論文 / 著書情報
Article / Book Information

Title	Integrating Generative and Contrastive Approaches for Human Action Recognition
Authors	Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, Koichi Shinoda
Citation	IEEE Access, vol. 13, , pp. 100095-100104
Pub. date	2025, 6
DOI	https://doi.org/10.1109/ACCESS.2025.3575707
Creative Commons	Information is in the article.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 March 2025, accepted 14 May 2025, date of publication 2 June 2025, date of current version 13 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3575707

APPLIED RESEARCH

Integrating Generative and Contrastive Approaches for Human Action Recognition

PABLO CERVANTES[1], YUSUKE SEKIKAWA[2], IKURO SATO[1,2], (Associate Member, IEEE), AND KOICHI SHINODA[1], (Senior Member, IEEE)

[1]Institute of Science Tokyo (formerly Tokyo Institute of Technology), Tokyo 152-8550, Japan
[2]Denso IT Laboratory, Minato City 105-0004, Japan

Corresponding author: Pablo Cervantes (cervantes@ks.c.titech.ac.jp)

This work was supported by the research project, "Development of Quality Foundation for Machine-Learning Applications," funded by Denso IT Laboratory Recognition and Learning Algorithm Collaborative Research Chair (Science Tokyo).

ABSTRACT This study introduces a novel approach to unsupervised skeleton-based human action recognition by integrating generative and contrastive learning methods. We propose a decomposition of representations, allowing for the preservation of detailed motion information for the generative learning objective while also extracting action features for the contrastive learning objective. By swapping contrastive representations between positive pairs (coining the name SwapCLR), we ensure that the generative and contrastive representations are complementary and both objectives contribute to learning a strong representation for downstream tasks like action recognition. Additionally, we address the challenge of noisy data in skeleton-based action recognition with a new saturating reconstruction loss, significantly reducing the impact of noise common to key-point detections. Our method demonstrates state-of-the-art performance in unsupervised action recognition on the NTU and PKU-MMD datasets, while also enabling generative downstream tasks such as motion in-painting and motion generation. Overall, these experimental results confirm the method's effectiveness and suggest its applicability to a variety of action analysis tasks.

INDEX TERMS Generative and contrastive, representation learning, unsupervised 3D action recognition.

I. INTRODUCTION

Human action recognition is a key component of any system built to interact with humans, such as healthcare monitoring, sports analysis, and human-computer interfaces. Skeleton-based action recognition is of particular interest for these applications because skeletons are light-weight representations of humans that generalize well between different subjects and scenes.

The complexity of human actions makes action recognition a challenging task. For skeleton-based action recognition this problem is exacerbated by the difficulty and cost of collecting annotated training data. Unsupervised action recognition addresses this issue by using unlabeled data and instead imposes prior knowledge through pre-text tasks, pseudo-labels, or data augmentation.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

Unsupervised learning is typically used to optimize a model's ability to infer robust and informative representations. This model is then fine-tuned on a small set of annotated data for a target task. Previous studies on unsupervised action recognition have explored the use of generative modeling with pre-text tasks such as motion prediction, jigsaw puzzle recognition, or denoising [15], [20], [42]. Conceptually these studies rely on the assumption that reconstructing motion details from a representation requires that representation to be informative about the motion's action class. Other studies have explored the use of contrastive learning [14], [21], [27], [30], [41], which enforces its representations to become invariant to augmentations that preserve the labels. As shown in fig. 1, both approaches focus on different, complementary features.

Combining generative and contrastive approaches is not trivial because the objectives of both approaches are in conflict. Contrastive approaches aim to remove superfluous information from its representations, while generative

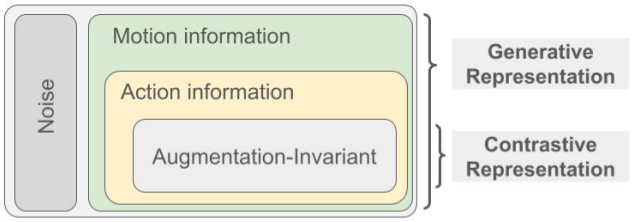

FIGURE 1. The information in recorded human actions consists of motion information and noise. The motion information is composed of components describing the performed action, the actor's identity, the style of the action, etc. In this work for the task of action recognition we focus on the action information. We argue that contrastive learning targets a subset of features within action information, whereas generative learning captures a more comprehensive superset. We therefore complement features from both approaches.

approaches aim to preserve all information in its representations. Neither approach matches the action information as shown in fig. 1. This conflict is avoided in previous studies [5], [12], [20], [34], [40] by decoupling the representations of both approaches. Although this avoids the conflict between both objectives, optimizing each objective independently may not find complementary features in the combination of both approaches.

We follow the conventional approach and decompose our representations into a generative and a contrastive representation to avoid the conflict. However, unlike previous approaches, we ensure that the features learned from both approaches complement each other by including the contrastive representation together with the generative representation during reconstruction. To ensure that the contrastive representation only picks up augmentation-invariant features, we swap the contrastive representations of positive pairs during reconstruction. This scheme ensures that both the generative and contrastive approach can find complementary solutions.

While generative modeling is a powerful tool for unsupervised learning, the sensitivity of the reconstruction loss to noise imposes some requirements on the quality of the training data. Specifically skeleton-based human action recognition depends on detected key-points which show various noise characteristics dependent the appearance, viewpoint or distance of the subjects to the camera. Some generative models such as [17] and [18] have been designed to improve their robustness to noise. However, the noise common in estimated human skeletons, particularly large distortions due to scene- or self-occlusions, is not well addressed.

We propose a saturating reconstruction loss that mitigates the impact of noise characteristic to estimated skeleton poses. More specifically, we observe that during training models are able to quickly reconstruct clean segments of a motion, while noisy segments are often not well reconstructed. To reduce the effects of noisy time-steps on our representations, we utilize a saturating reconstruction loss that discounts time-steps above a pre-determined threshold. We show that this improves the utility of our representations for downstream-tasks such as action recognition.

