論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A Study on Recovering Camera Motion and 3-D Structures from Sequential Monocular Images
著者(和文)	JIANGZijie
Author(English)	Zijie Jiang
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12808号, 授与年月日:2024年6月30日, 学位の種別:課程博士, 審査員:奥富 正敏,塚越 秀行,中臺 一博,田中 正行,原 精一郎,川上 玲
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12808号, Conferred date:2024/6/30, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

A Study on Recovering Camera Motion and 3-D Structures from Sequential Monocular Images

Zijie Jiang

Supervisor: Prof. Masatoshi Okutomi

Department of Systems and Control Engineering
Tokyo Institute of Technology

This dissertation is submitted for the degree of
Doctor of Engineering

May 2024

Declaration

I, ZIJIE JIANG, declare that this thesis titled, 'A Study on Recovering Camera Motion and 3-D Structures from Sequential Monocular Images' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Zijie Jiang
May 2024

Abstract

Recovering camera motion and 3D scene structures simultaneously from a monocular image sequence is a fundamental problem within computer vision. Existing methods exhibit limited robustness and accuracy when confronted with challenging input images, characterized by issues such as motion blur, texture-less regions, and dynamic objects. This thesis introduces targeted methodologies designed to effectively tackle these challenges in estimating camera motion and 3D scene structures from monocular image sequences. To summarize, we propose 1) a robust and efficient Structure from Motion pipeline for accurate recovery of camera motion and 3D structures under challenging environments by fusing relative pose information from a visual-inertial odometry, 2) a novel self-supervised monocular ego-motion estimation network based-on multi-layer fusion of RGB and inferred depth information, 3) a novel self-supervised monocular scene flow estimation network capable of jointly estimating depth, dense SE3 motion field and ego-motion estimation from monocular images, 4) a depth-aided neural radiance fields approach for novel view synthesis in challenging monocular gastroscopy. Extensive experiments demonstrate a substantial enhancement in both robustness and accuracy achieved by our method compared to previous works.

Acknowledgements

First of all, I would like to thank my supervisor Professor Masatoshi Okutomi for his expertise, assistance, guidance, and patience throughout the process of composing this thesis. Without you, this manuscript will not come close to completion. I would like to thank Dr. Hajime Taira whose help, guidance, and comment became an enormous contribution to my works. I also would like to thank all members in Okutomi-Tanaka lab, who have not only made my daily life enjoyable and memorable, but also provided timely aid to my experiments. Specifically, I would like to express my gratitude to Zihua Liu and Yizhou Li from the laboratory, for their discussions and assistance in my research project, as well as valuable insights in other new research areas.

Finally, but most importantly, I sincerely thank my friends and family, who consistently support and encourage me.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background

Recovering both camera motion and 3D structures from sequential monocular images is a fundamental problem in computer vision, which constructs a minimal 3D perception system that solely requires a monocular camera. Fig.1.1 illustrates the general problem setting of this thesis. Compared to other 3D perception systems such as Lidar, time-of-flight sensors, and structured light cameras, monocular cameras present advantages in terms of affordability, simplicity in sensor setup, and versatile applicability, thus favored by researchers and manufacturers. However, due to the inherent limitations of monocular cameras in directly measuring 3D information, recovering camera motion and scene structure from monocular image sequences typically poses a challenge. Encouragingly, research on this topic has made tremendous progress over the past decade. We first give a literature review of the evolution of this field from three perspectives: 1) traditional geometry-based approaches, 2) learning-based data-driven approaches, and 3) approaches based on neural implicit representation. Then, we provide a general analysis of the shortcomings of previous methods and introduce the main contributions of this thesis.

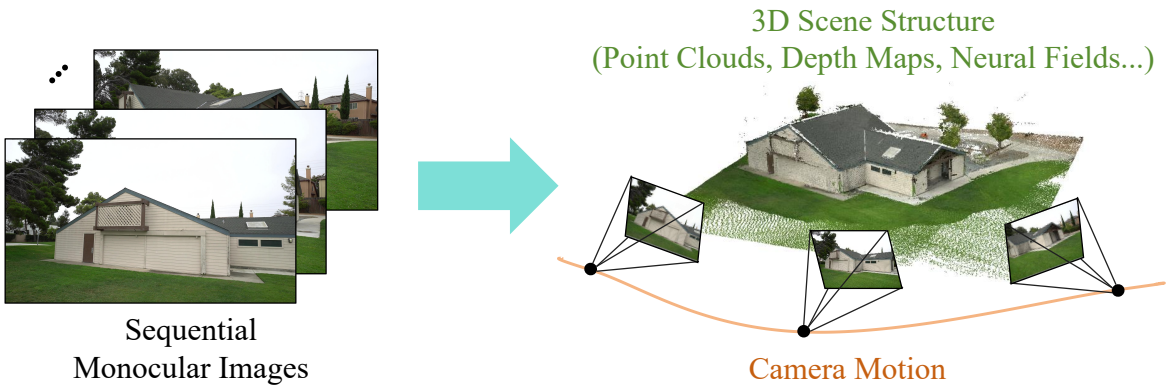


Fig. 1.1 **The general problem setting of this thesis.** We aim to recover both the camera motion and 3D scene structures from sequential monocular images. Depending on the approach, the 3D scene structures can take the form of point clouds, depth maps, or neural fields.

1.2 Literature Review

1.2.1 Traditional Geometry-based Approach

In the early stages, this topic is typically addressed by traditional geometry-based methods, commonly known as Structure from Motion (SfM) [138, 136] or visual simultaneous localization and mapping (vSLAM) [118, 117]. Fig. 1.2 demonstrates the general pipeline of a vSLAM algorithm. One key to the success of traditional SfM and vSLAM methods is the establishment of robust 2D-2D correspondences between image pairs [103], commonly referred to as feature matching in the context of SfM or feature tracking in the context of vSLAM. This procedure is achieved by computing the similarity between pixels or patches in two images, either based on hand-crafted feature descriptors like SIFT [103] or SURF [11], or directly based on pixel intensities [40]. The established 2D-2D correspondences are then used to recover the relative camera pose between the two images by estimating the fundamental matrix [105], and a triangulation method [60] is generally used to derive the associated 3D points from the 2D-2D correspondences and estimated relative pose. The above steps will be repeated to process the whole monocular image sequence, finally giving the camera pose of each image and the 3D structure represented by the reconstructed 3D point cloud. Typically, the recovered camera poses and 3D point cloud are jointly optimized multiple times using bundle adjustment [153] midway and at the end of processing, to obtain globally optimal results. With the improvement of various sub-modules, traditional SfM and vSLAM methods have made long-term progress

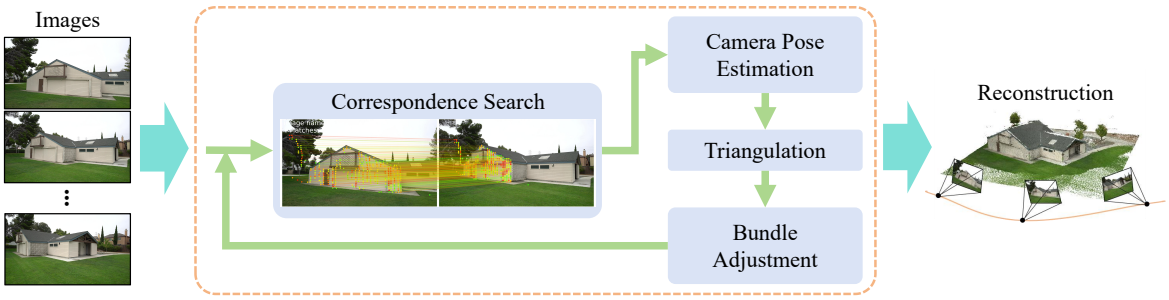


Fig. 1.2 **The general pipeline of a vSLAM algorithm.**

in the past years in terms of accuracy, efficiency, and application to large-scale datasets. At present, both traditional SfM methods and vSLAM methods have mature and easy-to-use open-sourced software [136, 117], that is suitable in a wide range of scenarios. However, these traditional methods still struggle to handle non-Lambertian and textureless scenes, where hand-crafted feature descriptors or pixel intensities are no longer reliable for establishing accurate 2D-2D correspondences across images. Besides, traditional methods have limited tolerance for the existence of noise in the estimated 2D-2D correspondences, and the accumulation of errors at each sub-stage of methods makes it difficult to obtain a stable solution to the optimization problem. All these factors make the traditional methods less robust in challenging scenarios and often yield suboptimal or failed results.

