point error (MPJPE), and the definition for ρ is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{5.17}$$

where:

- cov is the covariance
- X is the distance in the latent space
- Y is the MPJPE
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

5.3.4 Results

This chapter presents the results of an early qualitative analysis investigating the potential of the proposed method to detect discrepant motion differences, followed by more detailed results comparing different modules. Finally, the intermediate human poses generated by the motion manipulator are visualized for qualitative studies.

Case Study

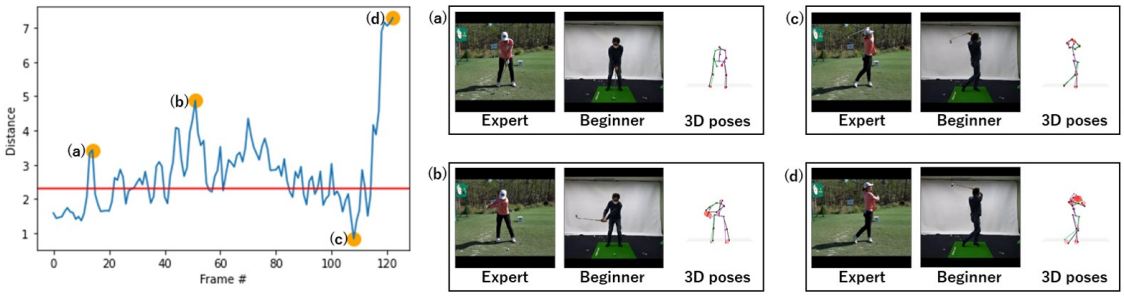

Figure 5.16: Case study with the V-TCC. The line graph shows the distance between two synchronized motions in the latent space. The red line in the graph indicates the threshold for discrepancy detection. The colored skeleton and black skeleton indicate the user's pose and expert's pose, respectively. The density and radius of red spheres indicate the degree of joint position difference between the two skeletons.

In our early study, we conducted a qualitative analysis by exploring the latent space to investigate whether the network could trace fine-grained differences. We first used the V-TCC to synchronize the swing motions of professionals and beginners. We then computed the distances between the aligned videos in the latent space and visualized the overlaid 3D human poses for qualitative comparison (Figure 5.16).

Table 5.1: Phase classification accuracy. This is the accuracy metric showing the ability of the network to classify any given motion frame to its corresponding phase.

Labeled data (%)	5%	10%	30%	80%
V-TCC (with background)	0.718	0.724	0.796	0.840
V-TCC (without background)	0.859	0.839	0.893	0.917
S-TCC	0.895	0.901	0.916	0.929
SA-TCC	0.881	0.902	0.913	0.918

Phase Classification Accuracy

We trained the V-TCC using the GolfDB video dataset with and without background subtraction. In contrast, S-TCC and SA-TCC were trained using the pseudo skeleton dataset with unit vector normalization. After training the four models, we computed the phase classification accuracy for each trained model by assigning 1%, 5%, 10%, 30%, and 80% of labeled data.

The results are presented in Table 5.1. As we trained the network properly, the phase classification accuracy was low when the given labeled data were insufficient and rose with an increase in the number of labeled data.

Correlation

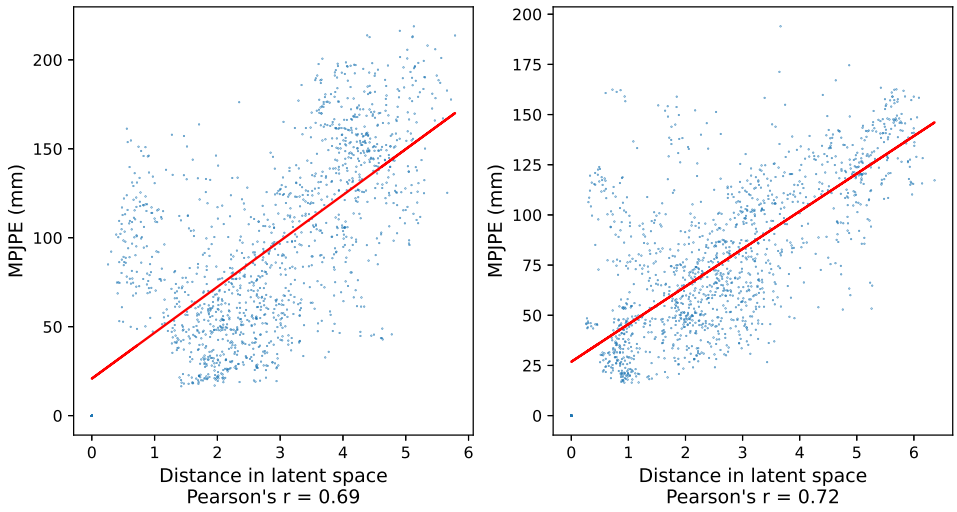

Figure 5.17: Pearson's correlation test for V-TCC. Left: normal videos. Right: videos without background.

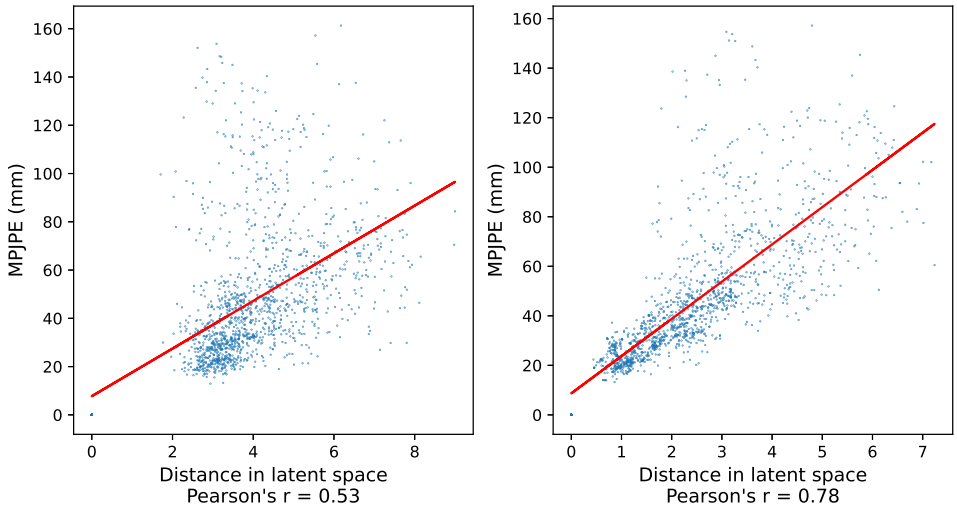

Figure 5.18: Pearson's correlation test with skeleton input. Left: S-TCC. Right: SA-TCC

We computed Pearson's correlation coefficient using the four models. For video inputs, a 0.69 Pearson's correlation coefficient was measured for regular video, and a 0.72 Pearson's correlation coefficient was obtained when the background was removed (Figure 5.17). For skeleton inputs, an over 0.76 Pearson's correlation coefficient was found when using the SA-TCC; however, the lowest Pearson's correlation coefficient with 0.51 was obtained from the S-TCC (Figure 5.18).

Motion interpolation

For the qualitative results, we computed and visualized the intermediate human pose between a pair of human poses considering the following three circumstances:

- The two poses were from a single person. The two poses were in different phases (Figure 5.19 (a)).
- The two poses were from different individuals. The two poses were in the same phase (Figure 5.19 (b)).
- The two poses were from different individuals. The two poses were in different phases (Figure 5.19 (c)).

5.3.5 Discussion

Case Study

In our early case study, we observed that in most cases, when the distance in the latent space was small, the difference between the two 3D poses was small(Figure 5.16 c). On the other hand, when the distance was considerable, the 3D poses showed more differences(Figure 5.16 b, d). This suggested that the network could distinguish between

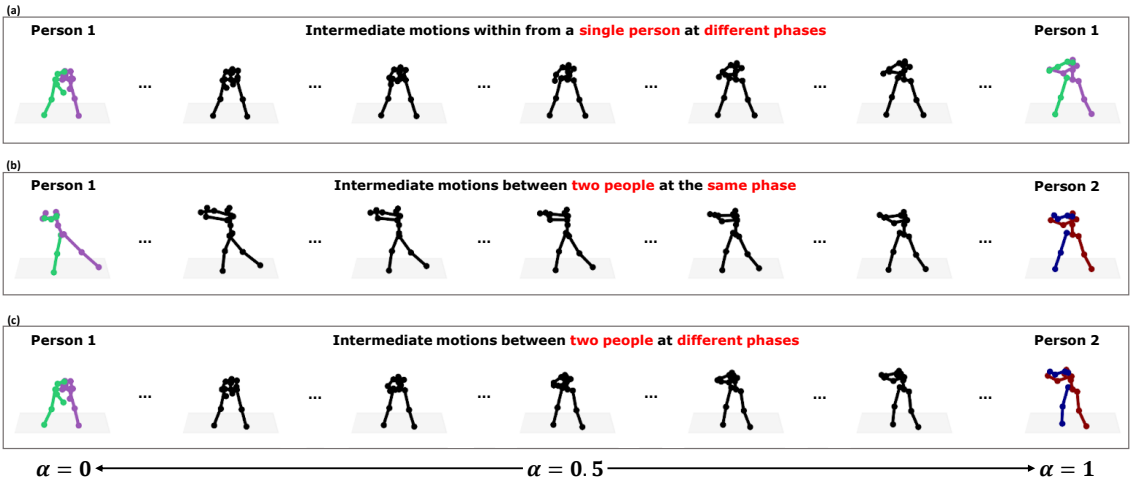

Figure 5.19: Motion manipulation. Unseen intermediate motions between two different 3D human poses are retrieved from the latent space using linear interpolation. α is the blending value. The same color of the skeleton denotes the same person.

beginners and professionals. Thus, we can use this feature to retrieve motion features that help users target the key moves they have to correct. However, as depicted in Figure 5.16 (a), the distance in the latent space remained large even in certain circumstances where the joint difference was negligible. This might be explained by the fact that the network focused not only on human motion but also on other motion features. In this case, the golf club's movement, which was also critical during the swing, might be the main factor causing the enormous distance in the latent space. This led us to the following quantitative studies, where we conducted statistical tests to examine the accuracy of the synchronization and the correlation between the latent space and joint difference under various conditions.

Phase Classification Accuracy

In this quantitative study, we discovered that the previous TCC implementation had limited precision in terms of synchronization. As our hypothesis suggests, synchronization quality might be influenced by various background information and other movements, such as the golf club's and human shadow's motions. When the background was removed, a significant increase in the accuracy of the V-TCC model was observed. This result supported our hypothesis that information outside the human region would affect alignment. Various types of information from the background, the movement of human shadows, and the motion of properties controlled by a human might be learned as features by the network, thus affecting the accuracy of the analysis. Because of the significant superiority of the without-background version, we considered it necessary to apply background subtraction before synchronization when building applications for real usage.

Next, considering the skeleton version, both S-TCC and SA-TCC outperformed V-TCC under all conditions. Given only 10% of the labeled data, both S-TCC and SA-TCC

achieved a phase classification accuracy of over 90%. This might be explained by the fact that the skeleton version had more compressed and precise information than the video version, where the color information might be noisy. Additionally, 3D poses represented high-level human features. Therefore, the features outside the human region were removed when retrieving 3D human poses, resulting in more precise accuracy for further golf swing analysis. Note that the skeleton data were the estimated results from the video data, and the estimation could not be perfect for every human pose. Despite imperfect skeleton inputs, the S-TCC and SA-TCC networks outperformed V-TCC. This result suggested that we could use the skeleton version for more precise implementation in actual usage.

Finally, comparing the two skeleton version models, S-TCC performed better than SA-TCC under most conditions. However, SA-TCC remained competitive in some circumstances (better than S-TCC with 10% labeled data). This led to the following correlation test, in which we discussed the ability of the networks to detect discrepancies between two motions.

Correlation

Discussing the correlation test, as shown in Figure 5.17, we observed a relationship between the distance in the latent space and the MPJPE (mean per joint position error). While the data pairs seemed scattered, a Pearson's correlation coefficient exceeding 0.69 indicated the network's capacity to distinguish whether the difference between two motions was small or large. However, as discussed in the case study section, in some instances the network failed to detect a discrepant motion (latent space showed little difference) even when the joint difference was significant. As shown in Figure 5.17, a higher correlation and more clustered distribution were found when the background and golf club were removed, suggesting that certain background or prop features influence how the network interprets differences.

For the two skeleton-based models, the SA-TCC yielded an over 0.76 Pearson's correlation coefficient, whereas the S-TCC only reached 0.51. Along with the phase classification results (Table 5.1), we see that although S-TCC excelled at synchronization, its ability to detect discrepant motions was limited—potentially hindering real-world application. In contrast, SA-TCC maintained acceptable phase classification accuracy *and* showed a superior correlation value, making it more effective overall.

A key reason SA-TCC achieves a higher correlation is that its attention mechanism focuses the network on critical frames and joints where differences truly matter. MPJPE directly measures how far two sets of 3D joints deviate, frame by frame. By dynamically allocating greater weight to segments where motion discrepancies are pronounced, attention aligns the latent-space encoding of "distance" with the actual joint offsets reflected in MPJPE. In contrast, the S-TCC approach, which relies on static weighting across all frames and joints, cannot adaptively emphasize crucial deviations or downweight noise. As

a result, when genuine skeletal errors occur, SA-TCC's latent distance grows more consistently in tandem with MPJPE, leading to a stronger Pearson's correlation coefficient. In other words, attention pinpoints the parts of the sequence that physically deviate the most, thereby making the latent-space distance track those measurable errors more faithfully.

Hence, we regard the attention-based implementation as the most suitable choice for practical usage. After confirming these findings, we opted to use SA-TCC as the final version of our TCC network for the decoder and overall application implementation, as detailed in the subsequent sections.

Motion interpolation

From the results shown in Figure 5.19 (a), we observed that the trained decoder could restore the ground truth of human poses. Furthermore, the motion decoder demonstrated its ability to generate new human poses unseen in the training dataset. This suggested that the TCC network could learn fine-grained features of the poses of a single person. Notably, the intermediate motion exhibited a continuous change from one pose to another. From this, We could infer that the human poses of a motion sequence were arranged in orders in the high-dimensional latent space.

Next, as depicted in Figure 5.19 (c), we found that the trained decoder could perform fair interpolation between different poses. Specifically, intermediate poses had the characteristics of two different input human poses. For example, we observed continuous changes in the distance between limbs from one human pose to another. This suggested that similar poses which pass through the TCC network were embedded in nearby points in the latent space regardless of the poses from different individuals.

Finally, as shown in Figure 5.19 (b), the motion decoder could generate meaningful intermediate poses at an aligned time. We observed that the human pose gradually changed from an abnormal to a standard form. This suggested that instead of teaching beginners directly with professionals' swing forms, we could provide them with a preliminary pose that is more acceptable for beginners to imitate.

5.4 AI Coach

To verify the proposed system's effectiveness in helping users learn a golf swing in the real world, we implement a GUI application and conduct a user study with golf beginners. As depicted in Figure 5.1, AI Coach comprises three modules: motion synchronizer, discrepancy detector, and motion visualizer. Figure 5.20 shows the workflow of AI Coach. The system first receives two motion inputs $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ and uses an encoder $\boldsymbol{G}$ to embed the input motions to the latent space where the two motions are represented as $\boldsymbol{L_1}$, $\boldsymbol{L_2}$. Here the encoder is trained to learn a latent space where similar motions appear close. Next, with the learned latent space, the motion synchronizer $\boldsymbol{MS}$ matches the timing of the two motions in the latent space by measuring the Euclidean distance between $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$. Then the discrepancy detector $\boldsymbol{DD}$ catches the two synchronized latent vectors $\boldsymbol{V_1}$, $\boldsymbol{V_2}$, measures the distance between them, and finds the most crucial frame where has the largest distance between them. Finally, using the results of the motion synchronizer and discrepancy detector, the motion visualizer renders the synchronized motions $\boldsymbol{X_{1Sync}}$ and $\boldsymbol{X_{2Sync}}$, along with overlaid error poses $\boldsymbol{E_{12}}$ that show the most crucial postures of the two motions at the timing detected by the discrepancy detector.

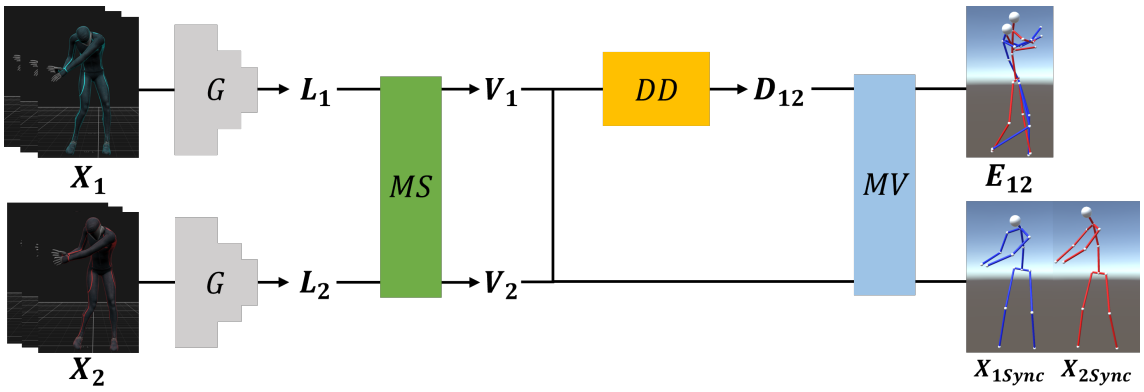

Figure 5.20: Workflow of AI Coach. $\boldsymbol{X}$ is the input motion sequence, $\boldsymbol{E_{12}}$ and $\boldsymbol{X_{Sync}}$ are the output error poses and the synchronized motions.

5.4.1 Motion Synchronizer & Motion Discrepancy Detector

For the motion synchronization and motion discrepancy detection module, we reuse the pipeline introduced in Section 5.2 To find the best presentation for AI Coach, we discuss the accuracy and effectiveness of the motion synchronizer and the discrepancy detector. Using the same experimental setup introduced in the previous section, we conclude the incorporated results of the accuracy and correlation in Table 5.2.

In this quantitative study, we discover that the S-TCC and SA-TCC can reach over 90% phase classification accuracy. This may be explained by the fact that the skeleton version owns the information more compressed and precisely than the video version, where the color information may be seen as noisy. In addition, the 3D poses represent high-

Table 5.2: Incorporated results of the phase classification accuracy and Pearson's correlation coefficient. The phase classification accuracy shows the ability of the network to classify any given motion frame to its corresponding phase. Pearson's r shows the correlation between the distance in the latent space and the MPJPE (mean per joint position error). LD: Labeled Data.

		Phase Classification Accuracy			
Condition	Pearson's r	5%LD	10%LD	30%LD	80%LD
V-TCC (w/ background)	0.69	0.718	0.724	0.796	0.840
V-TCC (w/o background)	0.72	0.859	0.839	0.893	0.917
S-TCC	0.51	0.895	0.901	0.916	0.929
SA-TCC	0.76	0.881	0.902	0.913	0.918

level human features. Therefore, the features out of the human region are automatically removed when retrieving 3D human poses, resulting in more precise accuracy for further golf swing analysis. Note that the skeleton data is the estimated result from the video data, and the estimation cannot be perfect for every human pose. Despite the imperfect skeleton inputs, the S-TCC and SA-TCC networks can still outperform the V-TCC. This result suggests that we may use the skeleton version for more precise implementation for our AI Coach.

Discussing the correlation test, as shown in Table 5.2, we can observe the correlation between the distance in the latent space and the MPJPE. An over 0.76 Pearson's correlation coefficient is found when using the SA-TCC. On the other hand, only a 0.51 Pearson's correlation coefficient can be obtained from the S-TCC. Therefore, along with the classification accuracy results, we find that although the S-TCC network has the most remarkable performance tackling the synchronization, its limited ability to detect the discrepant motion may be a drawback in actual application implementations. On the other hand, the SA-TCC has the superior Pearson's correlation coefficient over all the other competitors while keeping the acceptable phase classification accuracy. Overall, the skeleton model with the attention module implementation outperforms the video models in all aspects of evaluation. Herefore, based on the results, the attention module is the best choice for actual usage. We then use the SA-TCC as our final version of the TCC network for the application implementation.