In summary, in this work we propose SwapCLR which enables the use of noisy data without annotations to train state-of-the-art action recognition models. Our key contributions are as follows:

- We introduce an integration of a contrastive and a generative approach to learning 3D action representations. This integration resolves the conflict in the training objectives of contrastive and generative approaches while ensuring that both approaches complement each other.
- To overcome the sensitivity to noise, we propose a saturating reconstruction loss that is robust to noise characteristic for estimated keypoints of human skeletons and improves training generative models on highly noisy datasets.
- SwapCLR achieves state-of-the-art performance in unsupervised action recognition on the NTU and PKU-MMD datasets.

II. RELATED WORKS

A. SELF-SUPERVISED LEARNING

Self-supervised approaches can be categorized as either generative, contrastive or generative-contrastive [25]. While purely generative models perform well on tasks such as data reconstruction or generation, they are usually not competitive when it comes to discriminative downstream tasks such as action recognition.

Contrastive Learning is based on contrasting samples against each other and pulling the representations of semantically similar samples closer, while pushing semantically dissimilar samples apart. A common approach is to construct a batch of positive and negative pairs and apply the InfoNCE loss [6], [29], [32] to compare all samples within this batch to update the encoder.

These approaches often observe improvements from larger batch-sizes, which suggests that a greater diversity in the samples compared by the contrastive loss leads to improved results. Momentum contrastive methods [13] follow this intuition and efficiently increase the diversity, by constructing a large queue of representations through a momentum-model. By considering a larger number of samples the true underlying distribution of the data can be better approximated.

Negative pairs are often critical in preventing the representation space from collapsing because positive pairs only contract the space. However, the construction of negative pairs is rarely perfect and thus with a larger number of negative pairs the number of false negatives increases as well. Therefore approaches such as BYOL [10] and SimSiam [7] and VicReg [2] have explored methods that don't rely on negative pairs to avoid the collapsing problem.

Generative-contrastive methods have the benefit of not having to contend with the collapsing problem because most methods explicitly enforce a prior distribution over the representation space. Yet, the generative model can

improve performance for discriminative down-stream tasks by incorporating a contrastive approach. Studies such as [3] in particular highlight that generative and contrastive representations are complementary.

B. ACTION RECOGNITION

Action recognition can be performed with various modalities such as video, depth images or wearable sensors. Among these skeleton-based action recognition [8], [9], [19], [36] has proven particularly successful. One downside of these methods is that collecting and annotating training data is expensive.

Addressing the issue of requiring labelled training data, unsupervised contrastive methods for skeleton-based action recognition such as [14], [21], [22], [27], [28], [30], and [41] were developed. One line of research focuses on enhancing contrastive learning by expanding the pool of positive samples. Notable studies in this direction include CPM [41] and CMD [27], which leverage strategies such as mining nearest neighbors across different modalities to expand the set of positive pairs, providing a robust learning signal.

Another research avenue emphasizes incorporating domain-specific knowledge of the human body and human actions [14], [21]. For example, ActCLR [21] leverages the insight that human actions often emphasize a subset of the skeleton sequence, referred to as actionlets. By decomposing actions into actionlets and recombining them during training, ActCLR introduces a domain-informed training paradigm that has demonstrated significant efficacy.

Unsupervised generative methods such as [1], [15], [16], [26], [33], and [37] employ pre-text task such as masked motion prediction. Reference [1] specifically addresses the issue of generative models when dealing with noisy data, but reconstructing masked latent representations rather than the noisy raw data.

Combining generative and contrastive approaches to learn 3D action representations has been explored in previous studies such as [5], [12], [20], [34], and [40]. These studies avoid the conflict between both objectives by using projection layers between the generative representation and the contrastive representation. This allows the generative representation to maintain all motion details, while the projection layers can focus on the features relevant for the contrastive objective, effectively decoupling both objectives.

We argue that resolving the conflict between contrastive and generative objective by decoupling them does not take full advantage of their integration. Instead, we decompose the representations, similarly to previous studies, but integrate both representations in the solutions of both generative and contrastive objective. This ensures that the representations learn complementary features.

III. METHODOLOGY

In this section we first introduce the general architecture of SwapCLR (section III-A) and how the latent representation is decomposed into a contrastive and a generative representation. Then, we define the optimization objectives (section III-B) including a saturating reconstruction loss that improves the performance of the optimized model for downstream tasks such as action recognition. Finally, we describe the training procedure using contrastive representations exchanged between positive pairs (section III-C).

A. SKELETON-BASED ACTION RECOGNITION

We consider human actions represented by a sequence of low-dimensional skeletons. Formally, we denote a skeleton pose of sequence i at time t as $x_t^i \in \mathbb{R}^{P\times B}$ where P is the number of joints and B is the dimensionality of the joint representation. In this work we focus on fixed-length sequences and denote sequence i of length T as $\mathbf{x}^i = \{x_t^i\}_{t=1}^T$. Because some action samples contain multiple actors, each actor's actions are stacked resulting in actions of the shape $\mathbf{x}^i \in \mathbb{R}^{M\times T\times P\times B}$ with M as the number of actors.

We employ an auto-encoder architecture as outlined in fig. 2, constructed of an encoder E parameterized by θ_{enc} and a decoder D parameterized by θ_{dec}. In the case of multiple input and output modalities we describe the ensemble of encoders and decoders with the same notation. A latent representation c is inferred as

$$c^i = E(\mathbf{x}^i; \theta_{\text{enc}}) \text{ with } c^i \in \mathbb{R}^{D\times M\times C}. \tag{1}$$

with D as the number of input modalities and C as the number of dimensions of the representation.

This latent representation is processed with a latent model f parameterized by θ_{var} and θ_{clr} resulting in the generative latent representation z_{gen}^i

$$z_{\text{gen}}^i = f(c^i, \theta_{\text{var}}, \theta_{\text{clr}}) \text{ with } z^i \in \mathbb{R}^{D\times M\times Z} \tag{2}$$

with Z as the number of dimensions of the representation.

Given a latent representation z_{gen}^i the decoder constructs a motion $\hat{\mathbf{x}}$ as

$$\hat{\mathbf{x}}^i = D(z_{\text{gen}}^i; \theta_{\text{dec}}). \tag{3}$$

With multiple input modalities the representations of all modalities are averaged before reconstruction, resulting in a representation $\hat{z}_{\text{gen}} \in \mathbb{R}^{M\times Z}$. For the encoder, we choose a STGCN network [35] and for the decoder we choose an MLP for its simplicity and demonstrated performance on a motion generation task [4].