1.2.2 Learning-based Data-driven Approach

On the other hand, with the remarkable success of deep learning in various 2D visual tasks, *e.g.*object detection [50, 64] and image classification [86], researchers have begun exploring the integration of deep learning techniques into the 3D task of recovering camera pose and scene structure from monocular images, giving rise to a series of learning-based data-driven methods. At first, based on the characteristics of deep learning that can learn priors from a large amount of data, deep learning began to be used to solve some ill-posed problems, such as predicting depth from a single image. [38] first succeeded in training a convolutional neural network for depth prediction from a single image. Many excellent works [44, 129, 130, 14] have appeared in the follow-up to achieve better performance in monocular depth prediction by improving the network structure. This emerging monocular depth estimation technique can be naturally integrated into traditional SfM and vSLAM methods as an approach to generate RGB-D sequences from monocular sequences, thus traditional RGB-D-based methods can be applied [147]. Another feature of deep

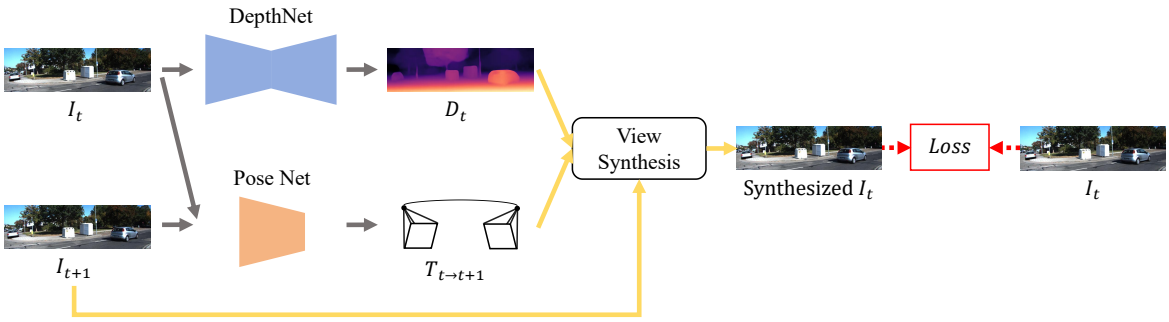


Fig. 1.3 **Self-supervised depth and camera motion estimation pipeline in SfM-Learner [190].**

learning is that it enables end-to-end learning, which inspired research on regressing relative camera poses directly from image pairs. DeMoN [155] is one of the first works to address the task of recovering both camera motions and 3D structures from monocular images as an end-to-end learning problem. The network architecture proposed in DeMoN imitates the pipeline of traditional SfM methods, where a bootstrap network is leveraged to estimate the initial depth and relative camera pose, and the subsequent series of iterative networks and a refinement network are leveraged to refine the existing depth and relative camera pose estimates. Subsequent works [148, 146] achieve further performance improvements by replacing the original generic network structure used to simulate feature matching, pose estimation, and iterative optimization with more interpretable network structures designed based on geometry and optimization theory. These methods have achieved more accurate and comprehensive estimates of camera trajectories and 3D structures compared to traditional approaches in small-scale reconstruction problems. However, due to the absence of some core modalities compared to a full SLAM system, such as loop closure and global bundle adjustment, a performance gap exists between them and traditional methods in large-scale reconstruction problems. DROID-SLAM [150] introduces a novel and full SLAM system equipped with an end-to-end neural architecture, and first outperforms prior traditional and learning-based methods by a large margin on challenging large-scale benchmarks. Despite the impressive achievements of these learning-based methods, all of them still require training in a supervised manner, necessitating a large amount of annotated data with ground-truth depth maps and camera poses. When the amount of training data is not sufficient, or there is a domain gap between the training data and the test data, the performance of these methods is significantly reduced.

To alleviate the reliance on labeled data, the application of self-supervised learning in addressing simultaneous camera motion and 3D structure recovery problems has also been continuously under attention and research. One core of self-supervised learning lies in the design of proxy loss. [190] and [52] concurrently proposed the utilization of the photometric loss based on differentiable image synthesis as the proxy loss. The distinction lies in that [52] employs stereo image pairs for training data, whereas [190] can leverage more easily available monocular image sequences for training. The network architecture of [190] consists of two independent networks, one for estimating the depth map from a single image and the other for estimating the relative camera pose from an image pair, as shown in Fig.1.3. Given a target image and a reference image, the depth map of the target image and the relative camera pose from the target image to the reference image are estimated separately. Then, a 'fake' target image is synthesized from the estimated depth of the target view and the relative camera pose, which follows the geometric constraints of Structure-from-Motion in a differentiable manner. The photometric difference between the original target image and the 'fake' target image constitutes the proxy loss to self-supervise the learning of the two independent networks to predict reasonable depth maps and relative camera poses. This initial design of the proxy loss was relatively simple and did not take into consideration various factors such as the existence of occlusion, dynamic objects, and non-Lambertian scenes during the loss calculation, which violates the assumption of multi-view consistency. Subsequent works have been devoted to engineering the proxy loss, such as robustly handling outliers in the loss [53], incorporating optical flow estimation [180], and imposing constraints for 3D consistency [15]. [53] proposed a simple yet effective photometric proxy loss for training using monocular image sequences, which select pixel-wise minimum photometric loss across the photometric differences calculated from all reference views. This strategy proves to be effective in handling outliers contained in the self-supervision signals due to occlusion, and demonstrates impressive improvements compared to previous works. In addition to the significant advancements that self-supervised learning methods have achieved in the design of proxy loss, another series of works [188, 186, 184] focuses on investigating enhancements in network architecture to improve the performance of self-supervised learning in monocular depth and camera motion estimation. The recurrent structure [192] and transformer-based structure [186] have been proven to be effective in improving the performance of monocular depth estimation in the context of self-supervised learning.

On the other hand, research dedicated to enhancing the network architecture for pose estimation within the context of self-supervised learning remains relatively scarce. At present, state-of-the-art self-supervised methods equipped with well-designed proxy loss and network architecture can achieve competitive performance in the task of monocular depth estimation compared to early supervision methods on benchmarks. Moreover, they exhibit advantages in tasks that involve generalization to unseen datasets with limited labeled data. However, in tasks related to long-term pose estimation, self-supervised methods still demonstrate a significant accuracy gap when compared to traditional monocular approaches. Another challenge for self-supervised depth and camera motion estimation arises from the presence of moving objects in most real-world datasets. These moving objects disrupt the multi-view consistency that the proxy loss relies on in self-supervised depth and camera motion estimation. Consequently, applying these methods directly to large, real-world dynamic datasets becomes challenging. The treatment of moving objects continues to be an active research area [93].

1.2.3 Approach based on Neural Implicit Representation

More recently, the rapid development of neural implicit 3D representation and differentiable rendering has injected new vitality into image-based 3D reconstruction. In contrast to traditional approaches that use point clouds or meshes for discrete and explicit representation of 3D scenes, neural implicit methods leverage neural networks to model continuous spatial signals, such as the structure, color, and material of 3D scenes. Seminal works [23, 113, 120] in neural implicit 3D representation typically use a single multi-layer perceptron (MLP) network for the scene representation, which takes 3D spatial coordinates as input and outputs the spatial attributes at corresponding coordinates, such as its occupancy or signed distance. This form of implicit representation is compact and memory-efficient, with these methods demonstrating impressive object-level reconstruction quality. Some follow-up works [25, 123] extend the single MLP representation to the combination of an MLP decoder and multi-level voxel grids of low-dimensional features, enabling the application to large-scale scenes at the cost of larger memory usage. However, the training of the early works still relied on 3D supervision.