5.4.2 Motion Visualizer

Using the results of the above modules, we finally implement the motion visualizer that gives users visual feedback showing the synchronized motions and the error poses. As our assumption, it is hard for beginners to recognize the difference between their and advanced golfers' forms because of their inconsistent timing and speed. Therefore, as shown in Figure 5.21, the motion visualizer displays the synchronized motions of both the coach and the user to help the user inspect the differences between the two motions. Furthermore, it is difficult for beginners to find a key point they must focus on to correct the form and improve their skills. To solve this problem, the motion visualizer generates overlaid error

poses that show the most important postures at the keyframe, which the users should carefully take care of to improve their forms (Figure 5.21 Right). The visualization of the overlaid poses can help users quickly target the postures they need to improve instead of searching in the whole motion sequence. As shown in Figure 5.21, several functions

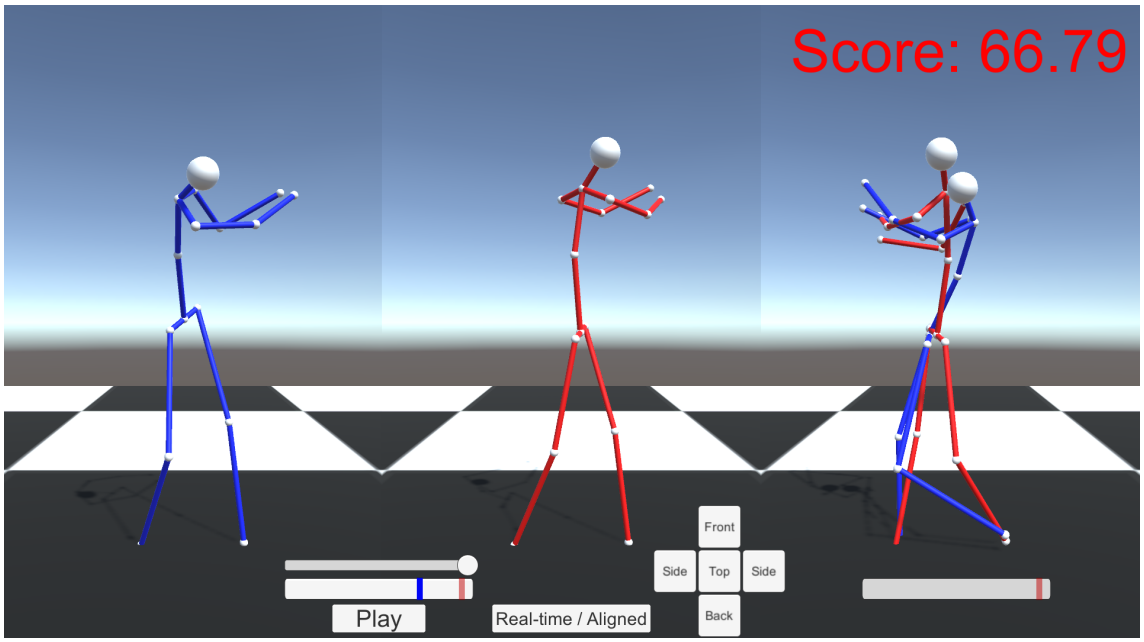

Figure 5.21: The motion visualizer. The blue indicator in the slider shows the current frame, and the red indicator shows the frame with the error poses. Left: coach's pose. Middle: user's pose. Right: error poses.

are implemented in the application, and the following sections will explain the functions' details.

Aligned Motion Playback

The coach's motion and the synchronized motion of the user are displayed in the application. As shown in Figure 5.21, the application provides a timeline slider for users to control the timing. Since the user's speed is matched with the coach's speed, the two poses are always at the same phase and timing during the replay. We aim to let users understand the differences between them and the coach by selecting various frames with the timeline slider and comparing the aligned 3D poses.

Error Poses & Scoring System

As a training system, the proposed application extracts a specific timing where the motions demonstrate the largest distance in the latent space. At that specific timing, the system overlays the user's pose on the coach's pose, as shown at the right in Figure 5.21. This specific timing is concerned as the most critical keyframe where and how the user is suggested to improve their swing motion. During training, users are recommended to revise their poses to match the coach's pose, especially at the keyframe suggested by the

system. Once the user improves their motion, they can update the system by reloading their latest swing motion. After the update, the system will generate another pair of error poses showing the next improvement the user should address if the previous pose is improved.

In addition, AI Coach generates a pseudo score showing the user's performance. After the two motions are synchronized, the system maps the average distance of the two motions in the latent to a performance score $\boldsymbol{S}$ ranges from 0 to 100:

$$\boldsymbol{S} = max(0, 100 - (\frac{\sum \boldsymbol{D_i}}{N})^2) \tag{5.18}$$

where $\boldsymbol{D_i}$ the distance between the two motions in the latent space at frame $\boldsymbol{i}$ and N the total frames of the two motions. Note that since the motions have been aligned, the total frame amounts of the two motions are the same as the motion sequence selected as the reference motion (in our case, the coach's motion). As our hypothesis, the distance in the latent space is correlated to the similarity of the two motions. Therefore, we aim to give users feedback with the performance score, showing them how well they performed, and users may be encouraged when the score increases if they improve their motion correctly.

Real-time Skeleton Visualization & Multi-view Manipulation

As a conventional method in many motor skill training, trainees can stand in front of a mirror and check their motions from the mirror in real-time. In the proposed application, the system also provides real-time visual feedback to the user using virtual cameras to render the 3D poses of the user. In addition, the system provides multiple view angles (front, back, forward side, backward side, and top) to let the users check their swing forms and the coach's poses from various angles.

5.5 User Study 1

We implement AI Coach with the SA-TCC network and the motion visualizer based on the previous results. The coach's motion is recorded using an invited trainer.

5.5.1 Hypothesis

In this user study, we investigate the proposed system's practical effect by raising the following hypothesis:

***H*1** A 3D visualization is more acceptable for training than a 2D presentation.

***H*2** AI Coach can better help users improve their skills than conventional methods.

***H*3** The discrepancy detection increases the user's mental workload for understanding the difference between motions.

5.5.2 Participants

Following the Institutional Review Board's approval, subjects were recruited from a local university population, most of whom were students and faculties. No rewards were provided for recruitment or performance. A total of 9 subjects participated in the study. 3 participants (3 males, age: 20 - 30) were invited to the pilot study. To counterbalance three conditions we examine in this study, we invited 6 individuals (3 females, 3 males, age: 20 - 30) to join the later user study. All of the participants were inexperienced in golf and were right-handed.

5.5.3 Pilot Study

In the pilot study, 3 participants were asked to practice swinging a golf club with AI Coach for 8 minutes and go through an interview after the training. In the pilot study, we first introduced how to perform a golf swing. Then, we asked the participants to practice with the proposed system. Before the practice started, we recorded the motion of the participant for the system to generate the error poses. After every 2 minutes of training, we recorded the latest motion of the participant and reloaded the system again with the motion. After the training with AI Coach, participants were asked to join a brief interview about the system's usability.

In the interview, two participants reported that the 2-minute practice interval was too short for recognizing the discrepancy motion and improving their postures. Also, the participants reported that they spent too much time getting used to the system, and the manipulation with a mouse cursor took most of the time during the training.

After the pilot study, we decided to have minor changes to the procedure of the study and the user interface. In particular, we extended the time interval before each system update and let the participant practice using the system before they started the training

process. In addition, we changed the mouse cursor manipulation to touch screen manipulation.

5.5.4 Conditions

To validate our hypothesis, we compare our system to a conventional training method using video replay and a real-time skeleton visualization method as a baseline skeleton-based visual feedback.

Video Playback and Virtual Mirror (Cond. V): As a conventional baseline, we design a video playback of a coach's swing motion, and a camera is used to provide a mirror view of the user.

Real-time Skeleton Visualization (Cond. S): In this condition, the system displays the 3D poses of the participant's real-time and the coach's playback motions. In addition, a timeline and view-changing functions are implemented the same way as the proposed AI coach system.

AI Coach (Cond. A): This is the training condition with the proposed system.

5.5.5 Hardware Setups

The setup for the user study is depicted in Figure 5.22. Either the appearance or the posture of the user is captured by a single RGB camera (Logitech V-U0028) or by the motion capture system, which comprises 12 motion capture cameras (Optitrack Prime 13W). We use Unity as the control interface. The Unity application receives either the image frames from the RGB camera or the tracking data from the motion capture system using the Optitrack Unity plugin. In each condition during the user study, the visual feedback is projected on the front projection screen by a projector (HITACHI LP-WU6500). As the user interface, participants are given a portable monitor (ASUS MB16AMT) with a touch screen to manipulate the functions in the application. An iron 7 golf club (Callaway DCB 55R) is used during the entire user study.

5.5.6 Procedure

The overview of the study procedure is described as follows:

1. Instruction of performing a golf swing and 5-minute free practice
2. Pre-training recording: Swing a golf club 5 times for comparison
3. Different order of the following 3 training conditions for 9 minutes each:
   Cond. V: The video condition
   Cond. S: The skeleton condition
   Cond. A: The AI Coach condition*
4. Post-training performance: Swing a golf club 5 times after each training condition

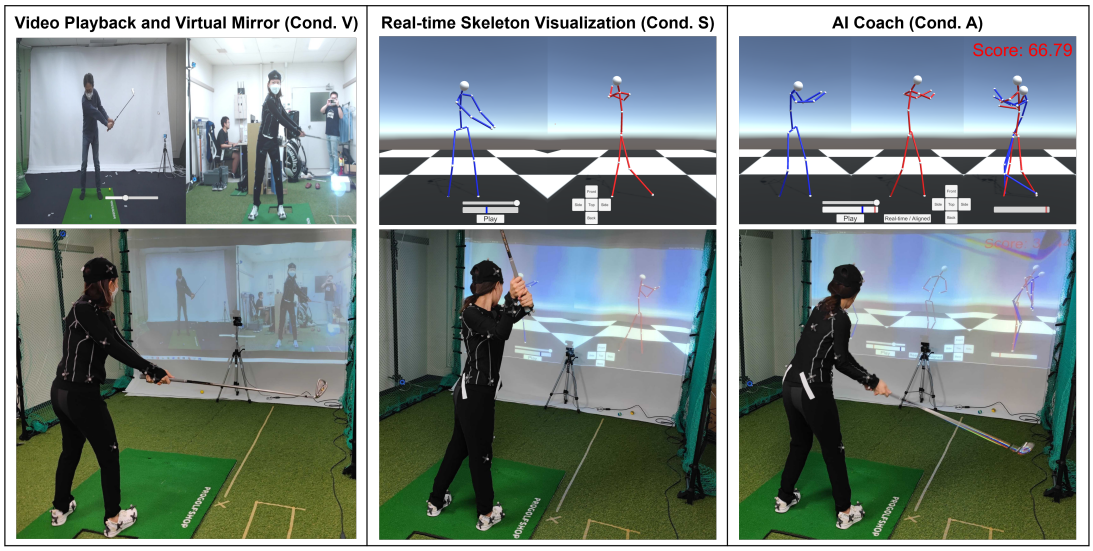

Figure 5.22: Three different setups for the user study. A webcam records the user's movement in the video condition, while the motion capture system captures the user's 3D motion in the skeleton and AI Coach conditions. The top row displays the application views in Unity. The bottom row depicts real-world training environments.

5. A questionnaire after each training condition, a post-study survey, and an interview at the end

* For the AI Coach condition, the motion discrepancy detection (the error pose) is updated every 3 minutes. Time for system updating is not counted.

Since all participants were inexperienced in golf, we gave a lecture about handling a golf club and performing a swing before the training process. After the brief instruction, we asked the participants to warm up and practice swinging for 5 minutes. After the free practice, participants were asked to perform 5 golf swings, and the swing motions were recorded as the pre-training baseline performance. Afterward, participants were asked to practice with one of the three conditions for 9 minutes and answer a post-training questionnaire. The participants would practice with all three conditions while receiving all combinations of training in different orders for counterbalancing. During the training, participants were free to rest at any time and were not forced to handle the golf club when practicing the swinging motion. After each training process ended, participants were asked to address 5 golf swings recorded as the post-training performance.

5.5.7 Results

Quantitative Results

For the quantitative results, we use three metrics to evaluate the performance improvement of the participants. All three metrics indicate the differences between the average of the 5 golf swings before and after each condition. The improvement of a metric shows how well the participants have improved their swing forms to be closer to the coach's motion,

compared with their performance before the training. Therefore, a greater improvement shows a better learning effect. We report the statistical results using a one-way ANOVA analysis to compare the effect of three different training conditions on the improvement of the golf swing. If we identify a significant effect, we then conduct Tukey's HSD posthoc test for comparing each pair of two conditions.

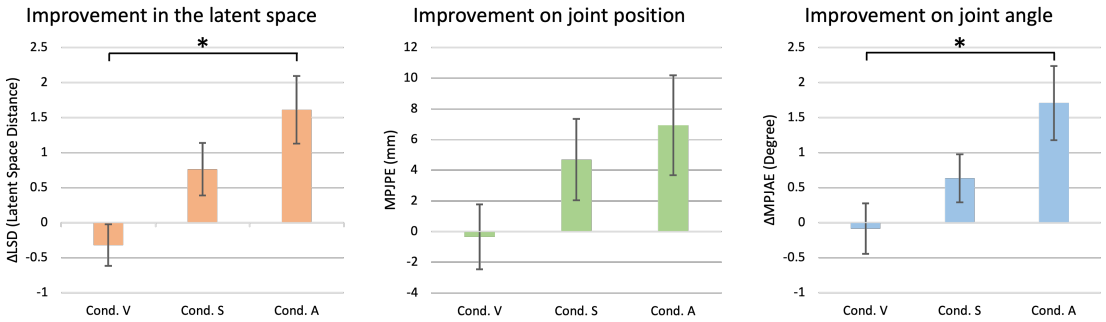

Figure 5.23: Average of the improvement after each training condition. LSD: Latent Space Distance. MPJPE: mean per joint position error. MPJAE: mean per joint angle error. Brackets indicate significant pairwise differences ($p < 0.05$ (*)). A greater improvement shows a better learning effect.

ΔLSD

LSD (Latent Space Distance) is the distance calculated between two aligned points in the latent space. This metric shows how much the users get close to the coach in the latent space. The average LSD of all participants before the training started was 6.402.

As shown in Figure 5.23, the improvement of LSD was -5.03% for Cond. V ($\Delta LSD = -0.322, SE = 0.298$), 11.90% for Cond. S ($\Delta LSD = 0.761, SE = 0.373$), and 25.1% for Cond. A ($\Delta LSD = 1.609, SE = 0.482$). In addition, the result of one-way ANOVA indicated a significant difference among the three conditions for $p < 0.05$ ($F(2, 10) = 5.08, p = 0.021$). Finally, Tukey's HSD Test revealed that the improvement of LSD was significantly different between Cond. V and Cond. A for $p < 0.05$ ($p = 0.016, 95\% C.I. = [0.353, 3.511]$). There was no statistically significant difference in the improvement of LSD between Cond. V and Cond. S ($p = 0.208$) or between Cond. S and Cond. A ($p = 0.368$).

$\Delta MPJPE$

MPJPE (Mean Per Joint Position Error) is the average positional error between joints of two aligned motions. This metric shows how well the users match their postures to the coach. The average MPJPE of all participants before the training started was 133.93 mm.

As shown in Figure 5.23, the improvement of MPJPE was -0.299% for Cond. V ($\Delta MPJPE = -0.400mm, SE = 5.55mm$), 10.47% for Cond. S ($\Delta MPJPE = 14.02mm, SE = 8.29mm$), and 13.60% for Cond. A ($\Delta MPJPE = 18.21mm, SE = 6.24mm$). However,

the result of ANOVA suggested that the improvement of MPJPE was not significantly different among the three conditions for p < 0.05 ($F(2, 10) = 1.71, p = 0.213$).

$\Delta MPJAE$

MPJAE (Mean Per Joint Angle Error) is the average angular error between joints of two aligned motions. This metric shows how well the users turn their bodies to copy the coach's movement.

As shown in Figure 5.23, the improvement of MPJAE was -1.08% for Cond. V ($\Delta MPJAE = -0.088, SE = 0.359$), 7.71% for Cond. S ($\Delta MPJAE = 0.633, SE = 0.343$), and 20.79% for Cond. A ($\Delta MPJAE = 1.70, SE = 0.530$). In addition, the ANOVA test revealed a significant effect for the condition for p < 0.05 ($F(2, 10) = 3.85, p = 0.045$). Finally, Tukey's HSD Test found that the improvement of LSD was significantly different between Cond. V and Cond. A for p < 0.05 ($p = 0.037, 95\% C.I. = [0.105, 3.484]$). There was no statistically significant difference in the improvement of LSD between Cond. V and Cond. A ($p = 0.523$) or between Cond. S and Cond. A ($p = 0.256$).

SUS and NASA-TLX

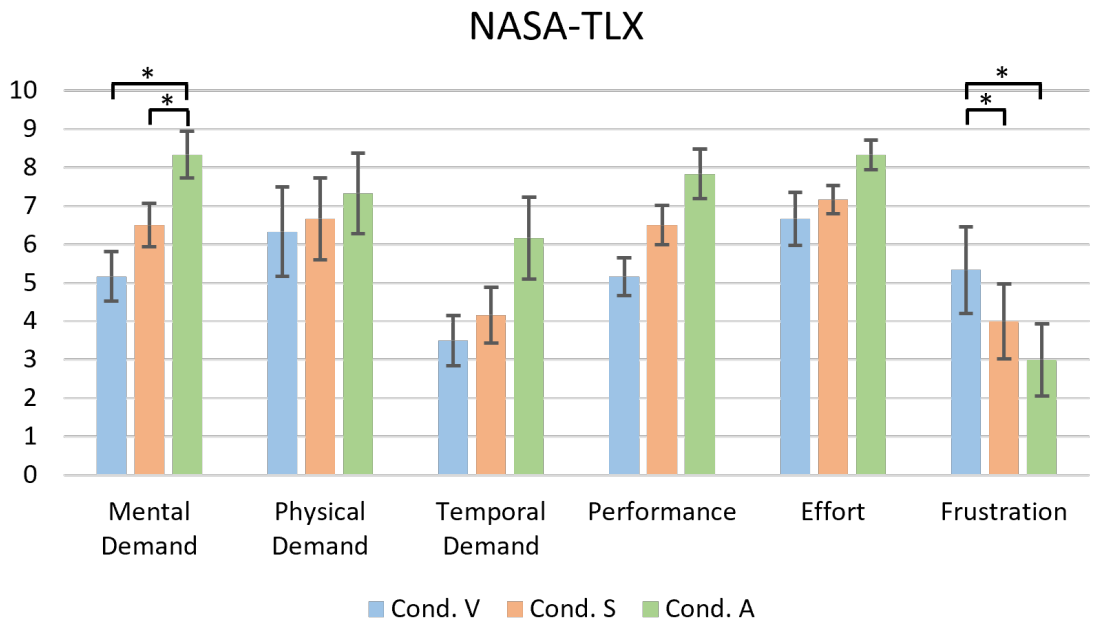

Figure 5.24: Average scores of NASA-TLX. Brackets indicate significant pairwise differences (p < 0.05 (*)).

To measure the usability of the proposed system, we use the system usability scale (SUS [25]) to measure the usability of the proposed system and the skeleton real-time visualization system. As a result, the SUS score was 80.833 for the proposed AI Coach and 77.083 for the skeleton real-time visualization system.

Furthermore, a more detailed analysis is conducted using NASA-TLX [58] to evaluate the participant's workload and mental/physical condition in each condition. We use a five-level Likert scale questionnaire to ask the participants' preferences regarding which system they would like to use for training and what they think about the functions implemented in the system. We report the statistical results using the Friedman one-way repeated measures analysis, followed by the Wilcoxon signed-rank test. Figure 5.24 depicts the results of the NASA-TLX. A Friedman one-way repeated measure analysis revealed that there was a significant difference among the three conditions in the question of mental demand ($\chi^2(2) = 11.27, p = 0.004, \alpha = 0.05$) and frustration ($\chi^2(2) = 7.23, p = 0.027, \alpha = 0.05$). On the other hand, no statistically significant difference was found in other questions: physical demand ($\chi^2(2) = 1.273, p = 0.529, \alpha = 0.05$), temporal demand ($\chi^2(2) = 3.391, p = 0.183, \alpha = 0.05$), performance ($\chi^2(2) = 5.182, p = 0.075, \alpha = 0.05$) and effort ($\chi^2(2) = 5.571, p = 0.062, \alpha = 0.05$).

After the Friedman test, a Wilcoxson signed-rank test revealed the significant difference between Cond. A and Cond. V ($Z = 2.214, p = 0.027$) and between Cond. A and Cond. S ($Z = 2.232, p = 0.026$) in the mental demand question. There was no statistically significant found between Cond. V and Cond. S ($Z = 1.841, p = 0.066$). On the other hand, the Wilcoxson signed-rank test showed a significant difference between Cond. V and Cond. S ($Z = 2.070, p = 0.038$) and between Cond. V and Cond. A ($Z = 2.032, p = 0.042$) regarding the frustration. There was no statistically significant found between Cond. S and Cond. A ($Z = 0.962, p = 0.336$).

Post-study Survey

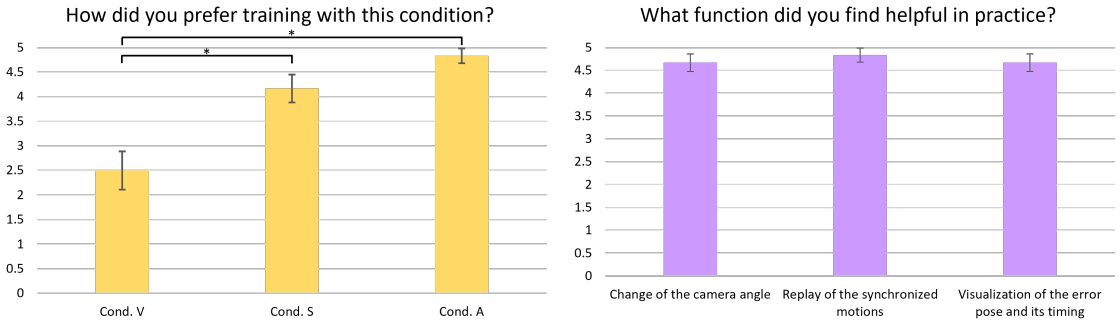

Figure 5.25: Average scores of the five-level Likert scale post-study survey. Brackets indicate significant pairwise differences (p $<$ 0.05 (*)).