The decomposition of latent representation c^i into a variational representation $z_{\text{var}} \in \mathbb{R}^{D\times M\times (Z/2)}$ and a contrastive representation $z_{\text{clr}} \in \mathbb{R}^{D\times M\times (Z/2)}$ is handled by the layers f_{var} and f_{clr} respectively. with $[\cdot|\cdot]$ denoting the concatenation operation.

$$f(c^i, \theta_{\text{var}}, \theta_{\text{clr}}) = [z_{\text{var}}^i | z_{\text{clr}}^i] \begin{cases} z_{\text{var}}^i = f_{\text{var}}(c^i, \theta_{\text{var}}) \\ z_{\text{clr}}^i = f_{\text{clr}}(c^i, \theta_{\text{clr}}) \end{cases} \tag{4}$$

The variational representation z_{var} is inferred by f_{var} through variational inference, assuming a Gaussian distribution over z_{var}. The parameters of this distribution are inferred

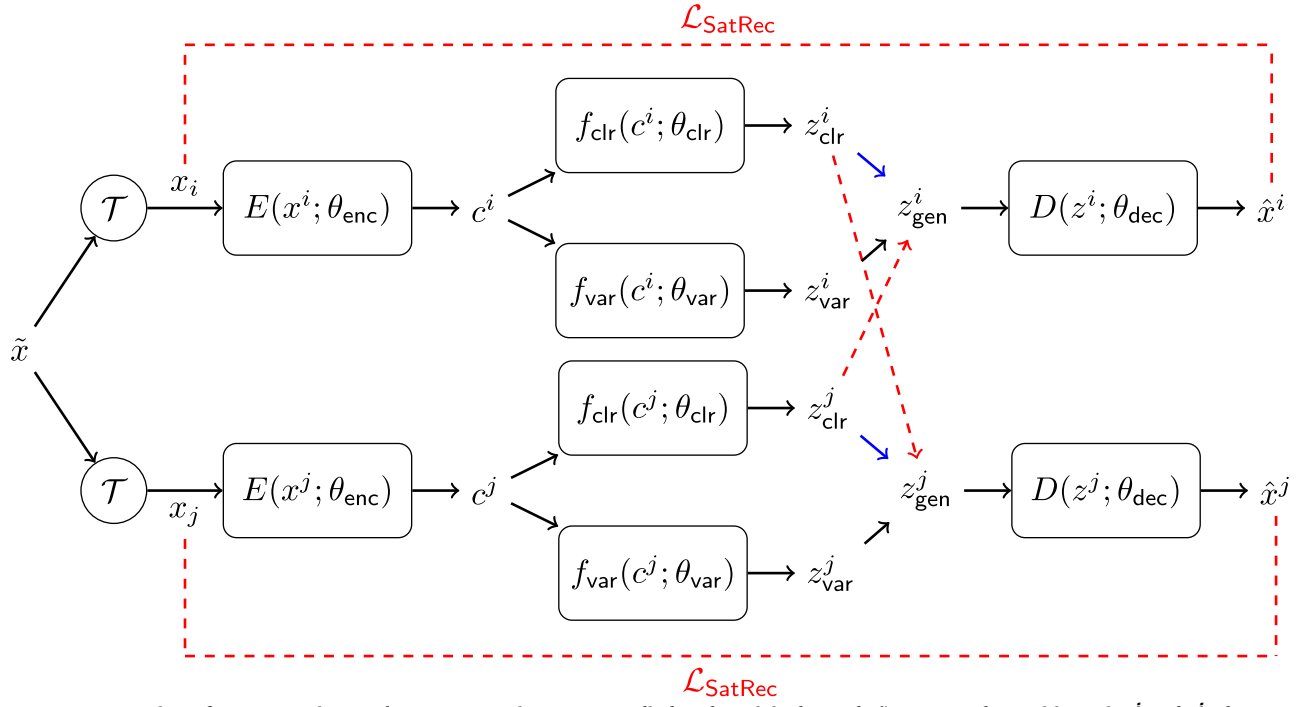

FIGURE 2. Overview of SwapCLR. First random augmentations τ are applied to the original sample $\tilde{x}$ to create the positive pair x^i and x^j. Then contrastive and variational representations z_{clr} and z_{var} are inferred and a generative representation z_{gen} is constructed from these two components. During training, the generative representation z^i_{gen} is constructed by incorporating the contrastive representation z^j_{clr} of the positive partner as indicated by the red line, while during inference the samples own contrastive representation is used (blue line). This exchange of representations within a positive pair motivates the name SwapCLR.

through optimized linear projections $\mu(c;\theta_{\text{var}})$ and $\Sigma(c;\theta_{\text{var}})$. During training $z_{\text{var}} \sim \mathcal{N}(\mu(c;\theta_{\text{var}}), \Sigma(c;\theta_{\text{var}}))$ is sampled from this Gaussian distribution and during inference z_{var} is produced by $z_{\text{var}} = \mu(c;\theta_{\text{var}})$. The contrastive representation z_{clr} is produced by f_{clr} a simple MLP.

B. OBJECTIVE FUNCTION

The objective function consists of the saturating reconstruction loss $\mathcal{L}_{\text{SatRec}}$, the contrastive loss $\mathcal{L}_{\text{clr}}$, the variational regularization $\mathcal{L}_{\text{kld}}$ and a weight decay term for the contrastive representation $\mathcal{L}_{\text{wd}}$ giving the combined loss as

$$\mathcal{L} = \mathcal{L}_{\text{SatRec}} + \mathcal{L}_{\text{clr}} + \mathcal{L}_{\text{kld}} + \mathcal{L}_{\text{wd}}. \tag{5}$$

To be able to handle noisy data and particularly the noise common to keypoint detections of a human skeleton, we formulate a saturating reconstruction loss to discount large errors.

$$\mathcal{L}_{\text{SatRec}} = 2\left(\frac{1}{1+e^{-\gamma\|\hat{\mathbf{x}}^i-\mathbf{x}^i\|_1}} - 0.5\right) \tag{6}$$

The hyperparameter γ governs the rate of saturation. When selecting γ, it's important to consider typical error magnitudes: large outliers should be placed in the saturated regime to discount their effect, while regular errors should remain in the linear regime where the loss is proportional to the error.