The emerging differentiable rendering [82] techniques have bridged the gap between neural implicit 3D representation and 2D image rendering, enabling the implicit reconstruction of 3D scenes from multi-view images. Representatively, neural radiance fields (NeRFs) [114] model a 3D scene as a field function that

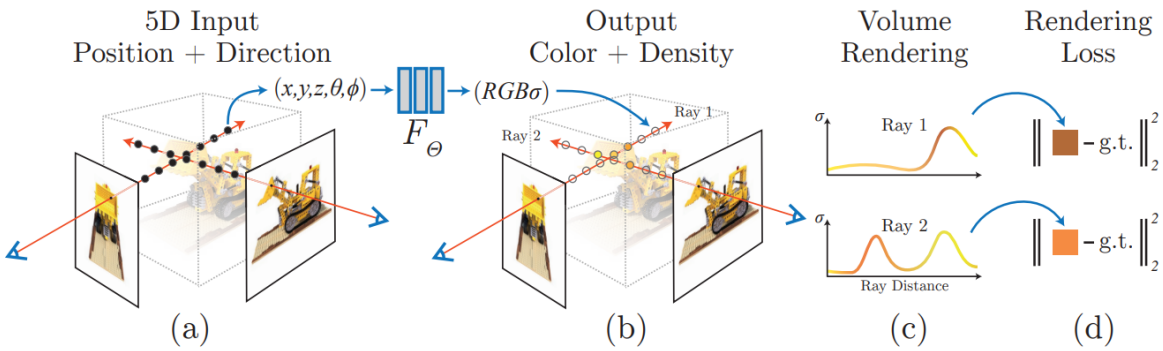


Fig. 1.4 **The overview of neural radiance field scene representation and its training [114].** The figure is adapted from [114].

takes a 5D coordinate (3D position plus 2D viewing direction) as input, predicting density and color as output. Utilizing differentiable volume rendering, it produces rendered images of the 3D scene from arbitrary viewpoints. The parameters of the MLPs representing this 5D field function are updated by computing the loss between rendered images and ground truth images of known viewpoints, until convergence. Several subsequent works [162, 45] combine surface-based implicit scene representation and volume rendering by transforming signed distance into volume density through a conversion function, which in turn improves the scene geometry extracted from neural radiance fields.

The aforementioned work focuses primarily on learning implicit scene representations from image collections and does not emphasize camera pose estimation. Typically, these methods preprocess image sequences using traditional Structure-from-Motion (SfM) techniques to obtain camera poses, which remain unchanged during the learning process of implicit scene representations. Recognizing the crucial impact of camera pose accuracy on the scene reconstruction results of neural radiance fields, some efforts [99, 169] concentrate on simultaneously optimizing noisy camera pose inputs and learning the neural implicit representation. The success of neural implicit functions in 3D scene representation has also spurred developments in the field of vision-based Simultaneous Localization and Mapping, wherein traditional 3D scene representation is replaced by neural implicit representation. iMap [142] introduces a real-time dense SLAM system that utilizes a single MLP network to compactly represent the entire scene In an effort to enhance the scalability of implicit scene representation, NICE-SLAM [193] employs a multi-level feature grid combined with a pre-trained decoder, enabling the construction of detailed geometry and textures for larger scenes. Both iMap and NICE-SLAM still depend on RGB-D inputs

to provide geometry cues for the convergence of neural implicit representation. NICER-SLAM [194], an extension of NICE-SLAM, relies solely on RGB inputs and leverages pre-trained monocular depth and normal estimation network for geometry cues, which also achieves real-time and competitive performance. [94] introduces a monocular RGB SLAM system that depends on multi-view consistency for geometry cues, removing the necessity for additional depth information or pre-trained monocular geometric models. Currently, neural implicit representation has achieved impressive results in expressing both indoor and outdoor scenes, while its application in more specialized domains, such as the medical field, is still under exploration.

1.3 Challenges and Thesis Contributions

In the previous chapter, we introduced the progress in recovering camera motion and scene structure from monocular image sequences through three avenues: 1) traditional geometry-based approaches, 2) learning-based data-driven approaches, and 3) approaches based on neural implicit representation. These three routes do not represent a simple progressive relationship but rather exhibit distinct advantages and disadvantages depending on application scenarios and requirements. For traditional geometry-based approaches, their strength lies in stability, as their performance remains relatively consistent in simple scenes across different datasets. However, their drawback is a lack of robustness, often leading to failures in challenging scenarios, such as texture-less scenes or motion blur contained in the image data. Learning-based data-driven methods typically exhibit strong robustness, but their performance is dependent upon the scale and quality of the training data. The emerging approaches based on neural implicit representation provide a more compact and expressive scene representation compared to the first two avenues while ensuring their robustness in challenging scenarios remains a challenge.

In this thesis, We have proposed effective methods to address the shortcomings and challenges within each of the aforementioned avenues. Fig. 1.5 illustrates the organization of this thesis.

In Chapter 2, we initially tackle the challenges related to the robustness of traditional geometry-based methods in challenging scenarios. We present a robust and efficient Structure-from-Motion pipeline for accurate 3D reconstruction under challenging environments by leveraging the camera pose information from visual-inertial odometry. Specifically, we propose a geometric verification method to