Figure 5.25 depicts the results of the post-study survey. A Friedman one-way repeated measure analysis revealed a significant difference among the three conditions among the questions of condition preferences ($\chi^2(2) = 11.14, p = 0.004, \alpha = 0.05$). On the other hand, no statistically significant difference was found among the system functionalities questions ($\chi^2(2) = 2.00, p = 0.368, \alpha = 0.05$).

After the Friedman test, a Wilcoxson signed-rank test revealed the significant difference

between Cond. V and Cond. S ($Z = 2.264, p = 0.024$) and between Cond. V and Cond. A ($Z = 2.214, p = 0.027$) in the preference question. There was no statistically significant found between Cond. S and Cond. A ($Z = 1.633, p = 0.102$).

5.5.8 Discussion

Video & Mirror Training

The video and mirror training method has been treated as a typical method of copying the motion of advanced players. In our user study, all three metrics used for the quantitative study showed a slight improvement in the conventional video training condition. Furthermore, this training method had average minus effects on the user's skill improvement. This fact indicated the limitation of training with a video, and beginners could struggle to learn a new motor skill by looking at videos without help from coaches.

In the interview, most subjects reported that they could not understand the coach's movement and could not determine if they were performing the swing accurately. Also, while they considered the speed controller useful for training, the fixed camera angle limited them from inspecting the coach's detailed posture and the angular movement of the body parts. Supporting our hypothesis $\boldsymbol{H1}$, the video condition was not effective enough to help the users improve their movements.

Real-time Skeleton Visualization

The real-time skeleton condition visualizes the user's real-time posture and the coach's motions replay, providing various changeable camera angles for manipulation. In the user study, all three metrics indicated improvement in the participants. Unlike the video condition, the skeleton condition showed more potential in helping the users develop an accurate motion. In the post-study survey (Figure 5.25), we found a significant difference in the preference between the skeleton and video conditions. This showed that the participant was more willing to practice with a skeleton-based presentation. In addition, in the interview, 5 of the participant reported that the skeleton visualization was more intuitive to learn and imitate because of the simple presentation of 3D human poses and the ability to inspect the poses from different camera angles ($\boldsymbol{H1}$).

On the other hand, we did not find significant differences between the skeleton and video conditions in the statistical study. This fact revealed that the skeleton condition was helpful for training but needed to be more efficient for self-training, especially for beginners. In the interview, some subjects mentioned that they could somehow recognize the dissimilarity between them and the coach's motion but needed help finding where and how they could correct their poses. Therefore, the need for beginners to access help or guides remains in the early learning phases. Otherwise, beginners can quickly learn the wrong moves, which may represent an additional obstacle in the future as the wrong movements remain preserved in their memories.

AI Coach

As a step forward from the skeleton condition, the AI Coach condition generates a pair of error poses as a recommendation for users to correct their poses. In the user study, all three metrics outperformed, on average, the video condition and the skeleton condition. This fact showed the effectiveness of the proposed method showing the error poses as a correction recommendation. In the statistical study, the participants' angular improvement using the proposed system was more remarkable than conventional video training. Since the golf swing is a motor skill that requires a large amount of rotation, joint angle error is an essential factor in golf. During the practice, using the proposed system, users could correct their joint angles to match the coach, which would be effective for training.

In the post-study survey, we found an average greater preference for AI Coach than the skeleton condition and a significantly greater preference over the video condition. The proposed method also had a more excellent score on SUS than the skeleton condition. This fact suggested that the proposed system was accepted by the participant and could act as a proper assistant during training. In the post-study survey, the participants agreed that the synchronized motions and the error pose were helpful in practice. In the interview, the participants reported that the synchronized motion was helpful because they could check their poses and the coach at the exact same timing, which was more straightforward in understanding the differences between the two motions. Furthermore, the subjects reacted that the error pose gave them a good starting point to follow to correct their mistakes. With the recommendation for pose correction, they could focus on the correction instead of spending time finding the differences throughout the entire motion sequence. Therefore, we consider the system could act as a coach giving correction suggestions, and the suggestion could lead to a greater performance improvement compared with conventional training methods, so we accept our hypothesis $\boldsymbol{H2}$.

Latent Space Distance & Scoring System

As shown in the results, the LSD significantly improved in the AI Coach condition, which means the score increased after the condition. While we only observed a slightly better improvement for the latent space distance when training with the proposed system than in the real-time skeleton visualization, we found significant differences compared to the video condition. As discussed in the experiment chapter, the latent space distance correlates with the differences between poses. Along with the statistical results of the user study, we consider the scoring system to reflect the users' performance and be effective during training.

In the interview, participants reported feeling joyful when they saw their score increase. They felt more encouraged by the AI Coach than in other conditions because the score conveyed how well they had performed. Even if the score was low, the proposed system explained the user's performance with error poses, and that suggestion helped them keep

their motivation in training. This fact suggested that the scoring system played a vital facilitating role during the study, and we consider the scoring system valid for implementing a training system.

Mental demand vs. Frustration

In the NASA-TLX questionnaire (Figure 5.24), we observed significant differences in both mental demand and frustration, and we consider the contradiction found here interesting to discuss. The participant thought the task was more complicated for thinking in the AI Coach condition, while they did not need to think much during the video training (***H3***). On the other hand, training with the video condition made them feel more stressed and irritated than training with the proposed system. The fact that the mental demand was high during the practice with the proposed system may be because the participants were required to understand the suggestions given by the system and correct their mistakes within a relatively short time (3 minutes before each system update). On the other hand, the participants had 9 minutes for self-training in the other two conditions where the system was not being updated. However, the later results have shown that the participants felt the least frustrated when training with AI Coach. This may be because the proposed system was showing a simple instruction for the users, and in this way, the users could focus on correcting the suggested posture at the detected timing. Furthermore, most participants reported in the interview that the scoring system encouraged them during the practice. This fact suggested that the system effectively kept the participants' motivation during the training process and prevented them from being frustrated and mistaking the proper way to improve their skills.

Summary

We can summarize the results by confirming the hypothesis raised before the user study:

- The improvement of using the real-time skeleton training method was not significantly greater than using the video & mirror training method; however, the post-study survey revealed that the users preferred to train with the 3D skeleton-based methods (***H1***).
- The participant received a better improvement using the AI Coach system than the conventional video & mirror training method (***H2***).
- Participant was thinking more about correcting their mistakes with the help of AI Coach (***H3***); meanwhile, they felt less stressed using the AI Coach system since they could focus on a specific and straightforward correction.
- The improvement of the latent space distance in the AI Coach condition was significantly greater than in the video condition. Also, participants felt more motivated when training with AI Coach because of the score.

5.6 User Study 2

In the previous study, we compared the proposed system to conventional methods, such as a video replay condition, to assess its effectiveness. However, the proposed system incorporates multiple functions, including error pose visualization and motion synchronization, whose specific contributions to skill development remain unclear. In this study, we aim to investigate in greater depth how these functions—error pose visualization and motion synchronization—support trainees in improving their skills. As in the previous study, we use the AI Coach framework, which comprises the SA-TCC network and the motion visualizer, and employ motion data recorded from an invited trainer serving as the coach.

5.6.1 Hypothesis

In this user study, we examine the individual effects of the main components of our proposed method by formulating the following hypotheses:

***H*4** Aligned (synchronized) motion playback can help users more directly compare their movements to the target motion.

***H*5** The error pose visualization guides users to improve their skills by identifying and recommending a specific timing or frame for correction.

5.6.2 Power Analysis

The sample size for Study 2 was determined based on the results of Study 1. Specifically, Study 1 reported an F-statistic of $F(2, 10) = 3.85$, which corresponds to an estimated effect size of $f = 0.877$. Study 2 also involves three conditions, and a power analysis was conducted using G*Power to determine the minimum number of participants required.

For the power analysis, we specified the following parameters: effect size $f = 0.877$, alpha error probability (α) $= 0.05$, power ($1 - \beta$) $= 0.9$, number of groups $= 1$ (repeated measures design), and number of measurements $= 3$. The G*Power analysis yielded a noncentrality parameter $\lambda = 18.459$, a critical F-value of 5.143, and a total sample size recommendation of 4 participants. The actual power achieved with this sample size was calculated to be 0.841.

To ensure proper counterbalancing across the three conditions and maintain the statistical power of the study, we decided to recruit 6 participants.

5.6.3 Participants

Following approval from the Institutional Review Board, we recruited subjects from a local university population, including both students and faculty members. No incentives or rewards were provided for participation or performance. A total of six male subjects (age range: 20–30) participated in this second user study. All participants were right-handed and had minimal or no prior experience with golf.

5.6.4 Conditions

To validate our hypotheses, we compared our system to a conventional training approach involving video replay, as well as to a baseline method using real-time skeleton visualization. We defined three experimental conditions, each incorporating different feedback mechanisms:

Skeleton Visualization Baseline (Cond. SB) Serving as the baseline condition, this setup is identical to the skeleton visualization used in the previous user study. The system displays the participant's real-time motion and the coach's prerecorded motion side by side using 3D skeletons. A timeline and view-changing features are also provided, mirroring the functionalities offered in the proposed AI Coach system.

Skeleton Visualization with Motion Synchronization (Cond. SS) Building upon the baseline, this condition integrates a motion synchronization function. The user can switch to an alternative view in which the participant's previously recorded motion is aligned frame-by-frame with the coach's motion, enabling direct, synchronized comparison.

Skeleton Visualization with Motion Synchronization and Error Pose (Cond. SE) This condition represents the full proposed system, incorporating both the motion synchronization function and an error pose visualization. In addition to synchronized playback, the system highlights specific frames where the participant's posture deviates from the target posture and recommends precise timing for correction.

5.6.5 Hardware Setups and Procedure

We used the same hardware setups and followed a similar procedure to that in the previous user study. Each experimental session involved the following steps:

1. **Instruction and Warm-Up:** Participants received instructions on how to perform a golf swing, followed by five minutes of free practice.
2. **Pre-training Recording:** Each participant performed five golf swings, which were recorded for later comparison.
3. **Training Conditions:** Each participant experienced three different training conditions (SB, SS, SE), each lasting nine minutes, presented in a counterbalanced order.
    - *Cond. SB*: Baseline skeleton condition
    - *Cond. SS*: Skeleton visualization with motion synchronization
    - *Cond. SE*: Skeleton visualization with motion synchronization and error pose visualization

* For *Cond. SE*, the system updates the error pose visualization every three minutes. System update time is excluded from the nine-minute practice interval.

4. **Post-training Performance:** Participants performed five additional golf swings after each condition, providing "post-training" data.

5. **Questionnaire and Interview:** After completing each training condition, participants filled out a questionnaire, followed by a final post-study survey and interview at the conclusion of the session.

5.6.6 Results

Quantitative Results

We used three quantitative metrics to evaluate improvement in participants' swing performance, comparing the average of five golf swings before and after each condition.

- **Latent Space Distance (LSD):** Represents the distance between the participant's and coach's aligned motions in the latent space. Lower LSD indicates a closer match.
- **Mean Per Joint Position Error (MPJPE):** Measures the positional error for corresponding joints between two aligned motions. Lower MPJPE indicates improved posture alignment.
- **Mean Per Joint Angle Error (MPJAE):** Assesses angular discrepancies between corresponding joints in aligned motions. Lower MPJAE indicates closer replication of the coach's joint rotations.

A higher improvement value on each metric indicates that participants' swings more closely resembled the coach's motion after training. We conducted a one-way ANOVA to test the effects of the three conditions (*Cond. SB*, *Cond. SS*, *Cond. SE*). In cases where the ANOVA indicated significance, we performed Tukey's HSD post-hoc test for pairwise comparisons.

Figure 5.26: Average improvement in each metric after each training condition. LSD: Latent Space Distance; MPJPE: mean per joint position error; MPJAE: mean per joint angle error. Brackets indicate significant pairwise differences ($p < 0.05$). A greater improvement indicates a better learning effect.

Δ LSD

The average initial LSD before training was 4.352. After training:

- *Cond. SB*: 23.22% improvement ($\Delta LSD = 1.010, SE = 0.552$)
- *Cond. SS*: 24.76% improvement ($\Delta LSD = 1.077, SE = 0.476$)
- *Cond. SE*: 27.99% improvement ($\Delta LSD = 1.218, SE = 0.456$)

A one-way ANOVA ($F(2, 10) = 0.037, p = 0.962$) showed no significant differences among the three conditions.

Δ MPJPE

The average initial MPJPE before training was 51.48 mm. After training:

- *Cond. SB*: 12.89% improvement ($\Delta MPJPE = 6.641$ mm, $SE = 2.26$ mm)
- *Cond. SS*: 11.29% improvement ($\Delta MPJPE = 5.81$ mm, $SE = 2.18$ mm)
- *Cond. SE*: 14.25% improvement ($\Delta MPJPE = 7.33$ mm, $SE = 1.57$ mm)

A one-way ANOVA ($F(2, 10) = 0.117, p = 0.89$) revealed no significant differences among the three conditions.

Δ MPJAE

The average initial MPJAE before training was 51.48. After training:

- *Cond. SB*: 19.56% improvement ($\Delta MPJAE = 1.446, SE = 0.538$)
- *Cond. SS*: 17.56% improvement ($\Delta MPJAE = 1.298, SE = 0.411$)
- *Cond. SE*: 20.81% improvement ($\Delta MPJAE = 1.538, SE = 0.423$)

A one-way ANOVA ($F(2, 10) = 0.057, p = 0.944$) again indicated no significant differences.

Usability and Workload Assessments

We administered the System Usability Scale (SUS) [25] to measure overall usability. The SUS scores were:

- *Cond. SB*: 70.833
- *Cond. SS*: 74.583
- *Cond. SE*: 73.333

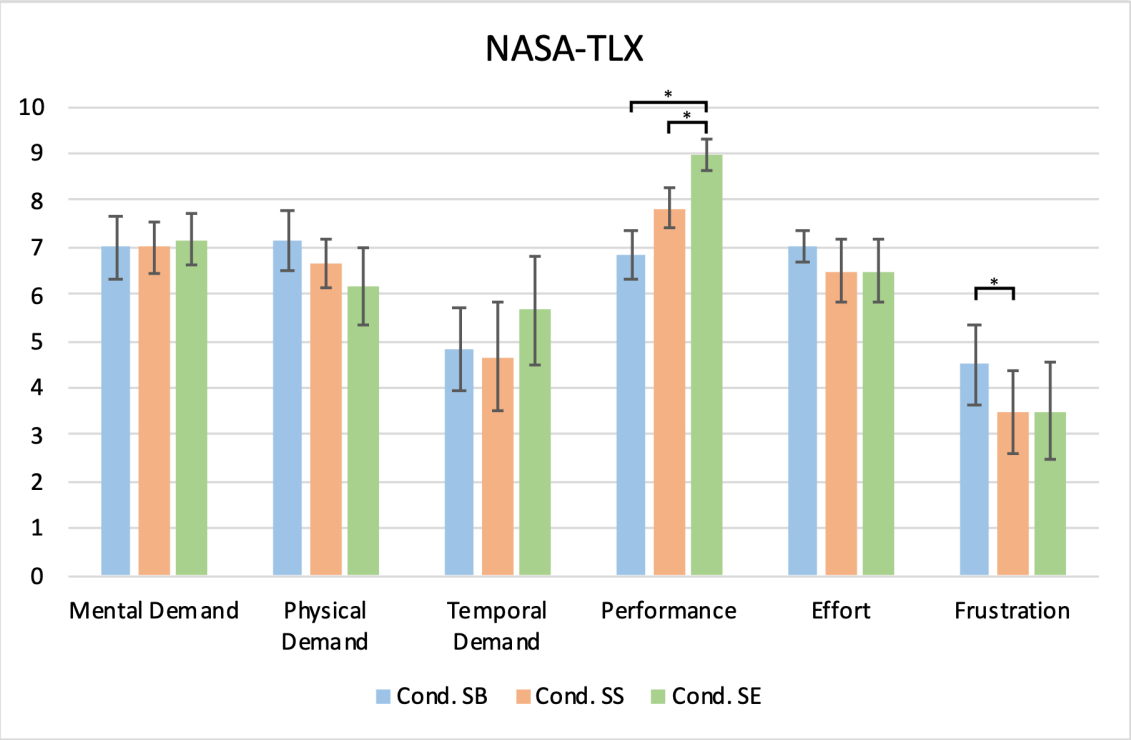

Figure 5.27: Average NASA-TLX scores for each condition. Error bars represent standard error.

All three conditions were rated above average in usability, with *Cond. SS* scoring the highest, closely followed by *Cond. SE*.

To further assess participants' cognitive and physical demands, we used the NASA Task Load Index (NASA-TLX) [58]. A Friedman one-way repeated measures test found no statistically significant differences in mental demand ($\chi^2(2) = 0.105, p = 0.948$), physical demand ($\chi^2(2) = 2.952, p = 0.228$), temporal demand ($\chi^2(2) = 2.375, p = 0.305$), or effort ($\chi^2(2) = 0.666, p = 0.716$). However, there were significant differences in perceived performance ($\chi^2(2) = 9.3636, p = 0.009$) and frustration ($\chi^2(2) = 7.6, p = 0.022$).

Subsequent Wilcoxon signed-rank tests revealed:

- **Performance:** Significant differences between *Cond. SB* and *Cond. SE* ($Z = -2.232, p = 0.026$), and between *Cond. SS* and *Cond. SE* ($Z = -2.121, p = 0.034$). No significance was found between *Cond. SB* and *Cond. SS* ($Z = -1.656, p = 0.098$).
- **Frustration:** Significant difference between *Cond. SB* and *Cond. SS* ($Z = -2.449, p = 0.014$), but not between *Cond. SB* and *Cond. SE* ($Z = -1.857, p = 0.063$) nor between *Cond. SS* and *Cond. SE* ($Z = 0, p = 1.00$).

User Experience Questionnaire Short Version (UEQ-S)

We also employed the User Experience Questionnaire Short Version (UEQ-S) [62] to measure participants' subjective impressions of the three conditions in terms of pragmatic quality, hedonic quality, and overall experience. Table 5.3 summarizes the results:

Table 5.3: UEQ-S scores for the three conditions. "P" = pragmatic quality, "H" = hedonic quality, "C2B" = comparison to benchmark [62].

	Cond. SB			***Cond. SS***			***Cond. SE***		
	M	*SD*	*C2B*	*M*	*SD*	*C2B*	*M*	*SD*	*C2B*
P	1.083	0.801	Below Average	1.416	0.930	Above Average	**1.791**	0.812	**Excellent**
H	-0.666	1.068	Bad	0.583	0.801	Below Average	**1.625**	0.932	**Excellent**
Overall	0.208	0.886	Bad	1	0.832	Above Average	**1.708**	0.692	**Excellent**

Table 5.4: Questions from the post-study survey.

Abbreviation	**Question**
Imitation	I found it easy to imitate the target movements shown in this training method.
Detection	I was able to detect my errors using this training method.
Correction	I was able to correct my errors using this training method.
Efficiency	This training method helped me train efficiently.
Motivation	This training method motivated me to engage more in the training.
Preference	I prefer practicing with this training method.

Cond. SE achieved the highest scores in all three UEQ-S categories, indicating a superior user experience relative to the other conditions.

Post-study Survey

We designed a customized questionnaire to capture participants' perceptions of the training methods. Table 5.4 presents the questions used. Each item was rated on a seven-point Likert scale.

Figure 5.28 displays the average Likert-scale responses. A Wilcoxon signed-rank test revealed significant differences among conditions for multiple questions:

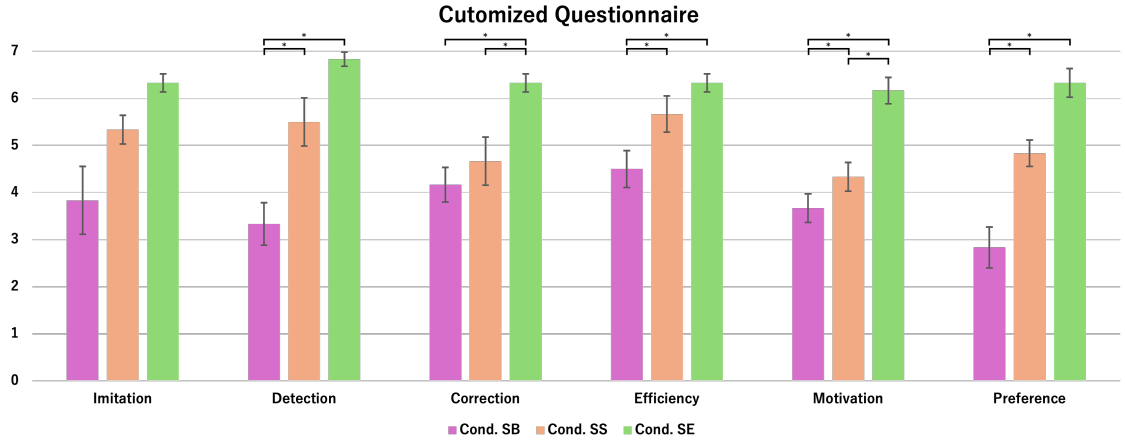

Figure 5.28: Average scores from the post-study survey for each condition. Error bars represent standard error. Brackets indicate significant pairwise differences ($p < 0.05$).