We formulate the contrastive loss for the contrastive representation z_{clr}, a set of positive pairs $\mathcal{P}$, a set of negative pairs as $\mathcal{N}$ and a temperature hyper-parameter τ as

$$\mathcal{L}_{\text{clr}} = -\log \frac{\sum_{p=1}^{\mathcal{P}} \exp((z^i_{\text{clr}} \cdot z^p_{\text{clr}})/\tau)}{\sum_{n=1}^{\mathcal{N}} \exp((z^i_{\text{clr}} \cdot z^n_{\text{clr}})/\tau)}. \tag{7}$$

The choice to remove the contribution term of positive pairs to the denominator is inspired by [38] and motivated by the intuition that hard positive pairs are discounted by this term.

Increasing the diversity of representations in the set of positive pairs is critical for good performance. Therefore, we construct the positive sets from contrastive representations across different modalities and across different actors. The negative pairs are constructed from all other samples in the same batch.

We employ the Kullbeck-Leibler divergence $\mathcal{D}_{\text{KL}}$

$$\mathcal{L}_{\text{kld}} = \mathcal{D}_{\text{KL}}(\mathcal{N}(\mu(c^i;\theta_{\text{var}}), \Sigma(c^i;\theta_{\text{var}}))\|\mathcal{N}(0, I)) \tag{8}$$

as regularization for the variational inference of the generative representation.

We also apply a weight decay as regularization for the contrastive regularization formulated as

$$\mathcal{L}_{\text{wd}} = \|z^i_{\text{clr}}\|_2. \tag{9}$$

C. TRAINING PROCEDURE

Contrastive learning seeks to learn informative representations by optimizing for invariance between augmented, yet semantically identical samples. This fosters robust and semantically meaningful representations. We seek to support

this contrastive objective by aligning it with the generative objective's reconstruction loss. More specifically, we swap the contrastive representations between two transformed samples during the reconstruction of each sample.

Denoting the indices of two transformed views of the sample as i and j, we train by swapping the contrastive representations of both samples during reconstruction such that the reconstruction of action $\mathbf{x}^i$ is inferred through

$$\bar{z}^i = [z^i_{\text{var}} | z^j_{\text{clr}}] \tag{10}$$

$$\hat{\mathbf{x}}^i = D(\bar{z}^i; \theta_{\text{dec}}). \tag{11}$$

Because the contrastive representation of the view j is agnostic to the transformation i, it can't contribute to the reconstruction of detailed motions. It can however provide information about the action class of the motion.

A more general formulation of this swapping scheme is to consider interpolation between the two contrastive representations with

$$\bar{z}^i_{\text{clr}} = \lambda z^i_{\text{clr}} + (1 - \lambda) z^j_{\text{clr}} \tag{12}$$

with $\lambda \in [0, 1]$ as the interpolation coefficient. However, in our experiments we have found little benefit from this. Similarly no improvements were found from a weighted average of more than two contrastive representations.

Note that this architecture does not explicitly avoid degenerate solutions where the decoder is insensitive to the contrastive representation and exclusively relies on the generative representation. The only mechanism that can steer the training away from this degenerate solution is the regularizing $\mathcal{L}_{\text{kld}}$ loss, discouraging over-reliance on the generative representation. However, while there is no other explicit mechanism discouraging this degenerate solution, there is also no reason for the decoder to ignore an informative representation. We verify in experiments that without a meaningful contrastive representation the reconstruction performance of the decoder significantly deteriorates.

IV. EXPERIMENTS

A. IMPLEMENTATION

1) MODALITIES

As input modalities for the encoder we follow previous studies [22] and consider the joint location (Joint), joint velocities (Motion) computed as the difference between two subsequent time-steps and joint connections (Bone) computed as the vector between two connected joints. For all modalities, missing joint data is replaced with zeros, and its corresponding reconstruction error is then excluded from the loss calculation. Variable-length sequences are interpolated to be fixed-length.

As output modalities of the decoder we predict joint locations (Joint), joint angles (Angles) represented as axis-angle rotations and joint velocities (Motion) computed as the vector between two subsequent time-steps.

We observe that while it is beneficial for the input joint representation to be in reference to the global origin, it is beneficial for the output joint representation to be in reference to a root joint, with only the root joint representing global motion. A similar discussion can be found in [15].

To compute the reconstruction loss for the axis-angle rotation, we perform forward-kinematics with the joint angles to produce a skeleton with joint locations that can be compared to the ground-truth. For the forward-kinematics we use bone-lengths of the sample to be reconstructed.

2) DATA AUGMENTATION

We use a shearing and temporal crop operation as data augmentations as introduced in [39]. The shearing augmentation simulates different camera angles, while the temporal crop encourages invariance to different performance speeds of the same action.

3) NETWORK ARCHITECTURE

The encoder is a ST-GCN as proposed in [35] with a latent dimension of 1024 and 9,871,364 parameters (per modality) in total.

The latent model decomposes the latent representation by the encoder into a contrastive and a variational representation, each with a dimension of 128. The concatenation of these representations forms the generative representation with 256 dimensions. The non-linear projections for this add up to 3,542,912 parameters (per modality).

The decoder is an MLP with 1,920,078 parameters per modality. This MLP follows the paradigm of implicit neural representations and thus receives the generative representation as well as a positional embedding with 256 dimensions and produces the skeleton representation of the time-step corresponding to the positional embedding.

4) TRAINING PROTOCOL

We train the encoder of our model with a SGD optimizer with a learning rate of 0.1, momentum of 0.9 and a weight decay of $1e^{-4}$ and the decoder with an ADAM optimizer with a learning rate of $1e^{-}3$ for 500 epochs. The learning rate of the encoder is dropped to 0.01 after 200 epochs.