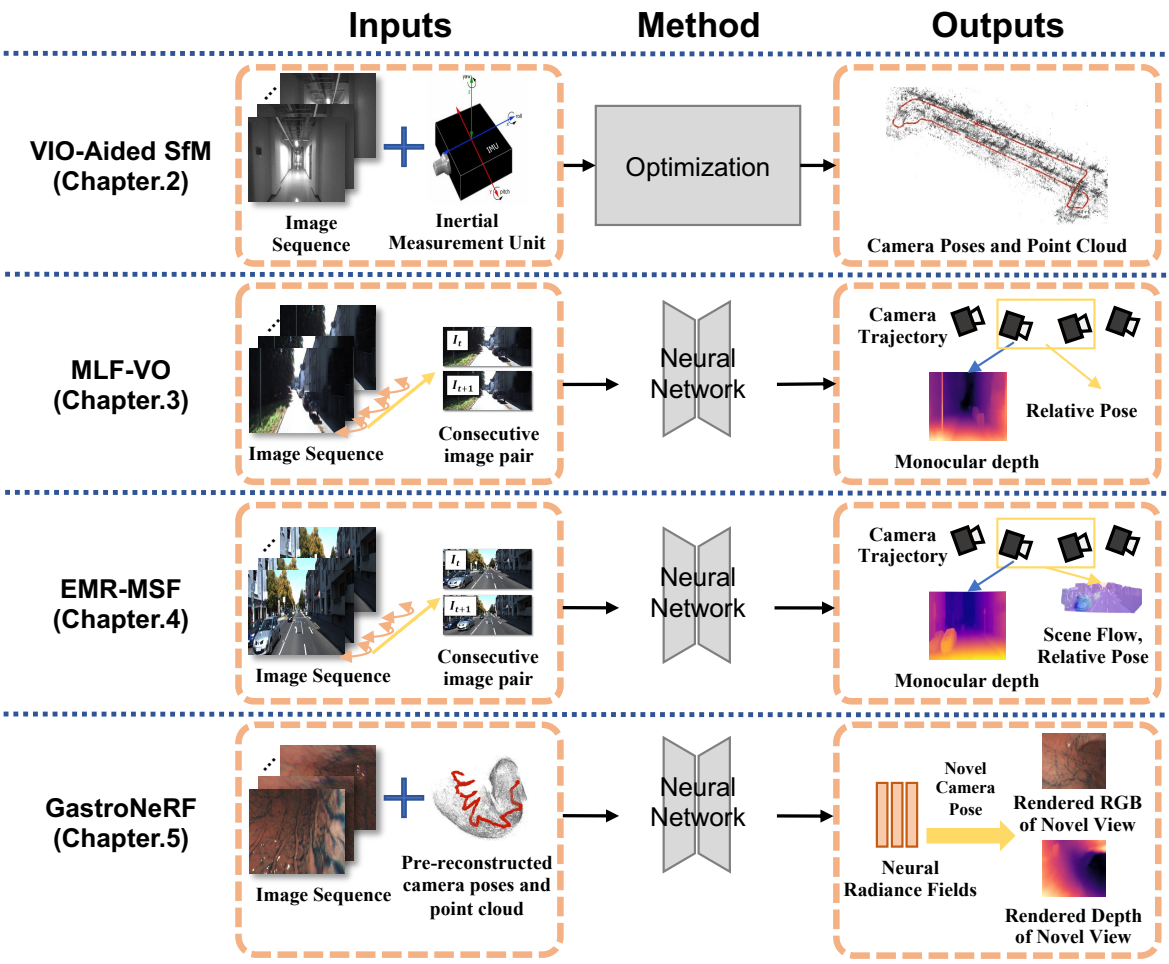


Fig. 1.5 **The organization of this thesis.**

filter out mismatches by considering the prior geometric configuration of candidate image pairs. Furthermore, we introduce an efficient and scalable reconstruction approach that relies on batched image registration and robust bundle adjustment, both leveraging the reliable local odometry estimation.

In Chapter 3, we delve into the impact of network architecture and feature fusion strategies on the performance of existing self-supervised learning approaches for camera motion and scene recovery. We propose a new framework called **MLF-VO** for self-supervised learning of depth and ego-motion estimation, which performs ego-motion estimation by leveraging RGB and inferred depth information in a Multi-Layer Fusion manner. Detailed studies on the design choices of leveraging inferred depth information and fusion strategies have also been carried out, which clearly demonstrate the advantages of our proposed framework.

In Chapter 4, to leverage real-world, unlabeled datasets containing numerous dynamic objects, we integrate 3D scene flow estimation, modeling the 3D motion of both static and dynamic objects, into the self-supervised framework for camera motion and depth estimation. we propose a superior model named **EMR-MSF** by borrowing the advantages of network architecture design under the scope of supervised learning. We further impose explicit and robust geometric constraints with an elaborately constructed ego-motion aggregation module where a rigidity soft mask is proposed to filter out dynamic regions for stable ego-motion estimation using static regions. Moreover, we propose a motion consistency loss along with a mask regularization loss to fully exploit static regions. Several efficient training strategies are integrated including a gradient detachment technique and an enhanced view synthesis process for better performance.

In Chapter 5, we apply the emerging technique of neural radiance fields (NeRF) to the challenging medial domain and propose **GastroNeRF**, which leverages monocular gastroscopic data for rendering photo-realistic images from novel viewpoints within the patient's stomach. We incorporate novel geometry-based supervision from reconstructed point clouds into the training of NeRF, which introduces the improved geometry-based loss to both pre-captured observed views and generated unobserved views to address the performance degradation due to view sparsity in local regions of monocular gastroscopy.

In Chapter 6, we conclude the dissertation and briefly discuss future works.

Chapter 2

VIO-Aided Structure from Motion Under Challenging Environments

2.1 Introduction

3D reconstruction with accurate geometry is desired for many different applications, such as robot navigation and industrial inspection. Structure from Motion (SfM) is a common technique to achieve this goal, which aims to recover 3D geometry and camera poses from image collections of a target scene [43, 136]. Given well-conditioned image collections, SfM can achieve highly precise 3D reconstruction assured by accurate camera pose estimation using rich local feature correspondences [41, 103] and subsequent global bundle adjustment [153] that refines the camera poses and structures. However, these approaches are vulnerable to the degradation of visual information such as the absence of textures and lack of overlapping views, consequently failing to find a good initial pose, resulting in incomplete or broken 3D structure (cf. Fig. 2.1).

On the other hand, thanks to the progress of sensor technology, imaging devices equipped with other built-in sensors such as inertial measurement unit (IMU), become widely available [84, 116]. In the field of robotics perception, various visual-inertial odometry (VIO) algorithms [116, 78, 91, 127] have been proposed to provide an accurate local camera pose estimation by fusing IMU measurements to image information. Even when the images cannot provide information about camera motion, VIO can still estimate the camera motion properly in a short time solely dependent on IMU measurements [127]. Estimated camera poses, however, do not necessarily satisfy consistencies in the whole scene since online VIO systems rarely

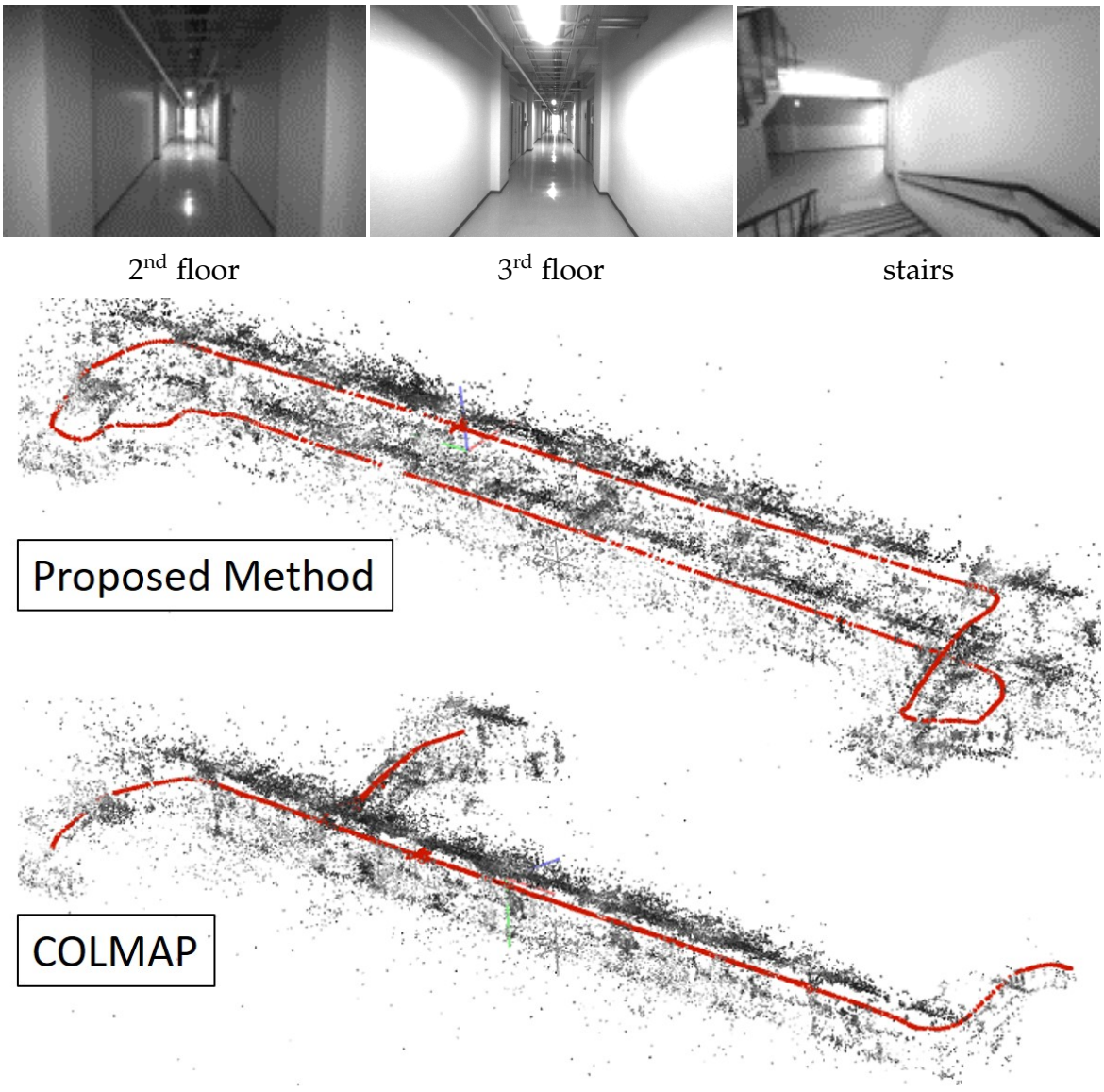


Fig. 2.1 **3D reconstruction results for a challenging indoor scene.** Gray dots are the reconstructed 3D scene points and red cones represent the estimated cameras relative to the model. COLMAP incorrectly merges two different floors due to their similar appearance. Also, the sequence appears weak visual connectivity (fewer feature matches) at the stairs part, which causes an unstable camera pose estimation. On the other hand, our proposed method provides an accurate reconstruction with aligned structures of different floors.

perform global bundle adjustment. Thus, due to the noises of IMU measurements, VIO often suffers from significant accumulated odometry errors. In addition, because of its purpose, VIO only produces a rather small 3D map for each frame than a globally consistent 3D structure obtained via SfM.