- **Detection:** Participants found it easier to detect errors in both *Cond. SS* and *Cond.*

SE relative to *Cond. SB* (*Cond. SB* vs. *Cond. SS*: $Z = -2.214, p = 0.027$; *Cond. SB* vs. *Cond. SE*: $Z = -2.214, p = 0.027$).

- **Correction:** Participants felt it was easier to correct errors in *Cond. SE* than in *Cond. SB* ($Z = 2.232, p = 0.026$) or *Cond. SS* ($Z = 2.060, p = 0.039$).
- **Efficiency:** Participants perceived both *Cond. SS* and *Cond. SE* to be more efficient compared to *Cond. SB* (*Cond. SB* vs. *Cond. SS*: $Z = -2.121, p = 0.034$; *Cond. SB* vs. *Cond. SE*: $Z = -2.032, p = 0.042$).
- **Motivation:** Participants were most motivated in *Cond. SE* (vs. *Cond. SB*: $Z = 2.232, p = 0.026$; vs. *Cond. SS*: $Z = 2.041, p = 0.041$), and also felt more motivated in *Cond. SS* than in *Cond. SB* ($Z = 2.000, p = 0.046$).
- **Preference:** Participants preferred *Cond. SS* and *Cond. SE* over *Cond. SB* (*Cond. SB* vs. *Cond. SS*: $Z = -2.220, p = 0.026$; *Cond. SB* vs. *Cond. SE*: $Z = -2.226, p = 0.026$).

These outcomes collectively indicate that while the baseline skeleton visualization (*Cond. SB*) provided a solid starting point, adding motion synchronization (*Cond. SS*) or combining synchronization with error pose recommendations (*Cond. SE*) offered more effective tools for detecting and correcting errors, thus enhancing overall efficiency, motivation, and the training experience.

5.6.7 Discussion

In this user study, we investigated how two key components of our AI Coach system—motion synchronization and error pose visualization—influence motor skill training, specifically in the context of golf swings. We compared three conditions: a baseline 3D skeleton playback (*Cond. SB*), a version augmented with synchronized motion playback (*Cond. SS*), and a full system that additionally provides error pose recommendations (*Cond. SE*). Although our quantitative metrics showed no statistically significant differences among the conditions, several important insights emerged regarding user behavior, subjective perceptions, and the overall impact of targeted feedback.

Quantitative Performance Metrics

The objective metrics (ΔLSD, $\Delta MPJPE$, and $\Delta MPJAE$) demonstrated numerical improvements in all three conditions, yet without statistical significance. One reason may be the short practice interval (9 minutes per condition), which could be insufficient for novice learners to internalize and visibly manifest changes in a complex motion like a golf swing. Additionally, our relatively small sample size (six participants) limited the statistical power. Even so, the slight trend favoring *Cond. SE* suggests that synchronized

playback *plus* targeted error feedback may hold promise for longer-term or more extensive training.

Although all conditions produced similar Δ results objectively, participants' subjective evaluations exhibited significant differences. This discrepancy can occur because performance-based metrics (e.g., ΔLSD or $\Delta MPJPE$) do not always mirror learners' *perceptions* of improvement. For instance, a relatively brief session may not produce large, measurable changes in swing form, especially if participants carry over knowledge or improved understanding from one condition to another in quick succession. Meanwhile, the user survey directly captures each system's features and ease of use; it more readily distinguishes which functions (e.g., error pose highlights) felt most helpful or motivational. In essence, the Δ metrics remain stable across conditions, but the *user experience* can still differ significantly, leading to notable contrasts in subjective ratings despite only minimal objective deltas.

Observed Usage Behaviors and User Preferences

Interestingly, participants reported that they did not frequently utilize the synchronized playback feature, even though it was available in *Cond. SS* and *Cond. SE*. Instead, they preferred the simpler side-by-side view of the target motion and their current real-time motion, which is essentially the *Cond. SB*-style interface. Some participants mentioned that it was difficult to pinpoint specific posture corrections by observing two playback videos (target versus their past motion), presumably because comparing full-body 3D skeletons, even when synchronized, introduced considerable cognitive load.

Error Pose Visualization as a Targeted Cue

By contrast, participants in *Cond. SE* actively used the error pose function. This feature highlights a single frame with an overlaid target-versus-user posture and specifies a "recommended timing" to correct their form. This approach appears to reduce cognitive demands by isolating the critical instant for correction rather than requiring users to infer it from continuous playback. Many participants practiced precisely that frame or timing, indicating that concise, frame-specific guidance helped them more easily locate and address their errors.

These findings suggest that while theoretically beneficial, advanced features like motion synchronization might be too complex or vague without additional guidance. In contrast, a specifically highlighted error pose offers a straightforward, actionable anchor point. This aligns with the broader observation that novices often benefit most from clear, incremental cues that directly map onto the corrections they need to make.

Subjective Usability and Workload

Although objective performance metrics did not differ substantially, participants' subjective responses in usability (SUS) and workload (NASA-TLX) offer valuable perspectives on how well the system features matched their expectations and training needs. All three conditions attained above-average SUS scores, indicating the baseline 3D skeleton visualization was inherently understandable and easy to operate.

Interestingly, *Cond. SS* achieved the highest SUS score, possibly reflecting that participants found synchronized playback *conceptually* intuitive (even if they did not use it heavily in practice) or appreciated having the *option* to compare motions if they wished. The NASA-TLX outcomes showed that mental, physical, and temporal demands did not significantly differ among conditions. However, participants reported higher perceived performance under *Cond. SE*, suggesting that the error pose function—albeit used intermittently—was particularly effective in boosting their confidence and sense of progress.

User Experience and Motivational Factors

Participants' positive reception of *Cond. SE* was also reflected in the UEQ-S questionnaire, where *Cond. SE* scored highest in pragmatic quality, hedonic quality, and overall user experience. The post-study survey further showed that participants felt more motivated, found it easier to detect and correct errors, and perceived their training to be more efficient when error pose recommendations were available. These findings highlight the importance of providing targeted, context-specific feedback that learners can quickly translate into action. The single-frame error pose offered a clear "anchor" for learners to see precisely what to fix and when, thereby reducing confusion and enhancing motivation.

Reconciling Objective Findings with Usage Patterns

The discrepancy between the limited adoption of synchronized playback and participants' positive impressions of the broader system suggests a crucial design takeaway: more detailed or technically advanced feedback is not always *better* if it lacks a clear, intuitive path for correction. Even though the synchronization feature had potential, its value might be fully realized only after learners progress beyond the novice stage or are given additional guidance to interpret the side-by-side skeletons. The trend favoring *Cond. SE* in objective metrics (albeit non-significantly) could partially be attributed to the occasional use of synchronization in tandem with the error pose marker.

Implications for Training System Design

From a practical perspective, our results underscore that feedback should be delivered in a way that clearly pinpoints actionable moments or frames for users. While synchronization offers a sophisticated visualization, its cognitive demands may outweigh its usefulness for novices. The single-frame error pose recommendation in *Cond. SE* epitomizes how

a streamlined, explicit cue can lead to stronger engagement and easier skill correction. Thus, for systems aiming to support novice motor training, it may be more beneficial to emphasize concise, context-specific feedback rather than rely on more elaborate but less straightforward modes.

Summary

In summary, the study shows that combining basic 3D skeleton visualization with targeted error pose feedback can significantly improve subjective learning experiences, motivation, and ease of error detection/correction. Although we did not observe statistically significant objective differences among the three conditions within the short training window, participants indicated that the error pose recommendations were particularly valuable. Meanwhile, synchronized playback was theoretically appealing but underutilized in practice. These insights collectively suggest that future AI coaching systems should balance complexity with clarity, prioritizing features that deliver actionable, frame-specific cues to help novice learners make tangible, confident strides in skill development.

5.7 User Study 3

This section details the design and methodology of our third user study, which involved a professional golf coach evaluating the effectiveness and validity of the AI coach system. The study aimed to assess the system's ability to identify critical timings, detect discrepancies, and improve specific phases and overall swing forms through targeted training. By comparing the AI's outputs with expert evaluations, this study addresses the subjective aspects of motion analysis that are often overlooked in automated systems.

5.7.1 Hypotheses

The study was guided by three hypotheses:

1. **H6:** The AI coach system's assessment of the "most critical timing" in a swing can be validated by a human professional coach's feedback. Specifically, we hypothesize that while there may be differences in how the system and the coach identify which phase to prioritize, there will be a correlation between the system's and the coach's judgments.

2. **H7:** The AI coach system's assessment of discrepancies between a beginner's motion and a target swing (advanced golfer) will be consistent with the coach's scoring of how different the two motions look in each phase.

3. **H8:** Training guided by the AI coach system (focusing on specific phases identified by the system) will result in noticeable improvement in both the specific phase trained and the golfer's overall swing form, as judged by the professional coach.

5.7.2 Participants and Setup

The study involved a single professional golf coach with over a decade of experience teaching both amateurs and professionals. The swings analyzed during the study were from five beginner golfers who had previously practiced with the AI coach system in Study 2.

The study employed a dual-device setup as shown in Figure 5.29. A touchscreen monitor displayed swing videos alongside their corresponding 3D skeleton visualizations, providing an interactive environment for detailed motion analysis. An iPad was used for completing questionnaires to record the coach's assessments. The swing data included 2D video and 3D skeleton views for both the beginner and advanced golfers.

A single advanced golfer's swing was used as the reference for discrepancy evaluations, ensuring consistency across all comparisons.

5.7.3 Procedure

The study was structured into three tasks: (1) identifying the most critical timing, (2) evaluating discrepancies, and (3) assessing training validity. All questionnaire responses

Figure 5.29: Study 3 setup. The professional golf coach use a touchscreen to evaluate the tasks and answer the questionnaires with an iPad.

were collected via an iPad application. Each response was recorded and associated with the corresponding swing ID. The swing videos and skeleton animations were displayed on the touchscreen to allow for interactive viewing (e.g., pausing, replaying key frames).

Task 1: Identifying the Most Critical Timing

In this task, the coach evaluated each beginner's swing (video and skeleton views) to determine the phase that required the highest priority for correction (Figure 5.30). The seven swing phases under evaluation were address, backswing, top, downswing, impact, follow-through, and finish.

The coach assigned rankings from 1 (least important) to 5 (most important) to each phase, allowing duplicate rankings to reflect equal importance. For each swing, the AI coach system's predicted critical timing was compared with the coach's rankings. A total of 20 swing evaluations (four swings per beginner) were conducted to investigate the alignment between the coach's judgment and the AI's predictions.

Task 2: Evaluating Discrepancies

As shown in Figure 5.31, the coach compared the beginner's swing to the target advanced golfer's swing, presented side-by-side with synchronized video and skeleton views. For each of the seven swing phases, the coach rated the magnitude of discrepancy on a scale from 1 (small discrepancy) to 7 (large discrepancy).

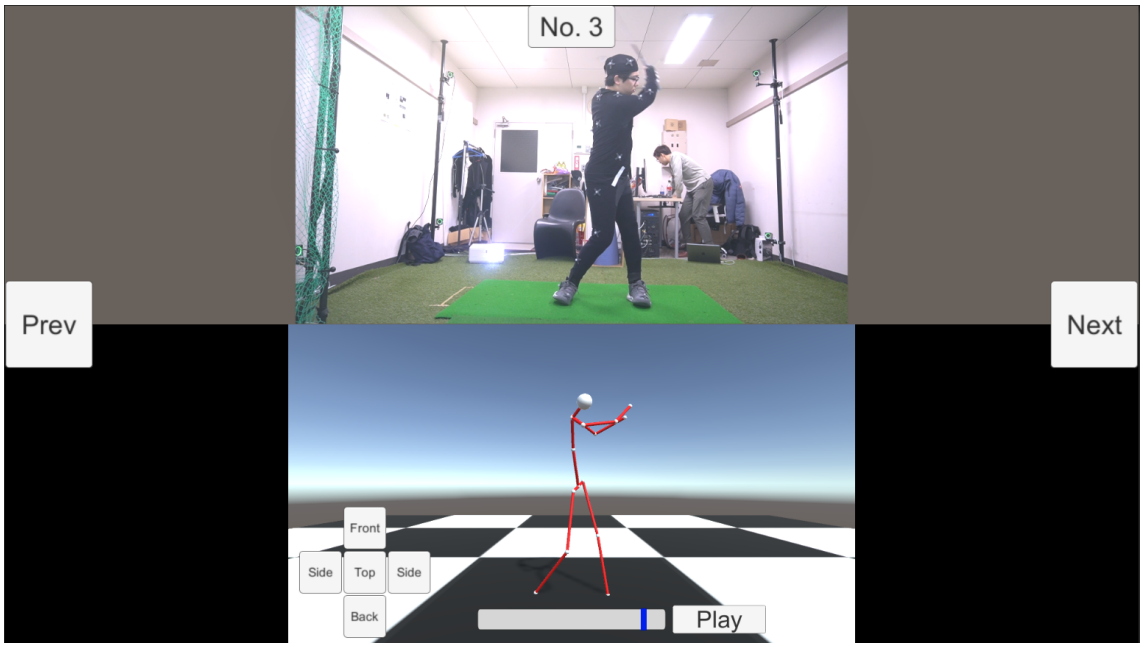

Figure 5.30: Task 1: Identifying the most critical timing in a single swing.

The AI coach system's latent-space discrepancy scores for each phase were compared with the coach's ratings. A total of 20 comparisons were performed, providing insight into the agreement between the AI's quantitative measures and the coach's subjective evaluations.

Task 3: Assessing Training Validity

In the final task, the coach assessed the effectiveness of AI-guided training. Each beginner's pre- and post-training swings were compared to evaluate improvements (Figure 5.32). Two questions were posed: 1. Did the targeted phase improve after training? 2. Did the overall swing form improve after training?

1. *Q1: Has the specific phase (e.g., finish) improved after training?*
2. *Q2: Has the overall swing form improved?*

The coach provided ratings for both questions on a scale from 1 (became worse) to 7 (significant improvement). A total of 15 evaluations (three per beginner) were conducted to determine whether the AI's recommendations led to meaningful improvements in both specific phases and the overall swing.

5.7.4 Analysis

The data consisted of the coach's questionnaire responses recorded for each swing. The analysis focused as below, which will be further dicsussed in the following sections:

1. **Compare Rankings (Task 1):** Evaluate the correlation between the coach's rankings (1–5) and the AI coach system's predicted critical phase. We will use measures

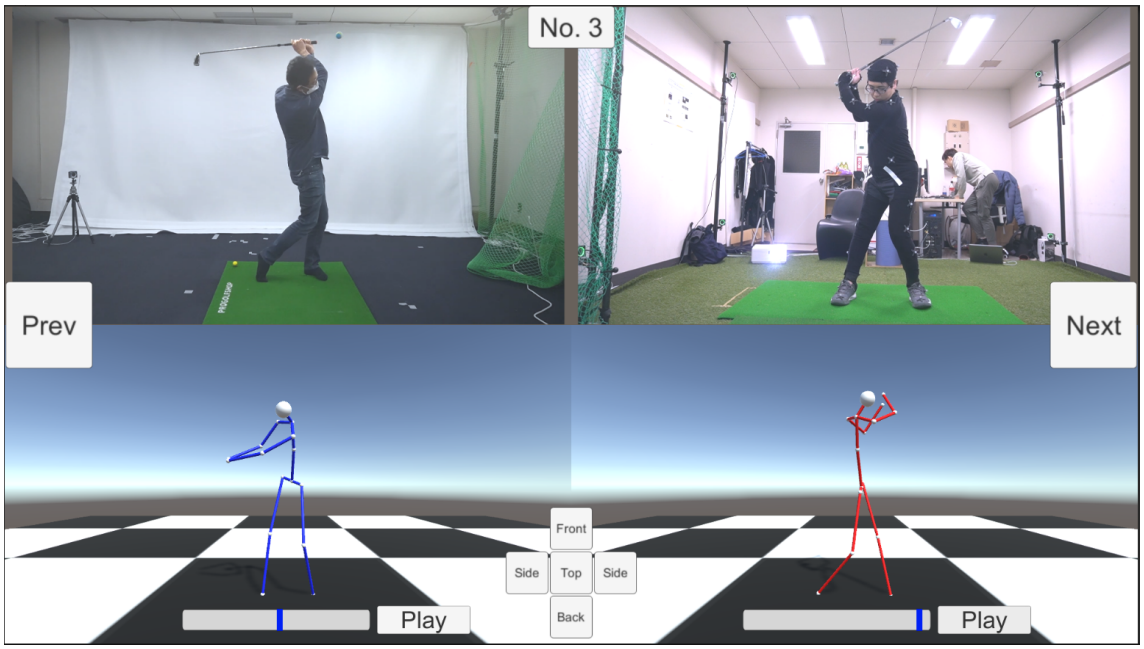

Figure 5.31: Task 2: Comparing the participant's swing aginst the target advanced golfer's swing.

such as Spearman's rank correlation.

2. **Compare Discrepancy Scores (Task 2):** Compute agreement measures (e.g., Pearson correlation) between the coach's discrepancy scores (1–7) and the AI-calculated discrepancy in the latent space.

3. **Assess Training Improvements (Task 3):** Conduct statistical tests (e.g., paired t-tests, Wilcoxon signed-rank tests) on the coach's responses to see if the improvements are perceived as significant in the specific phase and overall form.

5.7.5 Results

Task 1: Identifying the Most Critical Timing

In **Task 1**, the professional coach and our AI system each provided a rating of how "critical" each of the seven swing phases (address, backswing, top, downswing, impact, follow, finish) was for 20 distinct swings. The coach assigned an integer score from 1 to 5 for each phase, where 5 indicated "most important to correct." Meanwhile, the AI system produced continuous float scores for each phase, representing how drastically that phase deviated from an advanced reference swing in its latent-space analysis. To facilitate a direct comparison with the coach's 1–5 scale, we first normalized the AI's per-swing float scores into the range $[0, 5]$, then applied a ceiling operation ($\lceil \cdot \rceil$) so that any value above 4.0 became 5, aligning with the coach's discrete scale.

After this scaling, both the coach and the AI had integer (1–5) ratings for each phase of every swing. Figure 5.33 illustrates how frequently each phase was assigned the highest

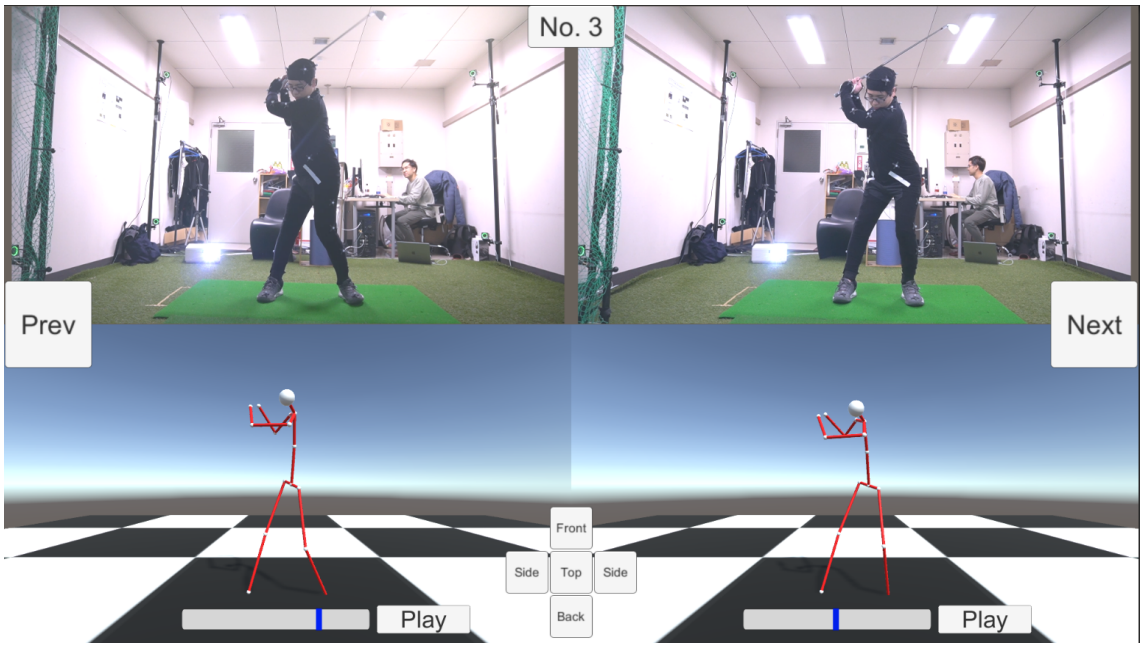

Figure 5.32: Task 3: Comparing pre- and post-training swings.

rating (5) by the coach vs. the AI in this *unweighted* scenario. The coach consistently emphasized the setup and early mechanics of the swing, marking *address*, *backswing*, and *downswing* with a 5 in 12 out of 20 swings each, while top, impact, follow, and finish seldom received such a priority. Conversely, the AI highlighted phases like *finish* (16 times) and *follow* (12 times) as most critical, rarely giving top priority to address or backswing. Only 25% of the swings displayed any overlap in which phase was labeled a 5 by both the coach and the AI, indicating a notable mismatch between the coach's instructional priorities (focusing on early mechanics) and the AI's purely geometric viewpoint (often flagging later, visually conspicuous deviations).