5) EVALUATION METRICS

To measure the quality of our representations, we use the linear evaluation protocol, which seeks to evaluate the linear separability of the produced representations. This is evaluated by optimizing a linear classifier (supervised) on the representations of the training set extracted by a frozen encoder and evaluating the accuracy of the classifier on the test set.

In the case of multiple encoders (one per modality) we train a linear classifier per encoder and average the scores of all classifiers for a joint prediction. Note, that this ensemble of linear classifiers is a linear classifier itself. We optimize the linear classifier over 500 epochs, with a learning rate of 0.1 (dropped to 0.01 after 200 epochs) with the Adam optimizer.

B. DATASETS

1) NTU RGB+D 60

[31]: NTU-RGB+D 60 (NTU-60) contains 60 action categories performed by 40 actors resulting in 56,880 sequences captured by three Kinectv2 cameras each from a distinct viewing perspective. Out of the 60 action classes, 10 classes involve interactions between two actors. The recorded data contain RGB videos, depth map sequences, 3D skeletal data, and infrared (IR) video. In this work, we only use the 3D skeleton data, which uses skeletons described by 25 body joints in 3D coordinates. Both cross-view (xview) and cross-subject (xsub) splits are available, where the training set contains distinct viewpoints or subjects respectively from the evaluation set.

2) NTU RGB+D 120

[24]: NTU-RGB+D 120 (NTU-120) extends NTU-60 by adding 60 classes resulting in 114,480 skeleton sequences in total. The number of action classes with two actors is increased by 15 and the number of actors is increased by 66. Both cross-subject (xview) and cross-setup (xset) splits are available, where the training set contains either distinct viewpoints or setups respectively from the evaluation set.

3) PKU MMD

[23]: PKU-MMD is a multi-modality 3D human action recognition dataset [23] consisting of two phases. We focus on the cross-subject evaluation protocol for Phase II (PKU-II) containing 5,332 skeleton sequences for training and 1,613 for testing.

C. EVALUATION

To show the effectiveness of our method, we compare its performance in table 1, table 2 and table 3 to the unsupervised state of the art in the downstream task of action recognition. We find that our method outperforms the state of the art on the NTU-RGBD xview and the PKU-II datasets and reaches comparable performance to the state of the art on other datasets.

The fact that our model does not outperform the state of the art on the cross-subject (xsub) and cross-setup (xset) splits of the NTU datasets suggests that the transformer model architecture in S-JEPA [1] can take better advantage of the larger dataset size of NTU120.

D. ABLATION STUDIES

To identify the impact of the different components of our proposed method, we perform ablation studies in table 4. Comparing the purely contrastive, encoder-only approach (first row) with the generative approaches (second and third row), we can observe that using a representation swapping scheme alone is not sufficient for the generative approaches to perform on-par with the contrastive approach. However, an integration of both approaches (last two rows) can outperform each approach by itself, providing evidence for our hypothesis that combining generative and contrastive methods can improve performance.

Comparing the two integrated methods (last two rows), one simple autoencoder and one autoencoder, with the proposed representation swapping scheme (SwapCLR), we find that

TABLE 1. Comparison to unsupervised action recognition methods on the NTU-RGBD-60 dataset. Methods are categorized as contrastive (clr) or contrastive + generative (gen + clr). 3s denotes three stream solutions with multiple skeleton modalities. Best performing results per dataset in bold.

Method	Type	NTU60	
		xview	xsub
AimCLR [11]	clr	79.7	74.3
GL-Transformer [15]	clr	83.8	76.3
CPM [41]	clr	84.9	78.7
CMD [27]	clr	86.9	79.8
3s-CrosSCLR [22]	clr	83.4	77.8
3s-AimCLR [11]	clr	83.8	78.9
3s-CPM [41]	clr	87.0	83.2
3s-SkeleMixCLR [39]	clr	87.1	82.7
3s-ActCLR [21]	clr	88.8	84.3
3s-CMD [27]	clr	90.9	84.1
S-JEPA [1]	gen	89.8	**85.3**
MS^{2}L [20]	gen + clr	52.6	-
CP-STN [40]	gen + clr	76.6	69.4
3s-SwapCLR (ours)	gen + clr	**91.1**	83.3

TABLE 2. Comparison to unsupervised action recognition methods on the NTU-RGBD-120. Methods are categorized as contrastive (clr) or contrastive + generative (gen + clr). 3s denotes three stream solutions with multiple skeleton modalities. Best performing results per dataset in bold.

Method	Type	NTU120	
		xsub	xset
AimCLR [11]	clr	63.4	63.4
GL-Transformer [15]	clr	68.7	66.0
CPM [41]	clr	68.7	69.6
CMD [27]	clr	70.3	71.5
3s-CrosSCLR [22]	clr	66.7	67.9
3s-AimCLR [11]	clr	68.6	68.2
3s-CPM [41]	clr	74.0	73.0
3s-SkeleMixCLR [39]	clr	70.5	70.7
3s-ActCLR [21]	clr	74.3	75.7
3s-CMD [27]	clr	74.7	76.1
S-JEPA [1]	gen	**79.6**	**79.9**
CP-STN [40]	gen + clr	55.7	54.7
3s-SwapCLR (ours)	gen + clr	73.7	75.1

TABLE 3. Comparison to unsupervised action recognition methods on the PKU-II. Methods are categorized as contrastive (clr) or contrastive + generative (gen + clr). 3s denotes three stream solutions with multiple skeleton modalities. Best performing results per dataset in bold.

Method	Type	PKU-II
		xsub
CPM [41]	clr	48.3
CMD [27]	clr	43.0
3s-CPM [41]	clr	51.5
3s-SkeleMixCLR [39]	clr	57.1
3s-CMD [27]	clr	52.6
MS^{2}L [20]	gen + clr	27.6
3s-SwapCLR (ours)	gen + clr	**57.2**

TABLE 4. Comparison of auto-encoders trained with various combinations of loss functions and models on NTU-60 xview. All methods besides the encoder-only model were also trained with the KLD and weight decay losses.