In this chapter, we aim to achieve a robust and accurate 3D reconstruction that can produce a globally consistent 3D model, even in visually severe situations, *e.g.*, highly texture-less indoor scenes, poor light conditions, and repetitive structures in industrial scenes. Assuming an input of sequential images and IMU measurements, we propose an SfM-based reconstruction pipeline that incorporates VIO estimation. Exploiting its robustness and local consistency, our system first estimates the camera odometry via a VIO algorithm and then integrates the relative camera poses into each step of the SfM pipeline. This allows us to robustly construct 3D scene representations even in visually severe situations. Furthermore, our proposed batch-wise image registration scheme with a new global bundle adjustment process, also aided by VIO estimation, ensures the global consistency of the obtained model at the marginal computational overhead. Fig. 2.1 provides a bird view of the 3D scene model obtained by our proposed method in a challenging scenario. Compared to an existing SfM-based reconstruction system [136], our method can produce an accurate 3D reconstruction, while dealing with repetitive scene natures and image sequences which has poor visual connectivity to each neighbor. Our contributions can be divided into three components:

- We propose a new geometric verification method that discards wrongly matched image pairs using the prior geometric configuration from VIO. This scheme is typically effective in the presence of dominant repetitive structures in the scene.

- Each image frame in the input is incrementally registered to the model by initializing its pose using VIO estimation. We then introduce a new cost function of bundle adjustment that refines the camera poses and 3D structures while balancing the vision-based and VIO-based penalties. Also, we effectively manage the computational cost for incrementally running the global bundle adjustment by designing the reconstruction pipeline in a batch-wise manner, while preserving the accuracy of the model.

- Finally, we evaluate the performance of the proposed pipeline using publicly available image (and IMU measurement) datasets which include various challenging situations such as weakly textured indoor scenes, industrial scenes

dominated by repetitive structures, and poor lighting conditions. Compared with SfM-based and VIO-based methods, the proposed method provides further accurate camera pose estimation, which results in a globally consistent 3D model.

2.2 Related Works

2.2.1 Structure from Motion

SfM has been widely used as a vision-based 3D reconstruction tool [175, 136], because of its robustness to various input scenarios [98, 43, 174] including unordered internet photos [138]. Several works achieved an accurate reconstruction for both camera motion and 3D structure, evolving each of the sub-systems in Structure-from-Motion (SfM) such as feature detection and matching [103, 36], camera pose initialization [90, 66, 136], multi-view triangulation [60, 61] and bundle adjustment [153, 175, 79]. On the other hand, Simultaneous Localization and Mapping (SLAM) systems have been developed as an alternative approach for vision-based 3D reconstruction, which is motivated to track the sequential image series input from a moving camera [29, 39, 118, 42, 40]. Ensuring its availability for real-time processing, SLAM usually employs on-the-fly consecutive camera poses estimation, and then refines them in post-processing, *e.g.*, by loop closure [145, 88, 87].

One common shortage of these vision-based 3D reconstruction methods is the lack of relevance in the absence of visual information. A potential approach to address the issue is to collect auxiliary camera pose information from other systems such as Global Positioning System (GPS) and Inertial Navigation System (INS), and utilize them for the initial structure and camera poses [71, 3]. These methods still rely on high-precision GPS/INS measurements for obtaining a good global initialization. Cui *et al*. try to eliminate the dependence using a track selection strategy and performing iterative triangulation and bundle adjustment [28].

2.2.2 Visual Inertial SLAM

In both computer vision and robotics perception, several VIO systems have been proposed to obtain more robust and accurate camera poses by fusing raw image information and IMU measurements in a single pose estimator [116, 78, 16, 49, 91, 127]. VIO systems usually seek the locally optimal cameras in a sliding-window fashion,

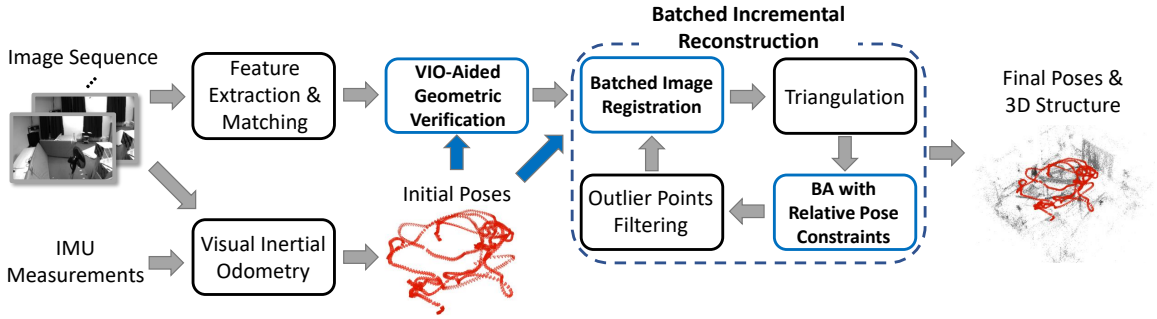


Fig. 2.2 **The overview of our proposed VIO-aided SfM system.** Our system takes sequential image collections and associated IMU measurements as inputs. We first obtain the initial camera poses of each image through a VIO system, and then incorporate them as prior information into the subsequent geometric verification and batched incremental reconstruction processes.

i.e., considering only recent measurements [37, 91, 127]. This configuration, however, loses global information and sometimes causes a long-term drifting in estimated poses, *i.e.*, cameras in the global coordinate system appear erroneous trajectories. Several works address the drift issue via loop closure [85, 110, 141] that attempts to detect a camera trajectory loop during the camera motion. A global pose graph optimization [141] typically follows the loop detection to alleviate the drift errors.

In this chapter, we aim to achieve high-quality reconstruction results in terms of both accuracy and robustness exploiting the camera poses obtained by a VIO system [127]. In contrast to previous methods [71, 3], the obtained camera pose information is utilized in a batched incremental manner, which alleviates the impact of accumulated odometry errors. In addition, we introduce a constrained bundle adjustment using the camera motion from VIO to achieve a globally consistent and more robust reconstruction result.

2.3 3D Reconstruction by VIO-Aided SfM

Fig. 2.2 illustrates our proposed pipeline for sparse 3D reconstruction from sequential image collections and associated IMU measurements. In what follows, we describe each part of our SfM-based reconstruction pipeline that incorporates the initial camera poses provided by a VIO system. First, we obtain camera poses of each image through an existing VIO system [127]. Though the original system offers absolute camera poses in the scene, we extract relative camera poses for utilization in the latter processes. Second, the camera poses are utilized as prior information for

geometric verification of image pairs (Sec. 2.3.1). Third, we register images into the global 3D model in a batch-wise incremental manner that iteratively expands the model using local geometries of a subset of images (Sec. 2.3.2). At the end of each batched process, both the obtained scene structure and registered camera poses are jointly optimized by also considering the VIO odometry to achieve both local and global consistency. The batch process is repeated until all images are registered.