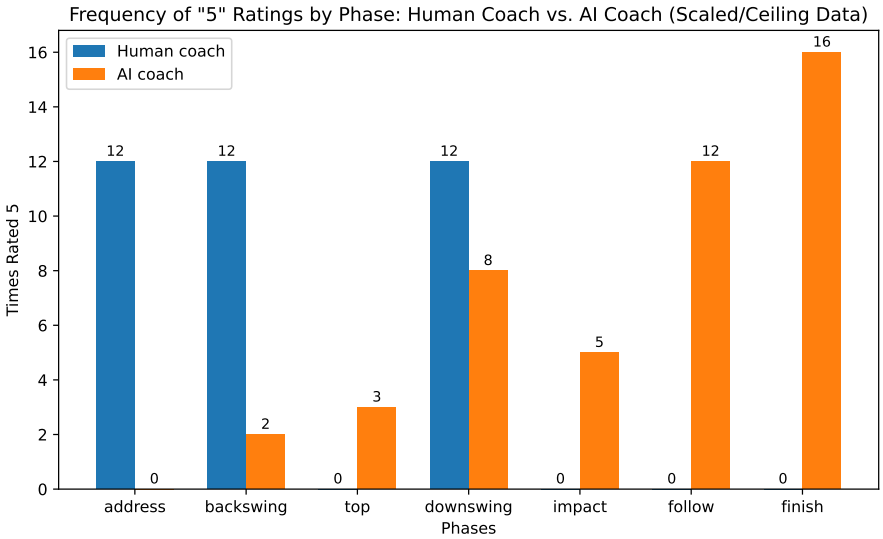

Figure 5.33: Comparison of "5" ratings (most critical) per phase from the coach and the AI's unweighted approach.

To address this discrepancy, we introduced a *hybrid-weighted* approach that multiplied each phase's raw AI score by a weight derived from the coach's Task 1 ratings for that phase. The result was then normalized and scaled back to 1–5, yielding a new distribution of top-phase ratings that incorporated both geometric deviation (the AI's strength) and professional emphasis (the coach's strength). Figure 5.34 shows how many times each phase received a 5 under this hybrid approach, plotted alongside the coach's original counts.

In this updated version, *downswing* became the universally top-rated phase, appearing with a 5 in all 20 swings, sometimes joined by *backswing* (7 times). The previously dominant finishing phases (finish, follow) faded, rarely or never reaching the top rating, indicating that domain knowledge significantly reoriented the AI's raw geometric scores. However, *address*, despite the coach's strong emphasis in earlier data, still never achieved a 5 under this weighting, suggesting that further calibrations (e.g., higher multipliers for address) or causal modeling (e.g., highlighting how early-phase errors can propagate) may be needed.

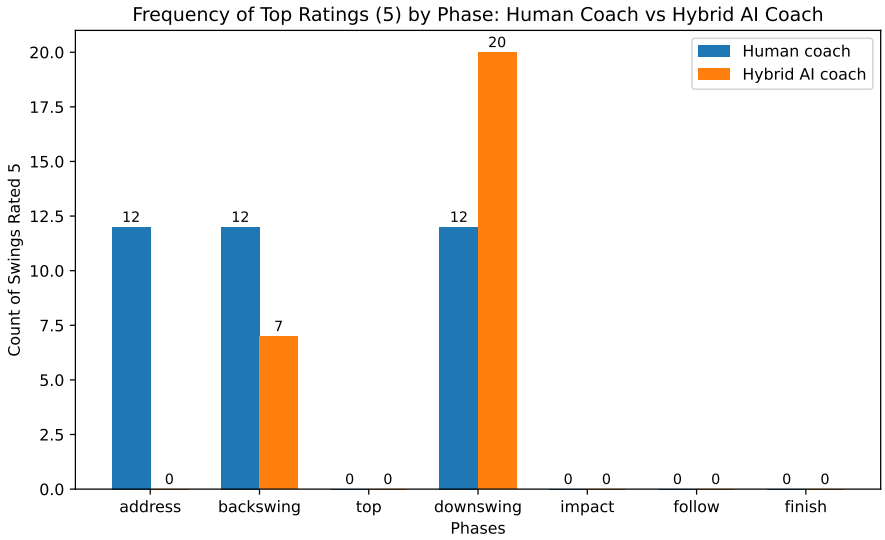

Figure 5.34: Comparison of "5" ratings (most critical) per phase from the coach and the hybrid-weighted AI approach.

In summary, Task 1 demonstrates how purely unweighted, geometry-based AI scores can diverge from a professional coach's longstanding instructional focus on early phases, but also shows that incorporating phase-level weights from the coach can partially reconcile these differences. While the hybrid method successfully shifts emphasis toward backswing and downswing, it does not yet capture every nuance of the coach's domain expertise, particularly the repeated stress on address as a foundational phase. Nonetheless, this progress highlights how merging data-driven discrepancy detection with an expert's phase priorities yields a more balanced ranking of which parts of the swing deserve immediate attention.

Task 2: Discrepancy Comparison

In **Task 2**, the professional coach and our AI system each provided estimates of how different a beginner's swing was from a target (advanced) swing across seven phases. The coach rated the discrepancy on a 1–7 scale for each phase (7 = large discrepancy), while the AI produced a float score indicating the latent-space distance between the two motions. Unlike Task 1, we did not transform the AI's scores. Instead, we compared the coach's raw 1–7 ratings to the AI's raw floats on a per-phase and per-swing basis:

- **Coach**: For each of the 20 swings, we identified the phase(s) in which the coach's 1–7 rating was *highest* (this highest score could be 7, 6, 4, etc., depending on the swing). We refer to this set as *CoachMax*.
- **AI**: For each swing, we located the phase with the single *largest* float value (if there had been ties, we would have included all tied phases). We label this set *AImax*.

By comparing *CoachMax* and *AImax* on a swing-by-swing basis, we could see whether both the coach and AI singled out any of the same phases as "most discrepant." We then compiled the frequency with which each phase appeared as top for the coach versus top for the AI.

Phase Frequencies

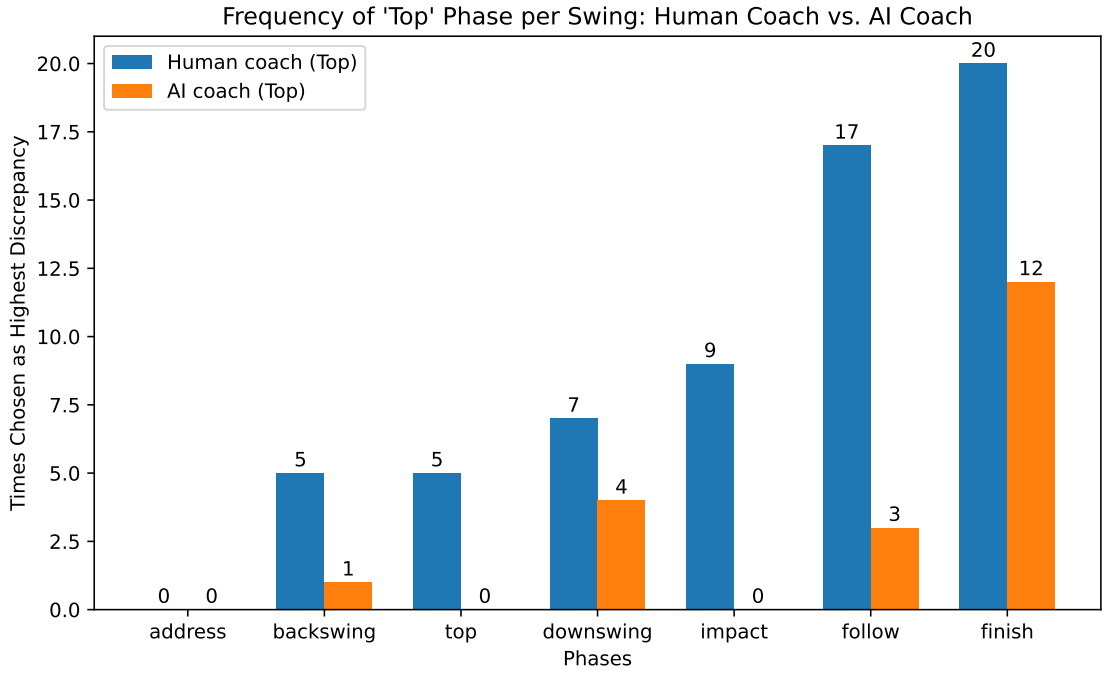

Figure 5.35: Comparison of the top rate (most discrepant) per phase from the coach and the AI after scaling the AI's data.

Firstly, as shown in Figure 5.35, looking at all 20 swings, we aggregated how often each phase was in *CoachMax*. Table 5.5 (column "Coach") shows the total counts:

- *finish* appeared as the coach's highest (or tied highest) in all 20 swings,
- *follow* was next (17 times),
- *impact* (9), *downswing* (7), *backswing* (5), *top* (5),
- *address* never reached the coach's highest rating.

Secondly, for each swing, the AI provided a single maximum float score. Table 5.5 (column "AI") summarizes:

- *finish* was the AI's top 12 times,
- *downswing* appeared 4 times,
- *follow* 3 times,
- *backswing* 1 time,
- *address*, *top*, and *impact* never emerged as the single highest.

Table 5.5: Number of swings (out of 20) in which each phase was the coach's top or the AI's top discrepancy.

Phase	**Human Coach**	**AI Coach**
address	0	0
backswing	5	1
top	5	0
downswing	7	4
impact	9	0
follow	17	3
finish	20	12

Overlap of Most Discrepant Phases

For each swing, we checked if

$$\text{CoachMax}(\textit{swing}) \cap \text{AImax}(\textit{swing}) \neq \varnothing.$$

We found that in **16 out of 20** swings, the coach and AI shared at least one top phase. In most cases, *finish* or *follow* appeared in both the coach's and AI's top sets. However, in 4 swings (#6, #10, #17, and #19), the coach chose a different highest phase than the AI, yielding no overlap.

Spearman Correlation on Flattened Data

To obtain an overall measure of how consistently the coach's 1–7 scores track the AI's continuous discrepancies:

1. We flattened the data across all **20 swings** and **7 phases** into two 1D arrays of length 140:
    - x: coach's numeric ratings (1–7),
    - y: AI's float scores.
2. We then computed a global Spearman rank correlation, yielding

$$\rho \approx 0.603, \quad p \approx 3.16 \times 10^{-15}.$$

This indicates a statistically significant, moderately strong *monotonic* relationship between the coach's discrepancy ratings and the AI's latent-space distances across all (phase, swing) pairs. While each swing has multiple phases (and thus the data are not fully independent), this correlation *does* provide an overarching sense that higher coach ratings typically align with higher AI scores.

In summary, Task 2 highlights a considerable alignment between the AI's latent-space discrepancy measurements and the coach's subjective evaluations of motion discrepancies across swing phases. The phase *finish* consistently emerged as the top-rated discrepancy for both the coach (20 swings) and the AI (12 swings), with moderate agreement on *follow* and *downswing*. Out of 20 swings, 16 showed at least one overlapping phase between the coach's highest discrepancy ratings and the AI's top-scoring phases, indicating strong alignment in the identification of critical discrepancies.

The overall correlation analysis further supports this alignment. A global Spearman rank correlation of approximately $\rho = 0.60$ (with $p \approx 3.16 \times 10^{-15}$) demonstrates a statistically significant, moderately strong relationship between the coach's 1–7 ratings and the AI's continuous scores. This suggests that the AI's data-driven approach reliably captures many of the same kinematic or visual discrepancies identified by an expert human evaluator.

Despite this agreement, notable differences remain. For instance, the AI occasionally focused on *downswing* or other phases, whereas the coach emphasized *finish* and *impact*. These differences may reflect variations in how the AI quantifies geometric deviations versus how the coach prioritizes discrepancies based on instructional principles or perceived impact on overall performance. Nevertheless, the results affirm the AI's potential as a complementary tool for detecting discrepancies, aligning well with human evaluations while offering unique, data-driven insights.

Task 3: Evaluating Training Validity

In **Task 3**, we investigated whether focusing on a specific, AI-recommended phase led to tangible improvements in that phase and in the overall swing form, as assessed by the professional coach. Each improvement was rated on a 1–7 scale (1 = worse, 4 = no change, 7 = strong improvement). Table 5.6 presents 15 evaluations reorganized by participant

Table 5.6: Collected improvement scores for Task 3. Each row shows (1) which phase the beginner practiced, (2) the coach's rating of that phase's improvement (1–7), (3) the coach's rating of the overall swing's improvement (1–7).

#	Practiced Phase	Phase Improved	Overall Improved
P1	top	3	3
P1	follow	2	2
P1	top	5	5
P2	finish	2	2
P2	finish	6	6
P2	follow	2	2
P3	top	3	3
P3	finish	5	5
P3	top	4	4
P4	follow	2	2
P4	top	5	5
P4	finish	4	4
P5	backswing	4	4
P5	finish	4	4
P5	backswing	4	4

(P1 through P5). Each row indicates which phase was practiced, the coach's rating for that phase's improvement, and the coach's rating for the entire swing.

Data Overview and Distribution. The table shows that the coach assigned an identical score to both the targeted phase and the overall swing in all instances, reflecting a holistic perspective in which changes in a specific segment are seen as directly influencing the entire motion. No entries received ratings of 1 or 7. Instead, improvements spanned the range of 2 (slight regression) to 6 (moderate positive change). Four of the 15 entries were rated 2, two entries were rated 3, five entries were rated 4, three entries were rated 5, and only one was rated 6. This distribution yields a mean score of approximately 3.67 and a median of 4, suggesting that the short training sessions largely resulted in little or no net improvement, with occasional moderate gains.

Participant-Specific Observations. Participant P1 alternated between practicing *top* and *follow*, initially seeing little improvement (ratings of 3 and 2) but later achieving a 5 on top, indicating moderate progress. Participant P2 focused on *finish* and *follow*, showing a jump from 2 to 6 on finish, yet remaining at 2 on follow. This fluctuation suggests that while participants can achieve one-off successes, they may not consistently carry them over to other phases or repeated attempts. Participant P3 trained top and finish with ratings ranging from 3 to 5 and 4, reflecting minor but inconsistent improvements. Participant P4 similarly targeted follow, top, and finish phases, registering mostly minimal gains (2 or 4), though top received a moderate 5. Finally, Participant P5 concentrated on backswing and finish, with all entries plateauing at 4 ("no change"), implying that deeply rooted

mechanics may require longer or more frequent practice to see visible progress.

Interpretation of Short-Term Gains. Overall, these results highlight the difficulty of effecting notable changes in complex motor skills like the golf swing over brief sessions. Although participants occasionally reached moderate improvement (5 or 6) for a specific phase, many attempts remained near 2–4, signifying no net gain or even regression. The uniform rating for phase and overall form in each row further indicates the coach's view that targeted corrections immediately affect the entire swing, yet the average participant struggled to transform short bursts of improvement into sustained benefits. In conjunction with the findings from Tasks 1 and 2, this pattern supports the need for extended practice intervals or repeated feedback loops to solidify the brief gains seen here. It also underscores how combining AI-based cues with domain knowledge—particularly for early-phase fundamentals—might yield more consistent progress if given ample time and reinforcement.

5.7.6 Discussion

This section synthesizes the findings from our three main tasks, emphasizing how the AI coach system compares to the human coach's priorities, how effectively it detects discrepancies, and whether its recommendations ultimately lead to meaningful improvements in short-term training. The results clarify both the strengths and the limitations of a purely data-driven approach, suggesting avenues for refining the system's emphasis and feedback mechanisms.

Differences in Correction Priorities (Task 1)

Task 1 initially revealed a stark mismatch between the professional coach's focus on early mechanics (address, backswing, downswing) and the AI's emphasis on visually prominent deviations in later phases (finish, follow). While the coach rated address, backswing, and downswing as "most important to correct" in 12 out of 20 swings each, the AI system (when scaled to a 1–5 range) singled out finish 16 times and follow 12 times, rarely elevating address or backswing to top priority. This discrepancy underscores how a purely geometric approach can overlook the causal importance of early-swing fundamentals.

To address this gap, we introduced a *hybrid* scoring model that reweighted each phase's AI score by a factor derived from the coach's Task 1 ratings. The revised distribution shifted the AI's highest ratings toward downswing (and sometimes backswing), aligning more closely with the coach's early-phase emphasis. However, address still never achieved the top rating, hinting that further or more causal weighting may be needed to fully capture the coach's repeated stress on setup mechanics. Overall, Task 1 demonstrated that while domain weighting can partially reconcile geometric discrepancy detection with expert insights, purely data-driven methods alone are unlikely to prioritize foundational

phases in the same way a professional coach would.

Agreement in Discrepancy Detection (Task 2)

In Task 2, we evaluated how closely the AI's float-based discrepancy scores tracked the coach's 1–7 discrepancy ratings across each swing phase. A moderate-to-strong Spearman correlation (approximately 0.60) showed that the AI effectively identified many of the same kinematic differences that the coach perceived, validating the fundamental latent-space approach. Sixteen of the 20 swings also shared at least one phase deemed "most discrepant" by both the coach and the AI. Nevertheless, mismatches arose when the coach considered certain phases (like impact or follow) highly dissimilar, but the AI did not assign them similarly large float values. These inconsistencies echo Task 1's findings, affirming that geometric detection can miss the strategic significance of certain phases. Future refinements—potentially integrating domain knowledge, biomechanical heuristics, or coach-driven weighting—might further align AI and expert evaluations.

Training Validity and Observed Improvements (Task 3)

Task 3 explored whether short training sessions focusing on AI-recommended phases produced visible improvements, as judged by the coach. We reorganized the 15 evaluations by participant (P1–P5), each practicing various phases like top, finish, follow, or backswing. The coach invariably assigned the same rating to both the targeted phase and the overall swing, reflecting a holistic stance that local corrections affect the entire motion. These ratings ranged from 2 (slight regression) to 6 (moderate improvement), with a mean around 3.7 and median of 4, indicating little net gain in most sessions. A few participants displayed brief successes (scores of 5 or 6 for one phase) but often reverted to neutral or regressive scores (2–4) in subsequent trials, suggesting that the limited practice intervals did not suffice to stabilize newly learned mechanics. Instances of repeated "4" (no change) in early fundamentals, such as P5's backswing, underline how altering deep-seated motions may require more extended or repeated feedback loops.

Collectively, these patterns point to two main conclusions. First, although the AI can occasionally prompt moderate improvements (especially on phases that are easier to adjust quickly), short sessions without repeated reinforcement rarely yield lasting gains. Second, the coach's holistic assessment—where any improvement in one targeted phase should raise the overall swing—did not materialize into consistent progress. Longer training interventions or real-time guidance might be necessary to capitalize on the AI's cues.

Overall Synthesis and Future Directions

Taken together, Tasks 1 and 2 confirm that the AI coach system can detect many of the same pose deviations as the professional coach, yet it may overemphasize later phases in its

unweighted form. The hybrid approach shows promise for rebalancing scores toward foundational mechanics, though address still remained underprioritized. Meanwhile, Task 3's modest or inconsistent short-term improvements highlight the importance of practice time in embedding subtle motor changes. These insights collectively suggest that purely geometric methods should be expanded with domain-specific weighting, causal models, or multi-session reinforcement to produce outcomes that align more closely with professional coaching philosophies. Future investigations could incorporate longer-term training trials, additional performance metrics (e.g., ball flight), and real-time corrections to determine whether learners who receive repeated, targeted feedback based on both AI detection and coach weighting can achieve more stable and durable skill improvements.

5.8 Discussion and Future Work

In this section, we provide an integrated discussion of our three user studies (Study 1, Study 2, and Study 3), followed by a summary of limitations in Section 5.8.3 and potential future applications in Section 5.8.4.

5.8.1 Overall Discussion

Our research aimed to design and evaluate an *AI coach* system for golf-swing training, progressively refining both the system's feedback methods and evaluation approaches through three sequential studies.

Study 1

In Study 1, we explored how beginners respond to three different conditions: conventional video and mirror training, real-time skeleton visualization, and an early prototype of the AI Coach equipped with error poses and a latent-space scoring mechanism. The results indicated that video-based training alone proved minimally effective, as participants struggled to detect posture discrepancies, particularly when relying on a single, fixed camera angle. Although the 3D skeleton visualization offered a more intuitive interface and garnered greater user preference, participants in this condition found it difficult to determine precisely where and how to correct their motions. In contrast, the AI Coach approach—featuring explicit error poses to highlight the most pronounced discrepancies and an engaging scoring system—outperformed both video and skeleton-based methods in terms of quantitative improvements (e.g., reduced joint-angle errors, lower latent-space distances) and subjective user responses. These findings suggested that targeted guidance, combined with a motivating feedback mechanism, can significantly enhance self-training for novice golfers who lack regular access to professional coaching.