Method	Loss	Acc [%]
Encoder-only	InfoNCE	86.7
Auto-Encoder (with swapping)	MSE	63.4
Auto-Encoder (with swapping)	SatRec	83.3
Auto-Encoder	SatRec + InfoNCE	90.1
SwapCLR	SatRec + InfoNCE	91.1

the proposed scheme leads to a significant performance improvement. This suggests that the representation swapping scheme in SwapCLR allows the generative and contrastive scheme to focus on complementary features that lead to more informative and robust representations.

The improvements from introducing the proposed saturating reconstruction loss can be clearly seen by comparing the experiments with a regular MSE loss (second row) and the saturating reconstruction loss (third row). This lets us conclude that the noise characteristic to key-point detections as found in NTU-60 is a significant issue for generative methods and is effectively addressed by the proposed saturating reconstruction loss.

a: MODALITIES

Previous studies have shown the effectiveness of considering multiple modalities as input to the encoder. We expand on these studies by investigating the effects of decoders with different generated modalities in table 5.

Our results align with the observations of previous studies that multiple input modalities are critical to reaching state-of-the-art performance. Multiple output modalities also encourage the model to learn more discriminative features and push our model to outperform the state-of-the-art. Note, that the performance improvements by combining multiple output modalities are modest, suggesting that the task of reconstructing different modalities does not encourage the encoders to learn more diverse features. Even an encoder with only a single modalities does not benefit from training with decoders with multiple output modalities (see the second row).

Comparing the different output modalities, we find that the modality that leads to the most informative representations is joint positions (J), followed by joint angles (A) and joint motions (M). A notable difference between the joint positions and the joint angles is that the reconstruction of joint positions requires information about the subject-specific bone lengths in the learned representation. Joint angles on the other hand are reconstructed through forward kinematics, with subject-specific bone lengths explicitly provided from the ground-truth data. The gap in performance between models with the joint position and joint angle as output modalities suggests that in the NTU datasets subject-specific features are informative about the action class, potentially explaining the gap in performance on cross-subject experiments compared to works like [1].

TABLE 5. Comparison of an auto-encoder model trained with various input and output modalities. The modalities are joint positions (J), joint motions (M), bones (B) and joint angles as axis-angles (A).

Input modalities	Output modalities	Acc [%]
J	J	86.8
J	J + A + M	86.9
J + M + B	J	90.8
J + M + B	A	90.3
J + M + B	M	88.0
J + M + B	J + A + M	91.1

TABLE 6. Comparison of representations with different dimensionality.

Representation size	Accuracy [%]
1024	91.1
512	89.6
256	87.8

TABLE 7. Comparison of the different representations produced by the proposed work. Investigated are c^i from eq. (1), z_{gen} from eq. (2) and z_{clr} from eq. (4).

Representation	Accuracy [%]
c^i	91.1
z_{clr}	80.6
z_{gen}	83.6

1) REPRESENTATION SIZE

Previous studies have shown that high dimensional representations are more informative and improve performance for downstream tasks. We observe a similar effect in our model as shown in table 6 and find that while large representations (1024) are needed to outperform the state-of-the-art, even small representations (256) reach strong performance.

Compared to contrastive methods, the reconstruction loss of generative-contrastive approach demands more detailed features and our method may benefit from even larger representations. However, in the interest of a fair comparison to other studies and due to the large memory requirements of models with larger representations we refrain from such experiments.

2) REPRESENTATION COMPONENTS

We also study the quality of different components of our representations in table 7. This investigation reveals that the most informative variable is the latent representation c^i before the decomposition. However, the contrastive representation z_{clr} and generative representation z_{gen} contain meaningful features for action recognition, which proves that the model has not collapsed to a degenerate solution, but is indeed relying on both the contrastive and generative approach to learn informative representations.

E. GENERATIVE DOWNSTREAM TASKS

Besides showing state-of-the-art performance on discriminative downstream tasks such as action recognition, our

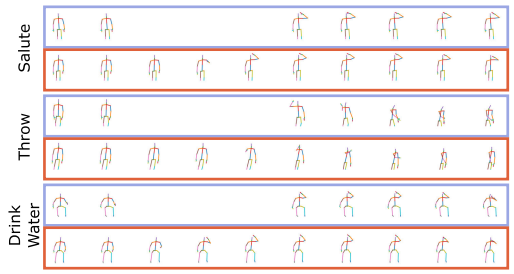

FIGURE 3. Motion in-painting experiments. From the groundtruth motion (blue) some frames were removed. The reconstruction of the motion (red) by our model is able to fill-in the missing frames.

FIGURE 4. Motion generation experiments. The groundtruth motion (blue) serves as seed motion that provides the contrastive representation, while a variational representation is randomly sampled from a normal distribution. The action label of the seed motion is preserved in the generated motion (red).

model is also able to perform generative downstream tasks. We qualitatively explore the tasks Motion In-Painting and Motion Generation.

1) MOTION IN-PAINTING

We explore the task of motion in-painting in fig. 3 with the goal to reconstruct frames that were missing in the original sample. For this we artificially remove the 20 frames from the 50 frame long input motions (removal means filling the frames with zeros). We then observe the reconstruction performance of the missing frames.

Our model is able to appropriately interpolated between frames in order to reconstruct the missing frames. This result suggests that our model is robust to missing frames and can infer missing information from context.

2) MOTION GENERATION

We also explore the task of motion generation in fig. 4 with the goal of generating new motions not included in the original dataset. More specifically, we perform conditional motion generation by sampling the representation of the new motion conditional on a groundtruth motion. By combining the contrastive representation z_{clr} of the seed motion with a variational representation z_{var} sampled from a normal distribution, we can generate a novel motion.

With this scheme our model is able to generate novel motions that are different from the seed motion but maintain the action label of the seed motion. This shows that the contrastive representation contains important cues about the action class that can be utilized by the decoder. Note that this allows our model to generate motions with a target action class fully unsupervised.

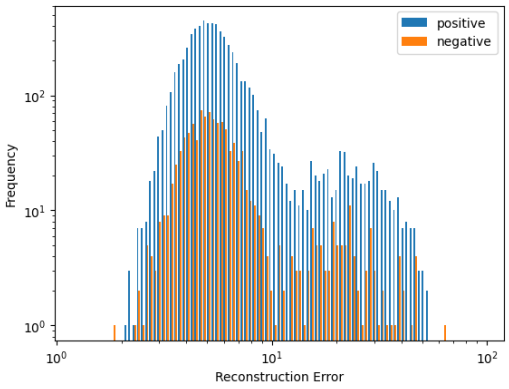

FIGURE 5. Histogram of the reconstruction error (MSE) on the NTU-RGBD60 datasets (test samples) of the joint position modality for positive (samples which where classified correctly) and negatives (samples which were classified incorrectly).