2.3.1 VIO-Aided Geometric Verification

The accuracy of the 3D structure reconstructed by SfM highly depends on the accuracy of detected correspondences between images. Therefore, after getting tentative matches from local feature matching, SfM systems generally introduce an outlier rejection scheme such as RANSAC that fits a transformation model between image points computed from randomly sampled matches [41, 61]. However, an incorrect transformation between images can still be estimated when the wrong matches are dominant. This typically happens in weakly textured indoor scenes, including visually similar objects in different places, *e.g.*, corridors, standardized doors, and furniture. Wrongly matched images offer false connections between actually distant places and can result in a collapsed 3D model. We address such ambiguous matches by exploiting the robust camera pose estimation of the VIO system. In what follows, we describe our new image-level verification scheme that discards wrongly matched image pairs using their relative camera poses provided by VIO.

Given an image pair $(I_t, I_{t'})$, we assume the relative pose from I_t to $I_{t'}$, which consists of a rotation matrix $\hat{\mathbf{R}}$ and a translation vector $\hat{\mathbf{t}}$, is provided as the prior information from VIO. We introduce the pixel-based Average Epipolar Error (AEE) that evaluates the local feature correspondences between images:

$$\begin{aligned} AEE(I_t, I_{t'}) &= \frac{1}{N_f} \sum_{k=1}^{N_f} d(f_{k,I_{t'}}, \hat{\mathbf{F}} f_{k,I_t}), \\ \text{where } \hat{\mathbf{F}} &= \mathbf{K}^{-\mathrm{T}} \hat{\mathbf{t}}_{\times} \hat{\mathbf{R}} \mathbf{K}^{-1}, \end{aligned} \tag{2.1}$$

where N_f is the number of tentative feature matches, f_{k,I_t} and $f_{k,I_{t'}}$ are the corresponding feature points between the image pair, and $\mathbf{K}$ is the camera intrinsic parameters. $[.]_{\times}$ denotes the matrix representation of the cross product and $d(.)$ denotes the perpendicular Euclidean distance between the image point and line. Here $\hat{\mathbf{F}}$ is the fundamental matrix that projects the image point f_{k,I_t} using known properties of the

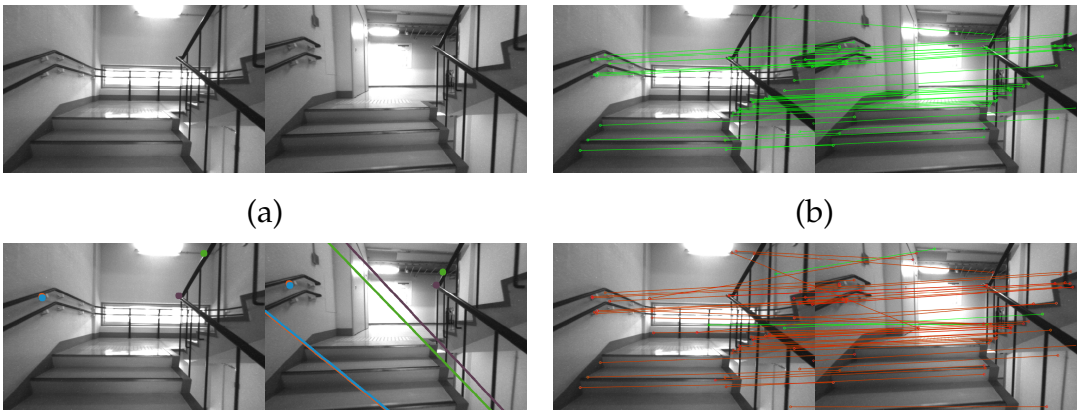
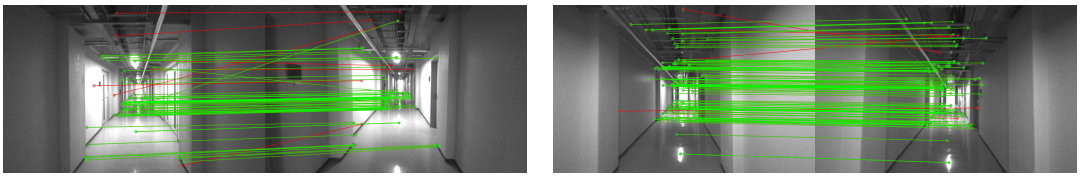


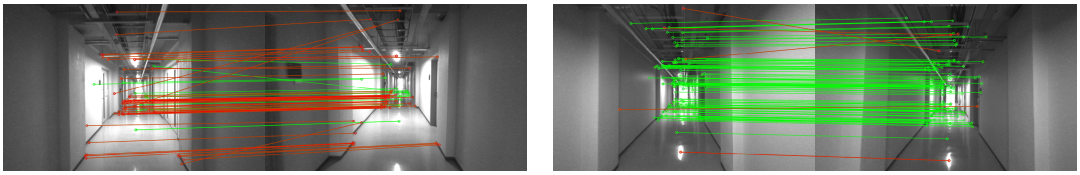
Fig. 2.3 **Illustration of our proposed image-level verification strategy.** (a) Example of a wrong image pair. (b) Due to visual ambiguity, incorrect feature correspondences are built on this image pair using tentative matches. (c) We draw several examples of incorrect correspondences and the corresponding epipolar lines, which are calculated based on VIO estimations in the right image. (d) The red and green lines indicate feature correspondences with large and small epipolar errors respectively.

image intrinsic and relative camera poses. Consequently, AEE evaluates the average of the distances between image point $f_{k,I_{t'}}$ and the epipolar line $\hat{\mathbf{F}} f_{k,I_t}$ projected by the prior $\hat{\mathbf{F}}$, which indicates the consistency between VIO-based geometric configuration and vision-based image relevance. We assume the image pair is wrongly matched if it has a larger AEE than the threshold T_{AEE}.

An example of the image-level verification strategy is illustrated in Fig. 2.3. The example of a wrong image pair is selected from the challenging sequence in Fig. 2.1. The local feature matching gives enormous incorrect feature correspondences due to the similar appearance of these two different places. We verify the feature correspondences using the prior geometric configuration of this image pair based on VIO estimations by drawing the corresponding epipolar line of each feature match. In Fig. 2.3 (c), we show several examples of the incorrect feature correspondences and their epipolar lines. The epipolar error of each match can be directly calculated by measuring the perpendicular Euclidean distance between the point and the epipolar line plotted on the right side. As a result, most detected feature correspondences in the image pair show great epipolar errors, as indicated in red lines in Fig. 2.3 (d), which give a large AEE (106.39 pixels for this image pair). Our verification scheme will discard such image pairs with larger AEE than the determined threshold T_{AEE}.



(a) RANSAC-based geometric verification



(b) Our VIO-aided geometric verification

Fig. 2.4 **Visualization of geometric verification under a challenging scene.** In the left column, we show the example of a wrongly matched image pair, which looks very similar but comes from different places (taken on different floors). For comparison, the right column shows the example of images taken from the same place after a loopback. For each pair, we draw lines of tentative matches by colors of green and red indicating inlier and outlier matches detected by each geometric verification.

Offering the new image-level verification strategy, we design a two-step geometric verification scheme. We first build candidates of image pairs for feature matching. For each image, we choose N_1 neighboring images based on timestamp and N_2 visually similar images by image retrieval. The tentative correspondences between these pairs are obtained by standard feature matching scheme [103]. As the first stage of verification, we compute the AEE of each image pair and discard the pair if it has a larger error than the threshold. Next, we evaluate the tentative correspondences of the remaining candidates by a RANSAC-based outlier rejection scheme [41, 61] which estimates an epipolar transformation between images. The candidates are finally approved if a sufficient number of inlier matches exist. Note that since the raw VIO estimations are not perfect, it is not guaranteed that all wrong image pairs can be rejected by our proposed image-level verification strategy. For the challenging sequence reported in Fig. 2.1 with dominant repetitive structures, our proposed image-level verification method can reject most wrong image pairs ($\sim$ 95%), whose tentative matches significantly violate the prior geometric configuration based on VIO estimations.