Study 2

Building on the insights from Study 1, Study 2 examined two specific features of the AI Coach system: motion synchronization (aligning the user's swing in real time with a reference swing) and error pose visualization (identifying a single frame at which the user's motion is most discrepant). A baseline 3D skeleton playback served as the control condition. The study found that participants seldom relied on the synchronization feature, possibly because comparing two fully animated skeletons demanded considerable cognitive effort. By contrast, the error pose feature was actively used and strongly favored, presumably because it pinpointed the exact moment requiring correction. Although overall objective improvements were not statistically significant, likely due to limited practice times and a small sample size, subjective feedback showed that providing concise, frame-specific cues was seen as more motivating and practically beneficial. These outcomes led us

to conclude that brief, focused feedback may be more valuable than complex, demanding features such as fully synchronized side-by-side playback.

Study 3

In **Study 3**, we investigated whether the AI's discrepancy detection aligns with a professional coach's subjective assessments and whether AI-suggested phases yield short-term training improvements. The results showed a moderate-to-strong correlation between the AI's latent-space scores and the coach's 1–7 discrepancy ratings, indicating that the AI successfully identified many of the same visually or kinematically significant errors. However, as in Task 1, the AI and coach diverged in *which* phases they deemed most critical to correct. The AI tended to emphasize visually prominent later phases (e.g., finish, follow), whereas the coach prioritized earlier mechanics (e.g., address, backswing).

To address this mismatch, we introduced a hybrid approach in Task 1 that multiplies each phase's raw AI score by weights derived from the coach's ratings. This revised method shifted the AI's highest scores largely toward downswing and occasionally backswing, bringing its selections closer to the coach's emphasis but still overlooking address. Thus, while domain-driven weighting reduced the AI's earlier bias toward finish, it did not fully capture the coach's holistic priorities.

In evaluating training effectiveness, we reorganized 15 brief training sessions by participant (P1–P5), each focusing on AI-recommended phases such as top, finish, follow, or backswing. The coach's identical rating for phase-specific and overall swing improvement suggests a belief that fixing one key phase should elevate the entire motion. Yet most scores remained near 3 or 4, signifying minimal net gains, with only occasional moderate improvements (5–6). In some instances, participants regressed (2–3) on subsequent tries. These modest, inconsistent outcomes likely reflect the short time frame of practice, insufficient for novices to adopt and retain meaningful changes to complex motor patterns. Extended interventions and repeated feedback loops, potentially combining the hybrid weighting with real-time or multi-session updates, may be necessary to ingrain the incremental successes observed in some individual trials.

5.8.2 Whole-Motion Imitation versus Specific "Key Tips"

While AI Coach generally takes the stance that imitating an overall, coherent swing pattern leads to effective skill improvement, one may ask whether it would be more beneficial to focus on only a few critical "key tips"—for instance, a specific wrist angle at impact, or hip rotation during the downswing. In other words, is the system missing an opportunity to isolate "the one crucial point" that, if corrected, quickly elevates a beginner's performance?

Our approach focuses on the entire motion because complex tasks, like a golf swing, involve multiple interdependent phases (address, backswing, impact, follow-through, and finish). A local correction (e.g., only adjusting wrist angles) may fail to address inter-phase

dependencies; subtle errors may propagate across consecutive frames. Hence, AI Coach synchronizes entire swing sequences with a reference and detects discrepancies across all phases. By examining the motion globally, we provide consistent feedback that respects biomechanical dependencies.

Although the system tracks the entire sequence, it does not ignore important localized clues. The attention module within AI Coach dynamically emphasizes critical frames or joints where a user deviates most significantly. Thus, while the system does not force an exclusive, single "tip" correction, it effectively weights and underscores particularly problematic areas. This enables the user to see which sub-actions matter most in real time (e.g., a breakdown at impact or insufficient hip turn at address).

Nothing in our framework precludes incorporating formalized coaching tips or domain-specific heuristics (e.g., ensuring a certain spine angle, recommended torque transfer, or stance alignment). In fact, such knowledge could refine the model's emphasis. If future versions of AI Coach incorporate explicit domain knowledge on essential "key tips," the system may further isolate and strengthen those single critical corrections while still retaining the broader, whole-motion synergy.

Overall, AI Coach aims to preserve a holistic view of complex motor skills while leveraging an attention mechanism to highlight the areas that most need correction. Adopting an exclusively "key tips only" approach might overlook critical interdependencies. Nevertheless, the method still captures key actions within the broader motion, offering a balance between whole-motion imitation and targeted correction.

5.8.3 Limitation

Despite the promise shown across these three studies, several important limitations remain:

1. Short Training Intervals. All studies provided relatively brief sessions (often under 10 minutes per condition), which may not capture the true potential of the AI coach over multiple days or weeks of practice. Complex motor skills like the golf swing typically require extended, repeated feedback sessions to solidify noticeable improvements.

2. Sample Sizes and Participant Diversity. Our user groups were relatively small (especially in Study 2), and many were novice golfers. Results might differ for more experienced players, who could better exploit advanced features like synchronization. Future experiments with diverse participant skill levels would clarify how novices vs. intermediates vs. advanced users respond to the AI's feedback.

3. Priority vs. Largest Discrepancy. Studies consistently showed that the *largest* geometric difference (the AI's perspective) may not always align with the coach's *priority* perspective. We have not yet integrated **domain-driven weighting** to shift the AI's

focus from purely geometric differences to *practically important* ones (e.g., address or backswing fundamentals).

4. Limited Representation of Equipment & Gaze. Some participants noted a lack of club and gaze-direction visualization. While we used a simplified skeleton for clarity, real-world golf coaching often requires understanding hand/grip, clubface orientation, and eye focus. Future systems should incorporate these elements to give more holistic feedback, especially at *impact*.

5.8.4 Future Applications

The proposed method can be used for self-training systems that detect discrepant motion frames using the distance between golf swings in the latent space and compare 3D human poses at the detected frames. As an application prototype, we combine the motion synchronizer, motion discrepancy detector, and motion manipulator into a single graphical user interface where users can select any professional's form from the database and compare the difference between their forms and the professional's. In addition, instead of directly imitating the selected motion, users can gradually imitate intermediate human poses using a motion manipulator.

As a future work, we also want to dig into the attention map learned by the SA-TCC network. As shown in Figure 5.36, in this particular frame, the attention network focuses more on the right wrist but less on the feet. Similar to this, we can produce a visualization of the attention map to show the importance of each body's parts (Figure 5.37). This may be a clue for discovering which part of the body should be focused on and what can be ignored. To go a step further, this might indicate the importance of body parts that the trainees should follow to revise their pose in an optimized way. In the future, we plan to conduct user studies to evaluate the training effectiveness of the proposed system and determine whether the crucial points shown by the attention maps can be used in real training scenarios. Moreover, we consider that using a high-quality motion capture system can enhance the solidity of the proposed system with high-precision human poses despite the high speed of motion, such as golf swings.

As this work focuses on proposing a novel flow for constructing a sports analysis tool, and because user studies may vary from domain to domain and are flexible for the system designer, further user studies evaluating the proposed system's efficiency have not been addressed in practice. As previously mentioned, there are many ways to design a proper way to provide feedback to users for training. Related works that use multi-modal feedback to alert users when performing the wrong way compared to professionals have shown their significance during training. In the future, we plan to combine our approach with other feedback systems and evaluate the effectiveness of the system. Furthermore, we plan to apply the proposed approach to other sports and skill training processes to explore the generality of the proposed method.

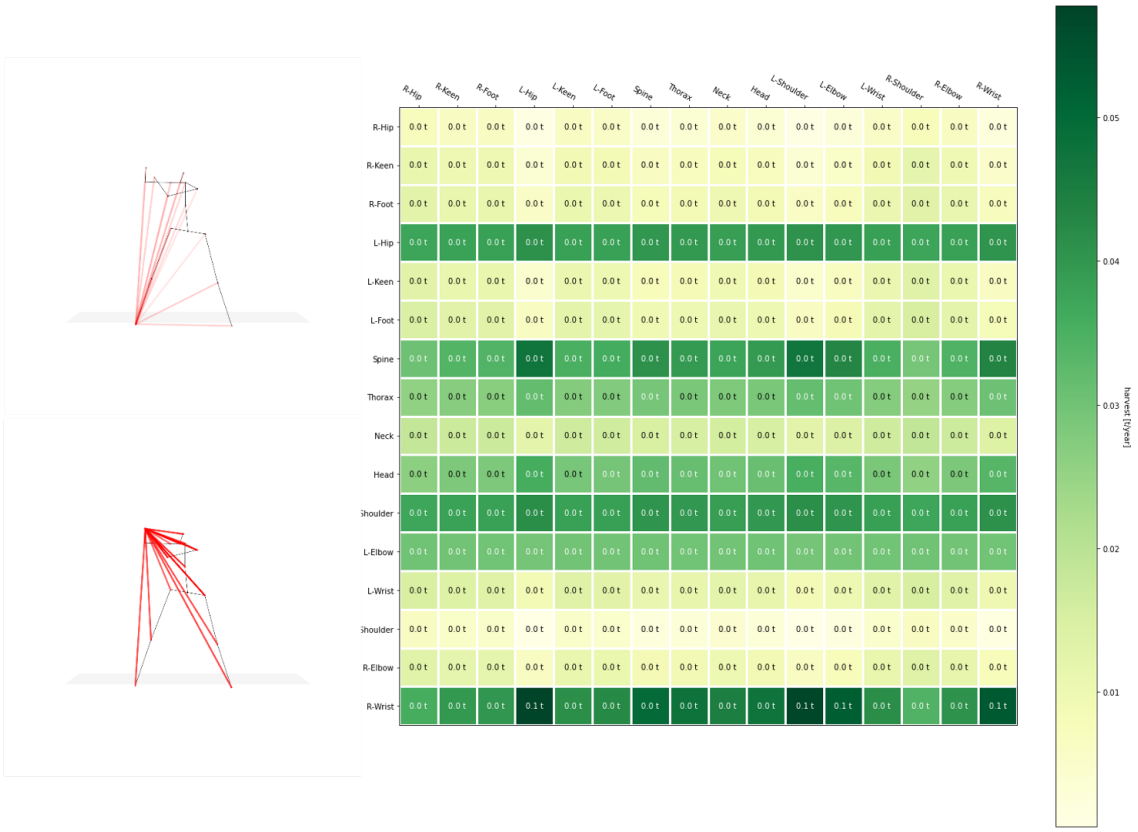

Figure 5.36: Visualization of the attention map.

5.8.5 Summary

Across these three studies, we have shown that:

- The AI's latent-space approach *largely* aligns with expert judgment on where the swing deviates (*Study 2 & 3*).
- Targeted *error poses* improve user *awareness* of how to correct their form and yield higher motivation (*Study 1 & 2*).
- There remain mismatches in *which phase* to prioritize and modest short-term improvements in actual training outcomes (*Study 3*).

In conclusion, the *AI coach* concept holds promise as an effective training tool, but further research on domain-driven weighting, longer-term studies, and more detailed instructions is necessary to realize its full potential and better replicate the nuanced expertise of professional coaches.

5.9 Conclusion

We proposed a golf swing analysis tool, powered by neural networks, to help users intuitively recognize the differences between their swings and those of professional players.

Figure 5.37: Visualization of body parts for revision suggestions. The attention-based network focuses on different body parts at three different phases (address, top, follow-through from left to right). The density and radius of red spheres show the intensity of the attention of the network.

The work is divided into three core components: **synchronization**, **discrepancy detection**, and **manipulation**. Building on these foundations, we developed and evaluated an *AI Coach* system in three successive studies, refining its features and assessing its efficacy in real user scenarios.

Motion Synchronizer First, we implemented a motion synchronizer that aligns motions with different phases and timings. Our experimental findings indicate that the proposed skeleton-based alignment method can outperform state-of-the-art video-based approaches. By accurately matching swings frame-by-frame, this module lays the groundwork for consistent motion comparison and subsequent analyses.

Motion Discrepancy Detector Second, using newly designed neural networks (specifically the SA-TCC architecture), we introduced a motion discrepancy detector to discover fine-grained differences between two golf swings in the latent space. Comparative analyses (e.g., examining 3D human poses at detected frames) showed that the discrepancy detector can reliably distinguish small vs. large differences. Moreover, our SA-TCC network achieved higher phase classification accuracy compared to the original TCC model, while demonstrating a strong correlation between latent-space distances and MPJPE. These findings validate the core capacity of our system to quantify swing variation in a way that aligns with actual kinematic errors.

Motion Manipulator Third, building on the synchronization and discrepancy detection, we proposed a decoder-based *motion manipulator*. This network can reconstruct human poses from the latent space and generate *intermediate* transitions that do not explicitly appear in the original dataset. In effect, it can propose "bridge" poses that gradually guide users from a beginner's form to a more advanced form, rather than forcing them to imitate a professional's pose in a single step.

Prototype Application and User Studies

Following these core technical contributions, we integrated the modules into a prototype application for golf swing self-training. The system allows users to visualize, compare, and interact with their own swings vs. various experts' swings. By highlighting step-by-step transitions, beginners can learn through incremental corrections rather than attempting a perfect motion outright. We conducted three studies (Study 1, Study 2, and Study 3) to examine the system's performance and usability.

- **Study 1** showed that our AI Coach (providing error poses and a latent-space score) outperformed both conventional video-based training and basic 3D skeleton visualization. Users reported higher motivation and clearer guidance for correcting their poses.
- **Study 2** delved deeper into *motion synchronization* and *error pose visualization*, revealing that users rarely leveraged detailed synchronization but strongly benefited from frame-specific *error poses* to reduce cognitive load and pinpoint exactly *where* and *when* to fix their swings.
- **Study 3** focused on comparing the AI's discrepancy detection to a professional coach's subjective scoring and evaluating whether short-term training improvements emerged when users practiced AI-recommended phases. Despite a solid correlation between AI-measured discrepancies and the coach's ratings, we observed notable differences in which phase each considered *most critical* to correct. Additionally, the short practice sessions in Study 3 yielded only modest improvements, underscoring the need for more extended training intervals to achieve significant skill gains.

Through these studies, we verified several key findings:

- Beginners benefit from direct feedback or guidance, especially in early-stage training.
- Our AI Coach system can yield superior training effects compared to traditional methods, offering incremental corrections that keep users motivated and engaged.
- Domain expertise remains vital: while our latent-space approach can reliably identify *large discrepancies*, professional coaches may prioritize earlier phases (*address* or *backswing*) for fundamental corrections. This gap prompts future integration of domain-weighted heuristics.
- In short-term user evaluations, the AI Coach showed promise in pinpointing swing differences and providing targeted *error poses*, but extended practice intervals are likely necessary to manifest robust, measurable improvements.

Overall, we have demonstrated that the proposed motion synchronizer successfully aligns swings on a phase-by-phase basis, offering improved performance over state-of-the-art video-based methods, particularly when employing 3D skeleton data. The motion

discrepancy detector (SA-TCC) efficiently captures detailed pose differences and exhibits a strong correlation with kinematic errors, surpassing previous TCC architectures in accuracy and reliability. The motion manipulator provides an intuitive approach to gradually transform a beginner's pose into a more advanced form by introducing an intermediate "bridge" step, thereby reducing the complexity inherent in directly imitating expert motions. Furthermore, the prototype AI Coach application that combines these modules not only highlights significant errors but also offers actionable frames for self-correction, a feature that our user studies confirmed to be both more helpful and more motivating than traditional video-based training methods.

In summary, this system establishes a solid foundation for discrepancy-based motor skill training by leveraging neural network latent spaces. Future research will concentrate on incorporating deeper domain knowledge, such as weighting early swing fundamentals, and on conducting extended intervention studies to verify long-term training outcomes. Through these enhancements, the AI Coach aims to evolve into a robust, accessible tool for guided self-training in golf and potentially other sports that demand complex motor skills.

Chapter 6

Discussion

In this thesis, we investigated methods to democratize skill acquisition by developing accessible and personalized training systems that leverage advancements in machine learning, computer vision, and interactive technologies. We proposed a three-step training pipeline aimed at enhancing motor skill learning for individuals without access to professional coaching. While we briefly introduced this pipeline and the concept of democratizing skill acquisition earlier, this chapter focuses on analyzing the effectiveness of our two developed systems—**Coach Navi** (one user study) and **AI Coach** (three user studies)—and the insights gained from their combined results.

6.1 Comparative Analysis

6.1.1 Distinct but Complementary Approaches

Both **Coach Navi** and **AI Coach** share the overarching goal of facilitating motor skill learning, yet they tackle different aspects of our proposed pipeline:

- **Coach Navi** primarily addresses:
    1. *Personalized Learning Target Selection*: By selecting an intermediate-level motion suitable for a user's current skill, rather than requiring immediate imitation of expert moves.
    2. *Representation with User's Appearance*: Through a personalized "ideal me" avatar, which helps users better visualize and adopt the target motions.
- **AI Coach** focuses on:
    1. *Fine-Grained Motion Comparison and Guidance*: Pinpointing discrepancies within the swing via neural-network-based analysis, highlighting specific timings and body parts needing correction, and offering targeted *error poses*.

While Coach Navi is particularly strong at presenting manageable, motivating learning targets, AI Coach excels in providing precise, data-driven feedback for incremental refine-

ment. Taken together, they illustrate two complementary approaches to democratizing skill acquisition.

6.1.2 User Studies Across Two Systems

- **Coach Navi Study**: Explored three training methods for beginner golfers—Video Playback (*Cond. V*), Skeleton & Avatar Visualization (*Cond. SA*), and **Coach Navi** (*Cond. CN*). Results indicated that Coach Navi outperformed traditional video approaches by providing personalized avatars and intermediate-level targets, thus increasing motivation and self-representation.
- **AI Coach Studies**:
  - *Study 1:* Compared AI Coach's error-pose-based guidance to conventional video and skeleton-based training. Found that providing explicit *error poses* and scoring increased user engagement and improved learning outcomes.
  - *Study 2:* Investigated advanced features such as motion synchronization vs. a simpler, single-frame error-pose highlight. Users favored *targeted* over *complex* feedback.
  - *Study 3:* Assessed how AI Coach's discrepancy detection aligns with professional coaches' judgments, and whether focusing on AI-recommended phases yields measurable improvements in short-term training. The AI's latent-space analysis correlated moderately to strongly with expert views, yet differences emerged in *which* swing phase each saw as top priority. Short practice sessions showed only modest immediate improvements, underscoring the need for longer training intervals.

6.2 Combined Results

The combined studies across Coach Navi and AI Coach lead to several overarching themes and insights.

6.2.1 Personalization Enhances Learning and Engagement

Intermediate-Level Targets. Coach Navi showed that when users attempt more reachable, intermediate motions (instead of advanced expert forms), they are less intimidated and more motivated. Similarly, AI Coach can be adapted to propose incremental improvements rather than big jumps to expert-level swings, aligning with the principle of scaffolding.

Customized Avatars and Feedback. Coach Navi's avatar-based approach provided higher self-representation, improving body awareness and motivation. AI Coach's *error*

poses similarly tailored each correction to the user's specific mistakes. In both systems, personalization significantly boosted engagement and user satisfaction.

6.2.2 3D Visualization Outperforms 2D Video

Participants across all studies consistently reported that 3D presentations (e.g., avatars, skeleton-based replays, or error-pose frames) offered superior clarity over 2D videos. The ability to manipulate viewing angles and see full-body details made it easier to grasp nuanced movements. Traditional video training often proved cumbersome for novices unaccustomed to mentally mapping 2D visuals to 3D physical motions.

6.2.3 Actionable and Targeted Feedback Reduces Frustration

AI Coach's single-frame or single-phase error cues particularly highlight the value of *targeted* feedback. While advanced features like synchronized multi-frame playback appeared theoretically useful, many participants felt cognitively overwhelmed. By contrast, an explicit cue (e.g., "Here is the critical timing to fix") was easier to digest and apply. Coach Navi similarly limited the complexity of the target motion by providing an intermediate skill level, which users found both less frustrating and more motivational.

6.2.4 Coach–AI Alignment and Priority Mismatches

From AI Coach's third study, we learned that while the AI's latent-space distances often correlate well with a coach's discrepancy ratings, each may differ in *which* phases matter most for improvement. The AI typically flags whichever phase shows the largest geometric difference, while professional coaches or domain experts sometimes prioritize earlier, foundational phases (like *address* or *backswing*). This gap underscores the importance of integrating domain knowledge into purely data-driven methods.

6.3 Internal Models: Motor and Sensory Models

In the context of motor skill acquisition, **internal models** are fundamental cognitive frameworks that enable individuals to plan, execute, and refine their movements. These models consist of two primary components: the **motor model** and the **sensory model**. Understanding how these internal models function and interact is crucial when discussing effective training systems. Our training systems, **Coach Navi** and **AI Coach**, significantly contribute to both the motor and sensory models, thereby enhancing the overall training process and facilitating more efficient skill acquisition.