F. FEATURE STUDY

We seek to understand the integration of the generative and contrastive approach by comparing the reconstruction and recognition accuracy performance. We plot the distribution of the reconstruction loss of correctly and incorrectly classified samples (positive and negative samples) in fig. 5.

The distribution of the reconstruction error shows a bimodal distribution, suggesting an additional factor that leads to poor reconstruction performance. This factor may be the presence of an other actor that induces occlusions. This hypothesis is supported by the average reconstruction loss per class fig. 6, which shows that the average reconstruction error for actions with multiple actors (especially *walking towards each other* and *walking apart from each other*) is significantly higher.

We observe that both positive and negative samples have a similar distribution, suggesting that noise doesn't affect the recognition accuracy. This provides additional evidence that our proposed saturating reconstruction loss allows our model to focus on clean features and noisy features don't significantly degrade recognition performance.

We also investigate the average reconstruction error for positive and negative samples per class in fig. 7. We observe that for most actions the reconstruction error for negative samples is similar as for positive samples. Outliers like *sitting down*, *standing up (from sitting position)*, *staggering* and *falling* show a significantly higher reconstruction error for negative samples, suggesting that here noise may degrade recognition performance. Interestingly, *walking towards each other* is the only action where positive samples have a significantly higher reconstruction error, while *walking apart from each other* shows a significantly higher reconstruction error for negative samples. This may suggest that some noise pattern in *walking towards each other* is informative about the action class.

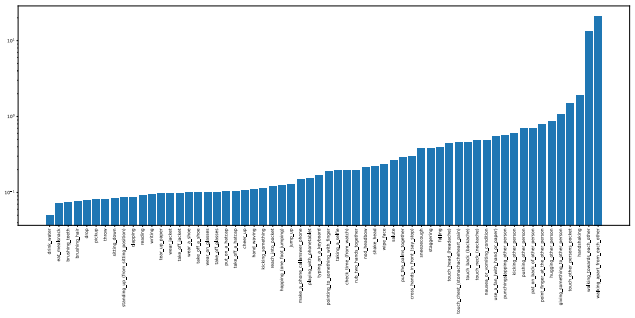

FIGURE 6. Average reconstruction error (MSE) per action class on the NTU-RGBD60 datasets (test samples).

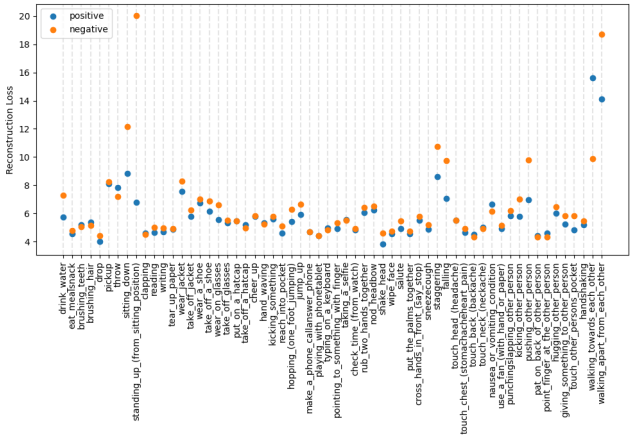

FIGURE 7. Comparison of the mean reconstruction error (MSE) on the NTU-RGBD60 datasets (test samples) for positive samples (samples which were classified correctly) and negative samples (samples which were classified incorrectly) per class.

V. LIMITATIONS AND FUTURE WORK

A limitation of using a generative approach for a discriminative downstream task is the additional requirement of training a decoder model that is discarded after training. Besides the additional computational and memory overhead of training the decoder, it increases the complexity of the model and requires careful tuning to achieve high performance.

Previous studies have shown improvements by adding a fine-tuning stage after the initial pre-training stage. A common strategy during this fine-tuning is to expand the set of positive pairs through clustering or other similarity criteria. Our generative approach would allow us to generate such positive pairs, effectively leveraging the information learned by the decoder to further improve the encoder. We leave the development of such a fine-tuning scheme to future work.

VI. CONCLUSION

In this work we propose an integration of a generative and contrastive approach to learn 3D action representations. This integration overcomes the conflict between the generative and contrastive objectives by constructing an action representation with both a generative and contrastive component. This allows the generative representation to contain all motion details, while allowing the contrastive representation to discard any features not invariant to augmentations. By swapping the contrastive representation between positive pairs and incorporating these during reconstruction, we ensure that the generative and contrastive representations learn complementary features.

Using a generative approach with a reconstruction loss induces a sensitivity to noise in the training data. 3D skeletons suffer from noisy keypoint detections and the extreme distortions can degrade the utility of the action representations. To address this issue, we introduce a saturating reconstruction loss, specifically designed to reduce sensitivity to large distortions.

We provide evidence to show the effectiveness of the method by reach state-of-the-art performance on the relevant datasets NTU and PKU-MMD and clearly showing the contributions of the different components of the proposed system in ablation studies.

ACKNOWLEDGMENT

This work is an outcome of a research project, Development of Quality Foundation for Machine-Learning Applications, supported by DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Science Tokyo). This work was supported by JSPS KAKENHI Grant Number JP23H00490.

REFERENCES

[1] M. Abdelfattah and A. Alahi, "S-jepa: A joint embedding predictive architecture for skeletal action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2024, pp. 1–15.

[2] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–15.

[3] F. Bordes, R. Balestriero, and P. Vincent, "High fidelity visualization of what your self-supervised representation knows about," 2021, *arXiv:2112.09164*.

[4] P. Cervantes, Y. Sekikawa, I. Sato, and K. Shinoda, "Implicit neural representations for variable length human motion generation," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 356–372.

[5] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint generative and contrastive learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2004–2013.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[7] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[8] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 180–189.