At the end of this section, a qualitative comparison between RANSAC-based geometric verification method [41, 61] and our method is shown in Fig. 2.4. Standard RANSAC estimates an epipolar geometry between images using tentative matches,

thus producing a large number of false-positive inliers when matches are dominated by wrong results (e.g. the detected matches in the example of a wrongly matched image pair in the left column of Fig. 2.4, which are visually similar but are taken on two different floors). In contrast, given fair VIO estimations and predefined threshold T_{AEE}, our verification scheme builds an epipolar constraint and evaluates the AEE of each image pair (e.g., 69.35 pixels for the example of a wrong image pair and 12.01 pixels for the example of a correct image pair after a loopback) to discard such an image pair that has a large AEE.

2.3.2 Batched Incremental Reconstruction

We are next to register all images into the global coordinate system and build 3D scene points for recovering the target scene. Given the initial camera poses obtained via VIO, a simple strategy to achieve this goal is to register all images at once using absolute camera poses from VIO and triangulate 3D points using feature correspondences [71, 3]. However, VIO estimations usually suffer from significant odometry drifts which turn into inaccurate absolute camera poses. Instead, we utilize the relative camera poses between consecutive images and incrementally build up the model. We also propose to perform the image registration in a batched manner, which suppresses the additional computation costs in a marginal range.

Batched Image Registration

We divide the sequential images into several consecutive k-frame batches in time order. The batch size k affects the final accuracy and computation time of our approach, which will be further discussed in Sec. 2.4.2. The initial camera pose set of batch i is denoted as $\mathbf{P}_i = \{\mathbf{p}\}$. In the i-th iteration, we register the i-th batch of images by computing a rigid transformation $\mathbf{T}_i$ to align this batch of images to the current model. We directly compute $\mathbf{T}_i$ as the relative camera pose between the first image in $\mathbf{P}_i$ to the last image in $\mathbf{P}_{i-1}$, which proves to be fast and effective.

Triangulation of 3D Scene Point

After a new batch of cameras is added to the existing reconstruction, we perform triangulation for the new tracks with equal or more than 2 cameras. We also re-triangulate the previous tracks which are extended by newly added cameras. We adopt a RANSAC-based triangulation method proposed in [136]. For each iteration

of processing a track, we randomly select two visible cameras in the track, and then check the view angle between the two cameras. We considered the selected two cameras to be well-conditioned if the angle is greater than T_a degrees, and use the DLT [61] method to determine the 3D scene point. After we get a 3D scene point triangulation, both the number of its consistent measurements from other cameras in the track and the corresponding view cheirality are checked. Note that all the cheirality [172] of cameras in a track should be positive and the measurements in a track are considered consistent with the current 3D scene point estimation if the corresponding re-projection error is less than T_r pixels. For each track, we find the best estimation of the 3D scene point that has the largest number of consistent measurements.

Bundle Adjustment with Relative Pose Constraints

After each batched image registration and triangulation, we refine the cameras and 3D scene points to guarantee the global consistency of the reconstruction. One general scheme for this purpose is global bundle adjustment [153] that minimizes the reprojection errors of the 3D scene points concerning the estimated cameras and their observed feature points. To handle the lack of visual constraints between consecutive cameras due to poor textures or motion blur, we propose a new objective for bundle adjustment that introduces additional relative pose constraints E_r based on the prior knowledge of the relative camera poses:

$$\begin{aligned}\underset{\mathbf{p}_i,\mathbf{X}_j,\mathbf{K}}{\arg\min} \sum_i \sum_j E_v(\mathbf{p}_i, \mathbf{X}_j, \mathbf{K}) + \\ \sum_i w_{i,i+1} E_r(\mathbf{p}_i, \mathbf{p}_{i+1} | \hat{\mathbf{p}}_{i,i+1}),\end{aligned} \tag{2.2}$$

where $E_v(\mathbf{p}_i, \mathbf{X}_j, \mathbf{K})$ is the standard bundle adjustment objective and is formulated as:

$$E_v(\mathbf{p}_i, \mathbf{X}_j, \mathbf{K}) = \rho\left(\left\|\mathbf{z}_{ij} - \pi\left(\mathbf{p}_i, \mathbf{X}_j, \mathbf{K}\right)\right\|^2\right), \tag{2.3}$$

where $\mathbf{K}$ is the camera intrinsic, $\mathbf{p}_i = [\mathbf{r}_i^{\mathrm{T}}, \mathbf{t}_i{}^{\mathrm{T}}]^{\mathrm{T}}$ is the 6D absolute pose vector of the i-th camera where $\mathbf{r}_i^{\mathrm{T}}$ is the axis-angle representation of the rotation, and $\mathbf{t}_i{}^{\mathrm{T}}$ is the corresponding translation vector. $\mathbf{z}_{i,j}$ is the feature point corresponding to the 3D point $\mathbf{X}_j$ observed by the i-th camera, and $\pi(.)$ denotes the projection function which projects scene points to the image plane. ρ is the robust function, *e.g.*, Cauchy function. This objective, purely based on visual information, sometimes gives

unstable results due to the weak visual connectivity between the scene points and images, *i.e.*, lack of image features. We thus introduce a relative pose constraint term E_r that penalizes the residuals of the camera motion concerning the prior knowledge of the relative poses:

$$E_r(\mathbf{p}_i, \mathbf{p}_{i+1}|\hat{\mathbf{p}}_{i,i+1}) = \left\|\mathbf{p}_{i,i+1} - \hat{\mathbf{p}}_{i,i+1}\right\|^2, \tag{2.4}$$

and $w_{i,i+1}$ is formulated by

$$w_{i,i+1} = \alpha e^{-\beta c_{i,i+1}}. \tag{2.5}$$

$\mathbf{p}_{i,i+1}$ is the relative pose between i-th and $(i+1)$-th cameras computed by $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$, and $\hat{\mathbf{p}}_{i,i+1}$ is the prior of the relative pose obtained from VIO. The term E_r penalizes the deviation between the estimated relative pose and its observation of the VIO. Since the relative camera poses obtained from VIO are possibly noisy, we hope to rely less on them when there is sufficient visual connectivity between the consecutive cameras and only rely heavily on them when the visual connectivity is weak, or in the absence of visual connectivity. Thus, we design a self-adaptive weighting factor $w_{i,i+1}$, where $c_{i,i+1}$ is the number of verified feature correspondences between the i-th and $(i+1)$-th camera, to balance the visual constraint and the relative pose constraint from VIO. α and β are two hyperparameters and their values are determined empirically. We use the Ceres Solver [1] for solving this nonlinear problem. After the bundle adjustment, the 3D scene points with large projection errors or small triangulation angles are filtered out for further robustness.

2.4 Experiments

2.4.1 Datasets and Implementation Details

Datasets

We collect several challenging sequences under different environments from two publicly available datasets [20, 81]. The EuRoC Dataset [20] contains sequences of images at 20 Hz and IMU measurements at 200 Hz captured mainly in indoor scenes. It also provides a ground-truth camera pose of each image which is obtained via VICON and Leica MS50. Each sequence is labeled as *easy*, *medium*, or *difficult* according to the illumination and camera motion. We use 7 sequences labeled with *medium* and *difficult* for evaluation, which capture separated scenes shown in Fig. 2.5. The OIVIO Dataset [81] consists of 36 sequences of images at 30 Hz and IMUs at

200 Hz recorded in weakly lighted environments. We select two sequences named "TUNNEL HANDHELD 3" and "MINE HANDHELD 1" (denoted by "Tunnel" and "Mine") for our experiments. In addition, we record our original sequence of walking in a challenging corridor and stair scene (named "Corridor and Stairs"). Our dataset contains 2677 frames of images at 30 Hz and IMU measurements at 200 Hz, captured by DUO MLX. It includes especially challenging cases that show significant degradations of visual information caused by texture-less areas, strong reflection, and motion blur. We arrange all image sequences by removing the static frames at the beginning and end and downsampling them to 10 Hz.