6.3.1 Motor Model Enhancement

The **motor model** represents the brain's internal representation of motor commands sent to muscles to execute movements. It involves planning the sequence of muscle activations,

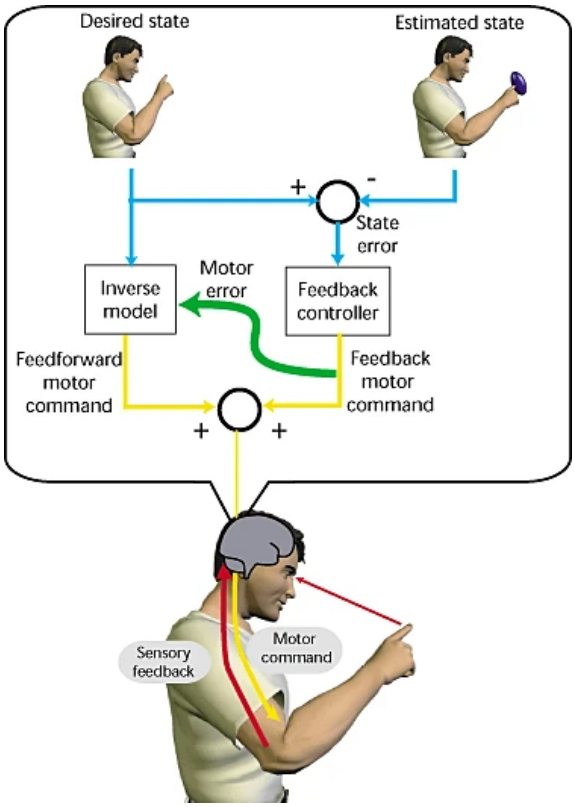

Figure 6.1: A schematic of feedback-error learning. [153]

predicting the outcomes of motor actions, and adapting these commands based on feedback 6.1.

Coach Navi can enhance the motor model through the following mechanisms:

- **Visualization of Ideal Movements:** Coach Navi provides users with personalized 3D avatars performing idealized movements. This feature serves as a mental blueprint, allowing users to engage in *motor imagery*—mentally rehearsing movements to enhance motor planning and execution [37]. By visualizing an idealized version of themselves, users can anticipate and coordinate muscle activations more accurately. Empirical data from our user studies indicate that participants utilizing the 3D avatar comparison reported improved movement accuracy and smoother execution, suggesting that mental rehearsal facilitated by personalized visualization contributes to a more refined motor model.

- **Personalized Learning Targets:** Coach Navi employs a motion navigator to select intermediate-level learning targets tailored to the user's current skill level. This approach ensures that users progress through skill development in manageable steps. Providing attainable challenges prevents users from being overwhelmed by the complexity of expert-level movements, fostering a gradual and structured enhancement of the motor model. Quantitative results demonstrated that users trained with Coach Navi showed significant improvements in *Mean Per Joint Angle Error (MPJAE)*

compared to those using traditional video methods. This indicates that personalized, intermediate targets facilitate more accurate and controlled motor command generation, reinforcing correct movement patterns and enhancing motor learning efficiency.

6.3.2 Sensory Model Enhancement

The **sensory model** processes and interprets sensory feedback from movements, enabling individuals to detect and correct errors. It integrates information from various sensory modalities such as vision, proprioception, and tactile feedback [128].

Coach Navi and **AI Coach** can enhance the sensory model through the following features:

- **Error Pose Visualization (AI Coach):** AI Coach highlights specific discrepancies between the user's motion and that of professional players by displaying error poses. This targeted feedback allows users to accurately identify and focus on precise aspects of their movements that require improvement. Users found the error pose visualization intuitive and effective in pinpointing exact areas needing correction, leading to more precise and efficient movement adjustments. Quantitative data from *AI Coach's Study 1* showed significant reductions in *Mean Per Joint Position Error (MPJPE)*, indicating effective sensory feedback processing.

- **3D Avatar Comparison (Coach Navi):** Coach Navi utilizes personalized 3D avatars to perform both the user's current and ideal movements. The multi-angle visualization enhances the user's ability to perceive and interpret their own motions in relation to the target, improving *body awareness* and *sensory feedback*. Participants reported higher levels of body awareness and self-representation, contributing to more accurate sensory feedback and more effective error correction. This is supported by improved scores in *Coach Navi's user study*.

- **Actionable Feedback and Scoring Systems:** AI Coach incorporates a pseudo scoring system that quantifies performance by calculating the average latent space distance between the user's motion and the expert's motion. This quantitative feedback provides users with clear metrics to assess their progress and understand the effectiveness of their adjustments. Users found the scoring system helpful in tracking progress objectively, reinforcing positive feedback loops and enhancing their motivation to continue training. In *AI Coach's Study 1*, the scoring system correlated strongly with user satisfaction and perceived improvement.

6.3.3 Integrated Contribution to Internal Models

By enhancing both the **motor model** and the **sensory model**, **Coach Navi** and **AI Coach** offer a comprehensive training experience that supports the entire motor skill

acquisition process:

- **Motor Planning and Execution:** Visualization tools such as personalized 3D avatars and error pose displays enable users to mentally plan and execute movements with greater accuracy. By providing clear mental blueprints and precise feedback, these tools ensure that motor commands are well-coordinated and refined.
- **Error Detection and Correction:** The systems' ability to highlight specific errors and provide quantitative scores enhances the sensory model's capacity for accurate error detection. This, in turn, informs the motor model, allowing users to adjust their movements based on reliable sensory feedback.
- **Feedback Loop Synergy:** The continuous interaction between enhanced sensory feedback and refined motor commands creates a robust feedback loop. Users receive actionable insights that inform their motor planning, leading to iterative improvements in their movements.

Empirical evidence from our user studies supports this integrated contribution. Users reported higher levels of motivation and engagement, attributed to the personalized and actionable nature of the feedback provided by both systems. Quantitative metrics, such as reductions in *MPJAE*, further validate the effectiveness of our systems in enhancing both internal models.

6.3.4 Future Enhancements to Motor Models

Looking forward, we plan to incorporate bioinformatic data, such as muscle activation patterns, into our systems. This integration aims to provide users with deeper insights into muscle control and coordination, further refining the motor model. By leveraging detailed physiological data, our systems can offer more personalized and precise guidance on muscle usage, enhancing users' ability to execute and control their movements effectively.

- **Muscle Activation Data Integration:** Incorporating electromyography (EMG) data to monitor muscle activation patterns during movements. Real-time feedback on muscle engagement allows users to adjust their muscle activations for more efficient and controlled movements. Understanding which muscles are under or over-utilized helps users fine-tune their motor commands to optimize performance and reduce the risk of injury [45].
- **Advanced Biomechanical Modeling:** Developing biomechanical models that account for individual differences in anatomy and physiology. Personalized biomechanical feedback can lead to more accurate and effective motor command adjustments, further enhancing the motor model's accuracy and adaptability. This approach accommodates users of different ages, body types, and fitness levels, making the system more inclusive and effective [128].

These future enhancements will deepen the interaction between motor commands and sensory feedback, fostering even more precise and controlled skill acquisition. By providing detailed insights into muscle control, our systems will enable users to develop more refined and efficient movement patterns, thereby enhancing the overall effectiveness of motor skill training.

6.4 Integration of *Coach Navi* and *AI Coach*

In this section, we summarize how the two systems developed in this research—*Coach Navi* and *AI Coach*—fit together, and what distinct functionalities each one provides. While each system can be used independently, combining them yields a more comprehensive training environment.

6.4.1 Coach Navi

- **Intermediate-Level Motion Target:** Coach Navi focuses on selecting an appropriate learning target for the user based on their skill level. Instead of forcing beginners to replicate an expert's advanced motion immediately, it finds a more attainable (i.e., intermediate) version of the motion, guiding users gradually toward higher proficiency.
- **Motion Style Transfer (MST):** The system also performs a style transformation to adapt the selected target to the user's own body shape, generating a personalized "ideal me" avatar. This allows users to visualize themselves performing an achievable, yet improved, form of the motion.
- **Latent-Space Navigation:** By leveraging a skill-labeled latent space, Coach Navi organizes motions according to their skill progression. Hence, users can step through progressively more advanced motions while preserving each motion's essential biomechanics.

6.4.2 AI Coach

- **Precise Error Detection:** AI Coach concentrates on detecting discrepancies in a fine-grained manner, pinpointing exactly where and when the user's motion deviates from the target. An attention-based network (SA-TCC) provides high accuracy in identifying these deviations.
- **Motion Synchronization and Error Pose:** The system can synchronize the user's swing with a prerecorded reference (e.g., an expert's motion), highlighting "error poses" that single out frames most in need of correction. This approach delivers actionable feedback, showing both the recommended timing and specific joints that require adjustment.

- **Frame-Specific or Segment-Specific Guidance:** Users can see precisely which body parts differ most from the ideal form, allowing them to focus on relevant corrections rather than searching the entire motion sequence for possible mistakes.

6.4.3 Combining *Coach Navi* and *AI Coach*

While *Coach Navi* and *AI Coach* each address different aspects of motor skill learning, their functions naturally complement one another:

1. **Target Selection via Coach Navi:** Users begin by obtaining an intermediate or user-friendly "ideal me" motion. Coach Navi ensures this target is neither too advanced nor too distant from the user's current capabilities, thus preventing intimidation or overload.

2. **Detailed Correction via AI Coach:** Once the target is set, AI Coach detects subtle discrepancies in the user's execution, synchronizing their motion with the reference and highlighting the specific frames that deviate. This helps users refine the intermediate motion more effectively.

3. **Adaptive, End-to-End Solution:** Taken together, the two systems yield a pipeline where users can (1) adopt a progressively advanced motion, then (2) receive targeted feedback on how to correct errors at a frame-level. This synergy supports continuous improvement without overwhelming beginners with an expert-level motion at the outset.

6.5 Future Work

6.5.1 Incorporating Inertial and Muscle Activation Data

While our current system primarily utilizes pose information (e.g., 3D joint positions) to analyze a user's golf swing, inertial factors and muscle activities can further enrich the training process. For instance, electromyography (EMG) data could capture muscle activation patterns, highlighting whether a player is exerting force appropriately or risking injury by overcompensating. Such data would be particularly valuable for beginners and users recovering from physical limitations, since subtle muscle imbalances could undermine an otherwise correct pose.

On a methodological level, incorporating muscle activation or force data poses additional challenges in terms of data representation and model complexity. Nonetheless, these challenges align well with the strengths of *machine learning* approaches: neural networks excel at assimilating multiple heterogeneous inputs. By extending our current framework to accept EMG or force data in tandem with joint-position sequences, we could offer more nuanced error detection and provide feedback not only on pose discrepancies but also on force management and muscular coordination. This multimodal expansion would enhance personalization, potentially giving targeted advice on whether users should redistribute force or adjust muscle tension at specific phases of the swing.

6.5.2 Outcome-Based Evaluations

Although this work focuses on detecting and correcting swing form, many learners measure success by the outcome of their shots, such as the ball's flight trajectory or a final score in a round of golf. Currently, our evaluation methods center on pose accuracy and motion discrepancy (e.g., MPJPE), but outcome-based metrics could offer a more comprehensive view of skill improvement.

In practice, however, beginners often struggle to make consistent contact with the ball, making a strictly outcome-focused evaluation more challenging in early training stages. One viable solution might be to employ VR simulations or specialized training environments that simplify ball contact, allowing novices to practice correct form while still tracking ball trajectory or distance in a controlled setup. Incorporating haptic feedback to simulate club impact and audio cues for ball contact may further enhance realism and user engagement. By bridging pose correction with direct outcome metrics, we could confirm whether an improved swing form genuinely translates to tangible performance gains such as increased consistency, better shot accuracy, or higher score improvements.

6.5.3 Comparing Hand-Crafted vs. Neural Network–Based Intermediate Motions

While this work generates intermediate motions using a latent-space approach, one could devise a simpler, hand-crafted method (e.g., linearly blending joint angles between beginner and expert forms). Such manual interpolations are straightforward and provide explicit control over how far to push each intermediate step. However, they can lack adaptability, especially when applied to new users with different body types or swing characteristics, and often produce less natural transitions.

By contrast, our neural network–based strategy automatically learns to generate smooth, personalized "mid-tier" motions. Once trained, the system can adapt to diverse users or additional skill levels with minimal extra effort. Future investigations could directly compare these two methods—hand-crafted interpolation vs. machine learning—to assess differences in motion realism, ease of user adoption, and final performance outcomes. Such a study would illuminate where each approach excels and how best to integrate or choose between them.

6.5.4 Real-Time or AR/VR-Based Feedback

Both Coach Navi and AI Coach could be enhanced by interactive, real-time solutions in AR or VR. Real-time overlays of avatar or error skeleton cues might allow users to correct errors on the spot, improving the training experience further. At the same time, care must be taken to avoid overloading users with too many simultaneous cues.

6.5.5 Domain-Weighted Priorities and Biomechanical Models

To address the mismatch in **which phase** to fix first, future AI algorithms could embed causal or biomechanical reasoning, giving extra priority to early-swing fundamentals if they propagate errors downstream. Such domain-weighted approaches may better reflect professional coaches' strategies.

6.5.6 Integration of Equipment Visualizations

Several participants noted the absence of the golf club or other sports equipment in the 3D representations. Including realistic equipment models and tracking data could help learners understand hand positions, grip, and clubface orientation, thereby offering more comprehensive training.

6.5.7 Longitudinal and Multi-Activity Evaluations

Extensive, longer-term studies are necessary to confirm whether these systems can drive lasting improvements. Research on skill retention, transfer to on-course performance, and motivation over multiple weeks or months would provide robust evidence of effectiveness.

Further, applying these techniques to different sports (e.g., tennis, baseball, dance) could demonstrate the broader applicability of both Coach Navi's *intermediate targeting* and AI Coach's *latent-space discrepancy* methods.

6.5.8 Potential Merger of Coach Navi and AI Coach Components

Finally, merging the approaches of Coach Navi (personalized, incremental target motions) and AI Coach (precise error detection and cues) into a unified system might offer learners a full pipeline: from selecting the best intermediate target, to identifying frame-specific corrections, to refining advanced swings over time. Such a cohesive platform could meet novices *and* intermediates where they are in the skill curve, promoting continuous improvement with tailored feedback.

Chapter 7

Conclusions

Beginners in sports often face significant challenges when striving to improve their skills, primarily due to limited prior knowledge and insufficient access to professional coaching. These difficulties are particularly evident in golf, where mastering an effective swing depends on executing intricate movements and precise timing. Traditional training methods, such as watching professional players in videos, typically lack the depth and personalization needed for novices to target their weaknesses effectively. Without expert feedback, learners struggle to identify the details of their performance that need adjustment, as well as how to make such adjustments in a way that accommodates their individual physical attributes.

In this thesis, we addressed these challenges by grounding our work in three specific research questions that guided our design of a more accessible and self-sufficient training framework. First, we asked *"Who to follow?"*, recognizing that expecting learners to imitate fixed professional template motions disregards individual biomechanical differences and fails to account for a user's current skill level. Second, we posed the question *"What to learn?"*, highlighting the disconnect that arises when novices try to emulate a motion performed by someone who has a different physique, muscle composition, or style. Finally, we considered *"How to improve?"* by emphasizing the need for clear, actionable feedback at the right moments within a motion, so that learners can focus on the most impactful elements of their technique.

To respond to these questions, we proposed a three-step training pipeline, which formed the conceptual backbone of our investigation and shaped two complementary systems: **Coach Navi** and **AI Coach**. *Coach Navi* implements the first two steps by personalizing learning target selection and rendering 3D avatar representations that match each user's appearance, thus catering to "who to follow" and "what to learn." Rather than intimidating beginners with highly advanced professional motions, Coach Navi finds and displays more intermediate forms that better suit the learner's current ability. Through a motion style transformation network, it also provides a tangible "ideal me" avatar that allows individuals to visualize how they should appear when performing the skill at an

attainable next level.

AI Coach implements the third step, focusing on "how to improve." By compressing both the user's and professional motions into a latent space and identifying the frames or segments of greatest discrepancy, AI Coach highlights exactly where an individual's swing deviates most. The system's error poses and scoring mechanism enable learners to see a frame-by-frame breakdown of their performance relative to a higher-skilled reference. User studies repeatedly showed that participants responded positively to this detailed, data-driven guidance, finding it more actionable and less frustrating than generic video playback.

We validated these systems through four user studies—one for Coach Navi and three for AI Coach—each of which contributed unique insights to the overarching question of *democratizing skill acquisition* in golf. Coach Navi's study established that tailoring learners' targets to a more moderate level of complexity (and aligning them with each user's body type) reduced cognitive overload and sustained motivation. AI Coach's studies collectively demonstrated that data-driven feedback, when presented in a focused manner, fosters stronger engagement, even though short practice intervals often limited the magnitude of observable improvements. The final study showed that while AI-driven discrepancy measurements can correlate well with professional coaching judgment, purely geometric approaches can sometimes overlook the foundational phases that experts consider essential for building an effective swing.

Taken together, these findings confirm that addressing *who to follow*, *what to learn*, and *how to improve* is vital for producing truly accessible, effective training systems. Coach Navi helps learners identify a manageable motion target and see themselves performing it, reducing the mismatch and intimidation tied to advanced demonstrations. AI Coach then pinpoints the crucial frames requiring attention, offering explicit cues and a scoring system that tracks progress. These systems engage both the motor and sensory models of skill acquisition, providing structure, clarity, and motivation for novice learners. Moreover, user feedback underscores the potential for domain-informed enhancements, such as weighting early-swing fundamentals more strongly or integrating AR/VR for real-time correction. Longer-term interventions and larger-scale evaluations would further reveal the extent to which these approaches translate into durable, real-world improvements.

In conclusion, by systematically tackling the questions of *who to follow*, *what to learn*, and *how to improve*, this thesis contributes a novel framework that unites intermediate-level personalized learning with fine-grained error detection. The synergy of Coach Navi and AI Coach provides a coherent pipeline that begins with approachable, user-specific targets and evolves into detailed, data-driven feedback mechanisms. This strategy enhances the feasibility of self-training and reduces reliance on constant professional oversight, paving the way for broader democratization of skill acquisition in golf and, potentially, other domains requiring nuanced motor skills.

References

[1] Blender motion capture retargeting tip. https://mx.pinterest.com/pin/blender-motion-capture-retargeting-tip--390124386468354516/. Accessed: 2024-11-05.

[2] Cygames motion capture studio. https://magazine.cygames.co.jp/archives/8209. Accessed: 2024-11-05.

[3] Marker placement protocols. http://www.lifemodeler.com/LM_Manual_2010/A_motion.shtml. Accessed: 2021-01-01.

[4] Smart golf training analysis system. https://www.itri.org.tw/english/ListStyle.aspx?DisplayStyle=01_content&SiteID=1&MmmID=1037333526356626457&MGID=1127157514330125717. Accessed: 2021-01-01.

[5] Vr ski coach. https://www.vogue.cs.titech.ac.jp/projects/digitalsports/vr-ski-coach. Accessed: 2021-01-01.

[6] *Visual Analysis of Humans: Looking at People*. Springer London, London, 2011.

[7] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 481–490, 2019.

[8] Z.A. Abbas and J.S. North. Good-vs. poor-trial feedback in motor learning: The role of self-efficacy and intrinsic motivation across levels of task difficulty. *Learning and Instruction*, 55:105–112, 2018.

[9] Kfir Aberman, Peizhuo Li, Sorkine-Hornung Olga, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62, 2020.

[10] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Trans. Graph.*, 39(4), August 2020.

[11] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[12] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed Human Avatars from Monocular Video . In *2018 International Conference on 3D Vision (3DV)*, pages 98–109, Los Alamitos, CA, USA, September 2018. IEEE Computer Society.

[13] S. Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. Multi-view pictorial structures for 3d human pose estimation. In *British Machine Vision Conference*, 2013.

[14] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[15] Kota Arai, Yutaro Hirao, Takuji Narumi, Tomohiko Nakamura, Shinnosuke Takamichi, and Shigeo Yoshida. Timtoshape: Supporting practice of musical instruments by visualizing timbre with 2d shapes based on crossmodal correspondences. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 850–865, New York, NY, USA, 2023. Association for Computing Machinery.

[16] Greg Atkinson and Alan M. Nevill. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, 26(4):217–238, 1998.

[17] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *2011 International Conference on Computer Vision*, pages 1092–1099, 2011.

[18] Arnold Baca and Peter Kornfeind. Rapid feedback systems for elite sports training. *IEEE Pervasive Computing*, 5(4):70–76, 2006.

[19] Gi-Seung Bang and Seung-Bo Park. Workout classification using a convolutional neural network in ensemble learning. *Sensors (Basel)*, 24(10):3133, May 2024.

[20] Renato Baptista, Girum Demisse, Djamila Aouada, and Björn Ottersten. Deformation-based abnormal motion detection using 3d skeletons. In *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2018.

[21] Roger Bartlett. *Introduction to Sports Biomechanics: Analysing Human Movement Patterns*. Routledge, London, 2nd edition, 2007.