[9] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2959–2968.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, B. Piot, k. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 21271–21284.

[11] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 762–770.

[12] M. Halawa, O. Hellwich, and P. Bideau, "Action-based contrastive learning for trajectory prediction," in *Proc. ECCV*, 2022, pp. 143–159.
[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.
[14] Y. Hua, W. Wu, C. Zheng, A. Lu, M. Liu, C. Chen, and S. Wu, "Part aware contrastive learning for self-supervised action recognition," 2023, *arXiv:2305.00666*.
[15] B. Kim, H. J. Chang, J. Kim, and J. Y. Choi, "Global-local motion transformer for unsupervised skeleton-based action learning," in *Proc. ECCV*, 2022, pp. 209–225.
[16] D. Kim, D. Cho, and I. S. Kweon, "Self-supervised video representation learning with space-time cubic puzzles," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8545–8552.
[17] S.-J. Li, H.-S. Zhu, L.-P. Zheng, and L. Li, "A perceptual-based noise-agnostic 3D skeleton motion data refinement network," *IEEE Access*, vol. 8, pp. 52927–52940, 2020.
[18] S. Li, Y. Zhou, H. Zhu, W. Xie, Y. Zhao, and X. Liu, "Bidirectional recurrent autoencoder for 3D skeleton motion data refinement," *Comput. Graph.*, vol. 81, pp. 92–103, Jun. 2019, doi: 10.1016/j.cag.2019.03.010.
[19] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 906–915.
[20] L. Lin, S. Song, W. Yang, and J. Liu, "MS2L: Multi-task self-supervised learning for skeleton based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2490–2498.
[21] L. Lin, J. Zhang, and J. Liu, "Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2363–2372.
[22] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3D human action representation learning via cross-view consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4739–4748.
[23] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:1703.07475*.
[24] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.
[25] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.
[26] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, and H. Li, "Masked motion predictors are strong 3D action representation learners," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10147–10157.
[27] Y. Mao, W. Zhou, Z. Lu, J. Deng, and H. Li, "Cmd: Self-supervised 3D action representation learning with cross-modal mutual distillation," in *Proc. ECCV*, 2022, pp. 734–752.
[28] O. Moliner, S. Huang, and K. Åström, "Bootstrapped representation learning for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4153–4163.
[29] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[30] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition," *Inf. Sci.*, vol. 569, pp. 90–109, Aug. 2021.
[31] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "Ntu RGB+D A large scale dataset for 3D human activity analysis," in *Proc. CVPR*, Jun. 2016, pp. 1010–1019.
[32] K. Sohn, "Improved deep metric learning with multi-class n–pair loss objective," in *Proc. NeurIPS*, vol. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 1–15.
[33] K. Su, X. Liu, and E. Shlizerman, "PREDICT & CLUSTER: Unsupervised skeleton based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9628–9637.
[34] Y. Su, G. Lin, and Q. Wu, "Self-supervised 3D skeleton action representation learning with motion consistency and continuity," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13308–13318.
[35] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–18.
[36] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 588–597.
[37] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Skeleton cloud colorization for unsupervised 3D action representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13403–13413.
[38] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," in *Proc. ECCV*. Springer, 2022, pp. 668–684.
[39] Z. Chen, H. Liu, T. Guo, Z. Chen, P. Song, and H. Tang, "Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition," 2022, *arXiv:2207.03065*.
[40] Y. Zhan, Y. Chen, P. Ren, H. Sun, J. Wang, Q. Qi, and J. Liao, "Spatial temporal enhanced contrastive and pretext learning for skeleton-based action representation," in *Proc. ACML*, 2021, pp. 534–547.
[41] H. Zhang, Y. Hou, W. Zhang, and W. Li, "Contrastive positive mining for unsupervised 3D action representation learning," in *Proc. ECCV*, 2022, pp. 36–51.
[42] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2644–2651.

PABLO CERVANTES received the dual master's degrees in science and engineering from RWTH Aachen University, Germany, and Keio University, Japan, in 2018. He is currently pursuing the Ph.D. degree in artificial intelligence with the Institute of Science Tokyo (formerly Tokyo Institute of Technology).

YUSUKE SEKIKAWA received the B.S. degree in electrical engineering from the Science University of Tokyo, Japan, in 2004, and the Ph.D. degree in computer science from Keio University, in 2020. From 2004 to 2009, he was with Japanese Ministry of Economy, Trade and Industry, as a Patent Examiner. From 2008 to 2012, he was a Software Engineer with Olympus Imaging Company Ltd. In 2012, he joined Denso IT Laboratory, as a Computer Vision Researcher. From 2014 to 2015, he was the MIT Media Laboratory, as a Visiting Scientist. His research interests include machine learning and neuromorphic image processing.

IKURO SATO (Associate Member, IEEE) received the Ph.D. degree in physics from the University of Maryland, USA, in 2005. After working as a Postdoctoral Fellow with the Lawrence Berkeley National Laboratory, USA, he joined the Research and Development Group, Denso IT Laboratory, Japan, in 2008, where he conducted research on image recognition and machine learning for automotive applications. Since 2020, he has also been a Specially Appointed Associate Professor with the Institute of Science Tokyo (formerly known as Tokyo Institute of Technology), Japan.

KOICHI SHINODA (Senior Member, IEEE) received the B.S. and M.S. degrees in physics from The University of Tokyo, Tokyo, Japan, in 1987 and 1989, respectively, and the D.Eng. degree in computer science from Tokyo Institute of Technology, Japan, in 2001. In 1989, he joined NEC Corporation, Japan, where he was involved in research on automatic speech recognition. From 1997 to 1998, he was a Visiting Scholar with Bell Labs, Lucent Technologies, Murray Hill, NJ, USA. From October 2001 to March 2003, he was an Associate Professor with The University of Tokyo. He is currently a Professor with the Institute of Science Tokyo. His research interests include speech recognition, video information processing, and machine learning. He is a fellow of IEICE and a Senior Member of IPSJ. He was the General Chair of APSIPA 2021, the Chair of IEEE SPS Tokyo Joint Chapter (2021–2022), and the President-Elect of the Information and System Society of IEICE (2024–2025).