[22] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, page 359–370. AAAI Press, 1994.

[23] Ernesto Brau and Hao Jiang. 3d human pose estimation via deep learning from 2d annotations. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 582–591, 2016.

[24] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, pages 8–15, 1998.

[25] John Brooke. Sus: A quick and dirty usability scale, 1996.

[26] Adrian Bulat, Georgios Tzimiropoulos, Jean Kossaifi, and Maja Pantic. Improved training of binary networks for human pose estimation and image recognition. *ArXiv*, abs/1904.05868, 2019.

[27] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021.

[28] Gerry Carr. *Mechanics of Sport*. Human Kinetics, Champaign, IL, 2nd edition, 1997.

[29] Dan Casas and Marc Comino Trinidad. Smplitex: A generative model and dataset for 3d human texture estimation from single image. In *British Machine Vision Conference*, 2023.

[30] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3643–3651, 2017.

[31] Steven Chen and Richard R. Yang. Pose trainer: Correcting exercise posture using pose estimation, 2020.

[32] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1221–1230, 2017.

[33] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8185–8194, 2020.

[34] Chia-Jung Chou, Jui-Ting Chien, and Hwann-Tzong Chen. Self adversarial training for human pose estimation. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–30, 2018.

[35] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5669–5678, 2017.

[36] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *ACM Trans. Graph.*, 27(3):1–10, August 2008.

[37] Jean Decety and Diana H. Ingvar. Brain structures participating in mental simulation of motor behavior: A neuropsychological interpretation. *Acta Psychologica*, 73(1):13–34, 1990.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[39] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[40] Bhat Dittakavi, Divyagna Bavikadi, Sai Vikas Desai, Soumi Chakraborty, Nishant Reddy, Vineeth N Balasubramanian, Bharathi Callepalli, and Ayon Sharma. Pose tutor: An explainable system for pose correction in the wild. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3539–3548, 2022.

[41] Han Du, Erik Herrmann, Janis Sprenger, Klaus Fischer, and Philipp Slusallek. Stylistic locomotion modeling and synthesis using variational generative models. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, MIG '19, New York, NY, USA, 2019. Association for Computing Machinery.

[42] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1801–1810, 2019.

[43] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3810–3818, 2015.

[44] Don Samitha Elvitigala, Denys J.C. Matthies, Löic David, Chamod Weerasinghe, and Suranga Nanayakkara. Gymsoles: Improving squats and dead-lifts by visualizing the user's center of pressure. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[45] Dario Farina, Sara Muceli, and Robert M. Enoka. The extraction of neural strategies from the surface electromyogram. *Nature Reviews Neuroscience*, 15(11):745–756, 2014.

[46] Zhiwen Feng, Zhi Liu, Xuejing Shen, Yue Wu, and Qun Zhao. Mining athletic performance via posture analysis with a markerless motion capture system. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1739–1748, 2014.

[47] P. M. Fitts and M. I. Posner. *Human performance.* Human performance. Brooks/Cole, Oxford, England, 1967.

[48] Simone Francia. *Classificazione di Azioni Cestistiche mediante Tecniche di Deep Learning.* PhD thesis, 04 2018.

[49] Juergen Gall, Bodo Rosenhahn, Thomas Brox, and Hans-Peter Seidel. Optimization and filtering for human motion capture. *International Journal of Computer Vision*, 87(1):75–92, Mar 2010.

[50] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[51] Paul S. Glazier and Keith Davids. Constraints on the complete optimization of human motion. *Sports Medicine*, 39(1):15–28, 2009.

[52] Mark A Guadagnoli and Timothy D Lee. Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *J. Mot. Behav.*, 36(2):212–224, June 2004.

[53] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. A Deeper Look into DeepCap (Invited Paper) . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(04):4009–4022, April 2023.

[54] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. DeepCap: Monocular Human Performance Capture Using Weak Supervision . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5051–5062, Los Alamitos, CA, USA, June 2020. IEEE Computer Society.

[55] Tavi Halperin, Ariel Ephrat, and Shmuel Peleg. Dynamic temporal alignment of speech to lips. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3980–3984, 2019.

[56] Perttu Hämäläinen. Interactive video mirrors for sports training. In *Proceedings of the Third Nordic Conference on Human-Computer Interaction*, NordiCHI '04, page 199–202, New York, NY, USA, 2004. Association for Computing Machinery.

[57] Ping-Hsuan Han, Yang-Sheng Chen, Yilun Zhong, Han-Lei Wang, and Yi-Ping Hung. My tai-chi coaches: An augmented-learning tool for practicing tai-chi chuan. In *Proceedings of the 8th Augmented Human International Conference*, AH '17, New York, NY, USA, 2017. Association for Computing Machinery.

[58] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[59] Shoichi Hasegawa, Seiichiro Ishijima, Fumihiro Kato, Hironori Mitake, and Makoto Sato. Realtime sonification of the center of gravity for skiing. In *Proceedings of the 3rd Augmented Human International Conference*, AH '12, New York, NY, USA, 2012. Association for Computing Machinery.

[60] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[62] Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. A benchmark for the short version of the user experience questionnaire. In *International Conference on Web Information Systems and Technologies*, 2018.

[63] Thuong N. Hoang, Martin Reinoso, Frank Vetere, and Egemen Tanin. Onebody: Remote posture guidance system using first person view in virtual environment. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, New York, NY, USA, 2016. Association for Computing Machinery.

[64] Jana Hoffard, Takuto Nakamura, Erwin Wu, and Hideki Koike. Pushtoski - an indoor ski training system using haptic feedback. In *ACM SIGGRAPH 2021 Posters*, SIGGRAPH '21, New York, NY, USA, 2021. Association for Computing Machinery.

[65] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Trans. Graph.*, 36(4), July 2017.

[66] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4), July 2016.

[67] Michael B. Holte, Cuong Tran, Mohan M. Trivedi, and Thomas B. Moeslund. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):538–552, 2012.

[68] Yuuki Horiuchi, Yasutoshi Makino, and Hiroyuki Shinoda. Computational foresight: Forecasting human body motion in real-time for reducing delays in interactive system. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, ISS '17, page 312–317, New York, NY, USA, 2017. Association for Computing Machinery.

[69] Jun-Da Huang. Kinerehab: a kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '11, page 319–320, New York, NY, USA, 2011. Association for Computing Machinery.

[70] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017.

[71] Atsuki Ikeda, Dong Hyun Hwang, and Hideki Koike. AR based Self-sports Learning System using Decayed Dynamic TimeWarping Algorithm. In Gerd Bruder, Shunsuke Yoshimoto, and Sue Cobb, editors, *ICAT-EGVE 2018 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. The Eurographics Association, 2018.

[72] Atsuki Ikeda, Yuka Tanaka, Dong-Hyun Hwang, Homare Kon, and Hideki Koike. Golf training system using sonification and virtual shadow. In *ACM SIGGRAPH 2019 Emerging Technologies*, SIGGRAPH '19, New York, NY, USA, 2019. Association for Computing Machinery.

[73] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[74] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 2022.

[75] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing.

[76] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the 2015 IEEE International Confer-*

ence on Computer Vision (ICCV), ICCV '15, page 3334–3342, USA, 2015. IEEE Computer Society.

[77] Theodore T. Kim, Mohamed A. Zohdy, and Michael P. Barker. Applying pose estimation to predict amateur golf swing performance using edge processing. *IEEE Access*, 8:143769–143776, 2020.

[78] Naoki Kimura, Keisuke Shiro, Yota Takakura, Hiromi Nakamura, and Jun Rekimoto. Sonospace: Visual feedback of timbre with unsupervised learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 367–374, New York, NY, USA, 2020. Association for Computing Machinery.

[79] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10236–10245, 2018.

[80] Duane Knudson and Craig Morrison. *Qualitative Analysis of Human Movement*. Human Kinetics, Champaign, IL, 2nd edition, 2002.

[81] Kyeong-Ri Ko and Sung Bum Pan. Cnn and bi-lstm based 3d golf swing analysis by frontal swing sequence images. *Multimedia Tools and Applications*, 80(6):8957–8972, Mar 2021.

[82] Jordan Koulouris, Zoe Jeffery, James Best, Eamonn O'Neill, and Christof Lutteroth. Me vs. super(wo)man: Effects of customization and identification in a vr exergame. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–17, New York, NY, USA, 2020. Association for Computing Machinery.

[83] Itaru Kuramoto, Yukari Nishimura, Keiko Yamamoto, Yu Shibuya, and Yoshihiro Tsujino. Visualizing velocity and acceleration on augmented practice mirror self-learning support system of physical motion. In *2013 Second IIAI International Conference on Advanced Applied Informatics*, pages 365–368, 2013.

[84] Matthew Kyan, Guoyu Sun, Haiyan Li, Ling Zhong, Paisarn Muneesawang, Nan Dong, Bruce Elder, and Ling Guan. An approach to ballet dance training through ms kinect and visualization in a cave virtual reality environment. *ACM Trans. Intell. Syst. Technol.*, 6(2), mar 2015.

[85] Adrian Lees. Technique analysis in sports: a critical review. *Journal of Sports Sciences*, 20(10):813–828, 2002.

[86] Rui Li, Ping Zhou, Changyin Xiong, and Yang Zhou. Deep learning for sports videos: Toward large-scale human-centric video analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2302–2315, 2019.

[87] Chen-Chieh Liao, Dong-Hyun Hwang, and Hideki Koike. How can i swing like pro?: Golf swing analysis tool for self training. In *SIGGRAPH Asia 2021 Posters*, SA '21 Posters, New York, NY, USA, 2021. Association for Computing Machinery.

[88] Chen-Chieh Liao, Dong-Hyun Hwang, and Hideki Koike. Ai golf: Golf swing analysis tool for self-training. *IEEE Access*, 10:106286–106295, 2022.

[89] Dario G. Liebermann, Leor Katz, Michael D. Hughes, Roger M. Bartlett, James McClements, and Ian M. Franks. Advances in the application of information technology to sport performance. *Journal of Sports Sciences*, 20(10):755–769, 2002.

[90] Tica Lin, Rishi Singh, Yalong Yang, Carolina Nobre, Johanna Beyer, Maurice A. Smith, and Hanspeter Pfister. Towards an understanding of situated ar visualization for basketball free-throw training. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[91] Ruofan Liu, Erwin Wu, Chen-Chieh Liao, Hayato Nishioka, Shinichi Furuya, and Hideki Koike. Synchronized hand difference visualization for piano learning. In *ACM SIGGRAPH 2022 Posters*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery.

[92] Tao Liu, Yoshio Inoue, and Kyoko Shibata. Development of a wearable sensor system for quantitative gait analysis. *Measurement*, 42(7):978–988, 2009.

[93] Zhi-Song Liu, Vicky Kalogeiton, and Marie-Paule Cani. Multiple style transfer via variational autoencoder. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2413–2417, 2021.

[94] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[95] Diogo C. Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[96] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017.

[97] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *arXiv preprint arXiv:2201.04439*, 2022.

[98] William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. Golfdb: A video database for golf swing sequencing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2553–2562, 2019.

[99] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017.

[100] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.*, 36(4), July 2017.

[101] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 527–544, Cham, 2016. Springer International Publishing.

[102] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[103] Takuto Nakamura, Daichi Saito, Erwin Wu, and Hideki Koike. Actuated club: Modification of golf-club posture with force feedback and motion prediction in vr environment. In *ACM SIGGRAPH 2020 Emerging Technologies*, SIGGRAPH '20, New York, NY, USA, 2020. Association for Computing Machinery.

[104] Fabian Nater, Helmut Grabner, and Luc Van Gool. Exploiting simple hierarchies for unsupervised human behavior analysis. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2014–2021, 2010.

[105] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.

[106] Kenta Ogawa, Shun Sawada, Kouichi Katsurada, and Hidehumi Ohmura. Automatic detection of poor tone quality in classical guitar playing using deep anomaly

detection method. In *Proceedings of the 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2023*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Institute of Electrical and Electronics Engineers Inc., 2023. Publisher Copyright: © 2023 IEEE.; 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2023 ; Conference date: 22-10-2023 Through 25-10-2023.

[107] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proc. ACM Comput. Graph. Interact. Tech.*, 4(3), September 2021.

[108] Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 4(3):1–17, 2021.

[109] Gemma S. Parra-Dominguez, Babak Taati, and Alex Mihailidis. 3d human motion analysis to detect abnormal events on stairs. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 97–103, 2012.

[110] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[111] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018.

[112] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017.

[113] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1253–1262, 2017.

[114] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7745–7754, 2019.

[115] Axel Pfister, Adrian M. West, Shaw Bronner, and J. Adam Noah. Comparative review of pose estimation techniques for motion analysis, mapping, and localization. *IEEE Transactions on Robotics*, 32(5):980–993, 2016.

[116] Pietro Picerno, Andrea Cereatti, and Aurelio Cappozzo. Joint kinematics estimate using wearable inertial and magnetic sensing modules. *Gait & Posture*, 33(4):476–482, 2011.

[117] Nuttachot Promrit and Sajjaporn Waijanya. Model for practice badminton basic skills by using motion posture detection from video posture embedding and one-shot learning technique. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, AICCC '19, page 117–124, New York, NY, USA, 2020. Association for Computing Machinery.

[118] Helge Rhodin, Nadia Robertini, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. A Versatile Scene Model with Differentiable Visibility Applied to Generative Pose Estimation . In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 765–773, Los Alamitos, CA, USA, December 2015. IEEE Computer Society.

[119] Jim Richards, editor. *Biomechanics in Clinic and Research: An Interactive Teaching and Learning Course*. Elsevier Health Sciences, Edinburgh, 2008.

[120] Nadia Robertini, Dan Casas, Helge Rhodin, Hans-Peter Seidel, and Christian Theobalt. Model-Based Outdoor Performance Capture . In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 166–175, Los Alamitos, CA, USA, October 2016. IEEE Computer Society.

[121] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. *COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder*, pages 382–398. 12 2020.

[122] Katsuhito Sasaki, Keisuke Shiro, and Jun Rekimoto. Exemposer: Predicting poses of experts as examples for beginners in climbing using a neural network. In *Proceedings of the Augmented Humans International Conference*, AHs '20, New York, NY, USA, 2020. Association for Computing Machinery.

[123] Katsuhito Sasaki, Keisuke Shiro, and Jun Rekimoto. Exemposer: Predicting poses of experts as examples for beginners in climbing using a neural network. In *Proceedings of the Augmented Humans International Conference*, AHs '20, New York, NY, USA, 2020. Association for Computing Machinery.

[124] Koya Sato, Yuji Sano, Mai Otsuki, Mizuki Oka, and Kazuhiko Kato. Augmented recreational volleyball court: Supporting the beginners' landing position prediction

skill by providing peripheral visual feedback. In *Proceedings of the 10th Augmented Human International Conference 2019*, AH2019, New York, NY, USA, 2019. Association for Computing Machinery.

[125] Luca Scofano, Alessio Sampieri, Giuseppe Re, Matteo Almanza, Alessandro Panconesi, and Fabio Galasso. About latent roles in forecasting players in team sports. *Neural Processing Letters*, 56(2):66, Feb 2024.

[126] Hirokazu Seki and Yoichi Hori. Detection of abnormal action using image sequence for monitoring system of aged people. *IEEJ Transactions on Industry Applications*, 122(2):182–188, 2002.

[127] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141, 2018.

[128] Reza Shadmehr and John W. Krakauer. A computational neuroanatomy for motor control. *Experimental Brain Research*, 185(3):359–381, 2008.

[129] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304, 2011.

[130] Isabelle M. Shuggi, Hyuk Oh, Patricia A. Shewokis, and Rodolphe J. Gentili. Mental workload and motor performance dynamics during practice of reaching movements under various levels of task difficulty. *Neuroscience*, 360:166–179, 2017.

[131] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1–2):4–27, March 2010.

[132] Leonid Sigal, Michael Isard, Horst Haussecker, and Michael J. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, 98(1):15–48, May 2012.

[133] Roland Sigrist, Georg Rauter, Laura Marchal-Crespo, Robert Riener, and Peter Wolf. Sonification and haptic feedback in addition to visual feedback enhances complex motor task learning. *Experimental Brain Research*, 233(3):909–925, Mar 2015.

[134] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *2011 International Conference on Computer Vision*, pages 951–958, 2011.

[135] Tomohiro Sueishi, Chikara Miyaji, Masataka Narumiya, Yuji Yamakawa, and Masatoshi Ishikawa. High-speed projection method of swing plane for golf training. In *Proceedings of the Augmented Humans International Conference*, AHs '20, New York, NY, USA, 2020. Association for Computing Machinery.

[136] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019.

[137] Yao Tang, Lin Zhao, Zhaoliang Yao, Chen Gong, and Jian Yang. Graph-based motion prediction for abnormal action detection. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, MMAsia '20, New York, NY, USA, 2021. Association for Computing Machinery.

[138] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3961–3970, 2017.

[139] Arshed Toma. *Characterization of normal facial features and their association with genes. PhD Thesis, Cardiff University. http://orca.cf.ac.uk/61852/.* PhD thesis, 03 2014.

[140] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.

[141] Dishita G Turakhia, Yini Qi, Lotta-Gili Blumberg, Andrew Wong, and Stefanie Mueller. Can physical tools that adapt their shape based on a learner's performance help in motor skill training? In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[142] Dishita G Turakhia, Andrew Wong, Yini Qi, Lotta-Gili Blumberg, and Yoonji Kim. Designing adaptive tools for motor skill training. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21 Adjunct, page 137–139, New York, NY, USA, 2021. Association for Computing Machinery.

[143] Laia Turmo Vidal, Hui Zhu, and Abraham Riego-Delgado. Bodylights: Open-ended augmented feedback to support training towards a correct exercise execution. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[144] Raquel Urtasun, David J. Fleet, and Pascal Fua. Temporal motion models for monocular and multiview 3d human body tracking. *Computer Vision and Image Understanding*, 104(2–3):157–177, November 2006.

[145] L. S. VYGOTSKY. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978.

[146] Huaijun Wang, Dandan Du, Junhuai li, Wenchao Ji, and Lei Yu. A cyclic consistency motion style transfer method combined with kinematic constraints. *Journal of Sensors*, 2021:1–17, 06 2021.

[147] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 374–382, New York, NY, USA, 2019. Association for Computing Machinery.

[148] Jingya Wang, Yiqiang Chen, Shangsong Hao, Xin Peng, and Liang Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.

[149] Xiaolin Wei, Peizhao Zhang, and Jinxiang Chai. Accurate realtime full-body motion capture using a single depth camera. *ACM Trans. Graph.*, 31(6), November 2012.

[150] Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, and Yong-Jin Liu. Autoregressive stylized motion synthesis with generative flow. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13607–13607, 2021.

[151] Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, and Yong-Jin Liu. Autoregressive stylized motion synthesis with generative flow. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13607–13607, 2021.

[152] A. Mark Williams, Paul Ward, and Craig Chapman. Training perceptual skill in field hockey: Is there transfer from the laboratory to the field? *Research Quarterly for Exercise and Sport*, 74(1):98–103, 2003.

[153] Daniel M. Wolpert and Zoubin Ghahramani. Computational principles of movement neuroscience. *Nature Neuroscience*, 3(11):1212–1217, Nov 2000.

[154] Mikołaj P. Woźniak, Julia Dominiak, Michał Pieprzowski, Piotr Ładoński, Krzysztof Grudzień, Lars Lischke, Andrzej Romanowski, and Paweł W. Woźniak. Subtletee: Augmenting posture awareness for beginner golfers. *Proc. ACM Hum.-Comput. Interact.*, 4(ISS), nov 2020.

[155] Erwin Wu and Hideki Koike. Futurepose - mixed reality martial arts training using real-time 3d human pose forecasting with a rgb camera. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1384–1392, 2019.

[156] Erwin Wu and Hideki Koike. Futurepong: Real-time table tennis trajectory forecasting using pose prediction network. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery.

[157] Xiangyu Xu and Chen Change. 3d human texture estimation from a single image with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13829–13838, 2021.

[158] Zou Yang, Sun Jie, Lu Shiqi, Cai Ping, and Niu Shengjia. Tangible interactive upper limb training device. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*, DIS '18 Companion, page 1–5, New York, NY, USA, 2018. Association for Computing Machinery.

[159] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2987–2997, 2020.

[160] Min Zeng, Le T. Nguyen, Baosheng Yu, Ole J. Mengshoel, Jun Zhu, Philip Wu, and Jie Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services*, pages 197–205. ACM, 2014.

[161] Chuanting Zhang, Haixia Zhang, Jingping Qiao, Dongfeng Yuan, and Minggao Zhang. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. *IEEE Journal on Selected Areas in Communications*, 37(6):1389–1401, 2019.

[162] Huan Zhao, Ashwaq Zaini Amat, Miroslava Migovich, Amy Swanson, Amy S. Weitlauf, Zachary Warren, and Nilanjan Sarkar. C-hg: A collaborative haptic-gripper fine motor skill training system for children with autism spectrum disorder. *ACM Trans. Access. Comput.*, 14(2), jul 2021.

[163] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2019.

[164] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *ArXiv*, abs/2012.06567, 2020.