Implementation Details

For obtaining the initial camera poses of our SfM system, we choose VINS-Mono [127], one of the state-of-the-art VIO systems. The VIO-aided geometric verification is implemented with OpenCV [17]. The batched incremental reconstruction part of our system is extended from COLMAP [136], another state-of-the-art SfM pipeline. Based on the image frame rate of the sequence, which is 10Hz in our case, the parameter N_1 in Sec. 2.3.1 is set to 40 for matching images within two seconds before and after the target image. The parameter N_2 is set to 30, which is slightly smaller than N_1. The batch size k in Sec. 2.3.2 is set to 50 and its probe tuning experiment is reported in Sec. 2.4.2. The error thresholds T_a and T_r for triangulation are set to 3 degrees and 8 pixels respectively. The hyperparameters α and β in Eqn. 2.5 are set to 1e3 and 0.003 empirically. All experiments are conducted on a desktop with an Intel Core i7-6700 CPU and a Geforce GTX 980Ti GPU.

2.4.2 Quantitative Evaluation for Reconstructed Odometry

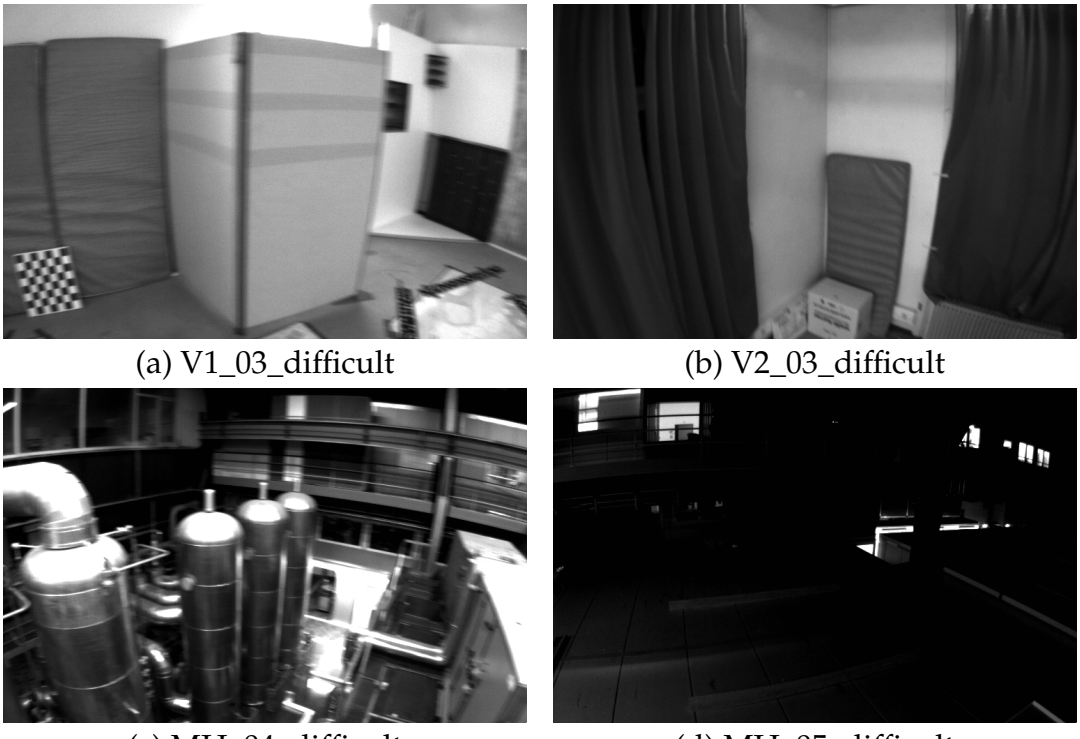
We first evaluate the accuracy of the reconstructed cameras by comparing their positions to the ground-truth locations. We compare our approach with other visual and visual-inertial methods including COLMAP [136], ORB-SLAM2 [118], DSO [40], OKVIS [91] and VINS-Mono [127] using implementations provided by authors on the 7 sequences from the EuRoC dataset where the camera location ground truth is available. Various challenging conditions including motion blur, texture-less area, and bad illumination are observed in these sequences, as shown in Fig. 2.5. Before

[1]We do not report results of OKVIS on V2_03_difficult because, despite our best efforts, it fails to estimate a reasonable trajectory for the sequence.

Table 2.1 **Performance of reconstructed trajectory.** For each comparison in the column, we report the root mean square error (RMSE) and the median error (ME) of the reconstructed camera positions in meters[1]. The best results are presented in bold and the second best are in blue. We additionally report the number of reconstructed cameras where the method does not provide a full trajectory for the input sequences.

Sequence		Visual methods						Visual-inertial methods					
		COLMAP[136]		ORB-SLAM2[118]		DSO[40]		OKVIS[91]		VINS-Mono[127]		Ours	
Name	Frames	RMSE	ME	RMSE	ME	RMSE	ME	RMSE	ME	RMSE	ME	RMSE	ME
V1_02_medium	756	0.043	0.040	0.064	0.063	0.598	0.213	0.067	0.062	0.060	0.057	0.022	0.019
V1_03_difficult	916	0.054	0.051	0.531	0.235	0.925	0.935	0.105	0.089	0.173	0.131	0.043	0.032
V2_02_medium	1025	0.029	0.032	0.056	0.056	0.092	0.080	0.081	0.058	0.124	0.103	0.014	0.012
V2_03_difficult	1028	0.041 (1014)	0.036 (1014)	0.079 (900)	0.073 (900)	1.386	1.008	-	-	0.191	0.153	0.029	0.021
MH_03_medium	968	0.040	0.034	0.038	0.032	0.172	0.135	0.146	0.143	0.080	0.067	0.035	0.029
MH_04_difficult	681	0.095	0.078	0.059	0.049	0.172	0.171	0.138	0.131	0.124	0.123	0.092	0.077
MH_05_difficult	701	0.084	0.064	0.068	0.055	0.102	0.093	0.261	0.227	0.133	0.110	0.083	0.072

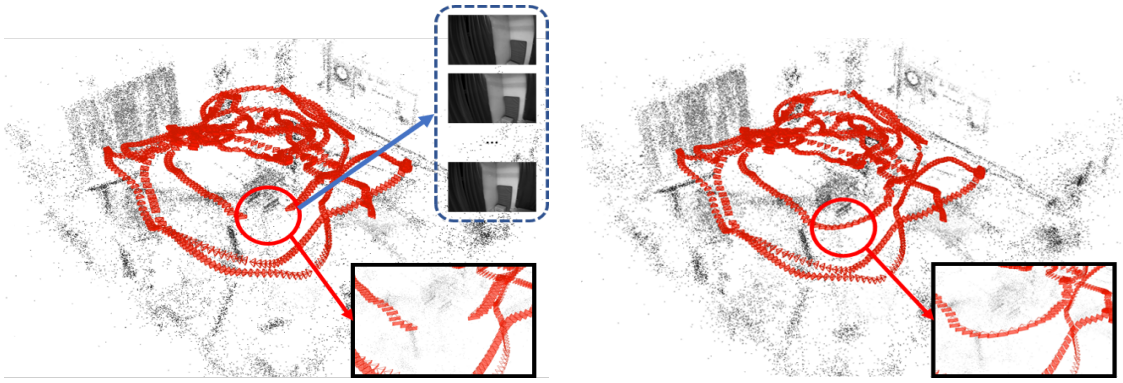
evaluation, we compute a similarity transformation [154, 56] to align the estimated trajectory to the ground-truth trajectory.



(a) V1_03_difficult (b) V2_03_difficult

(c) MH_04_difficult (d) MH_05_difficult

Fig. 2.5 Examples of images for the challenging environments in the EuRoC Dataset [20].



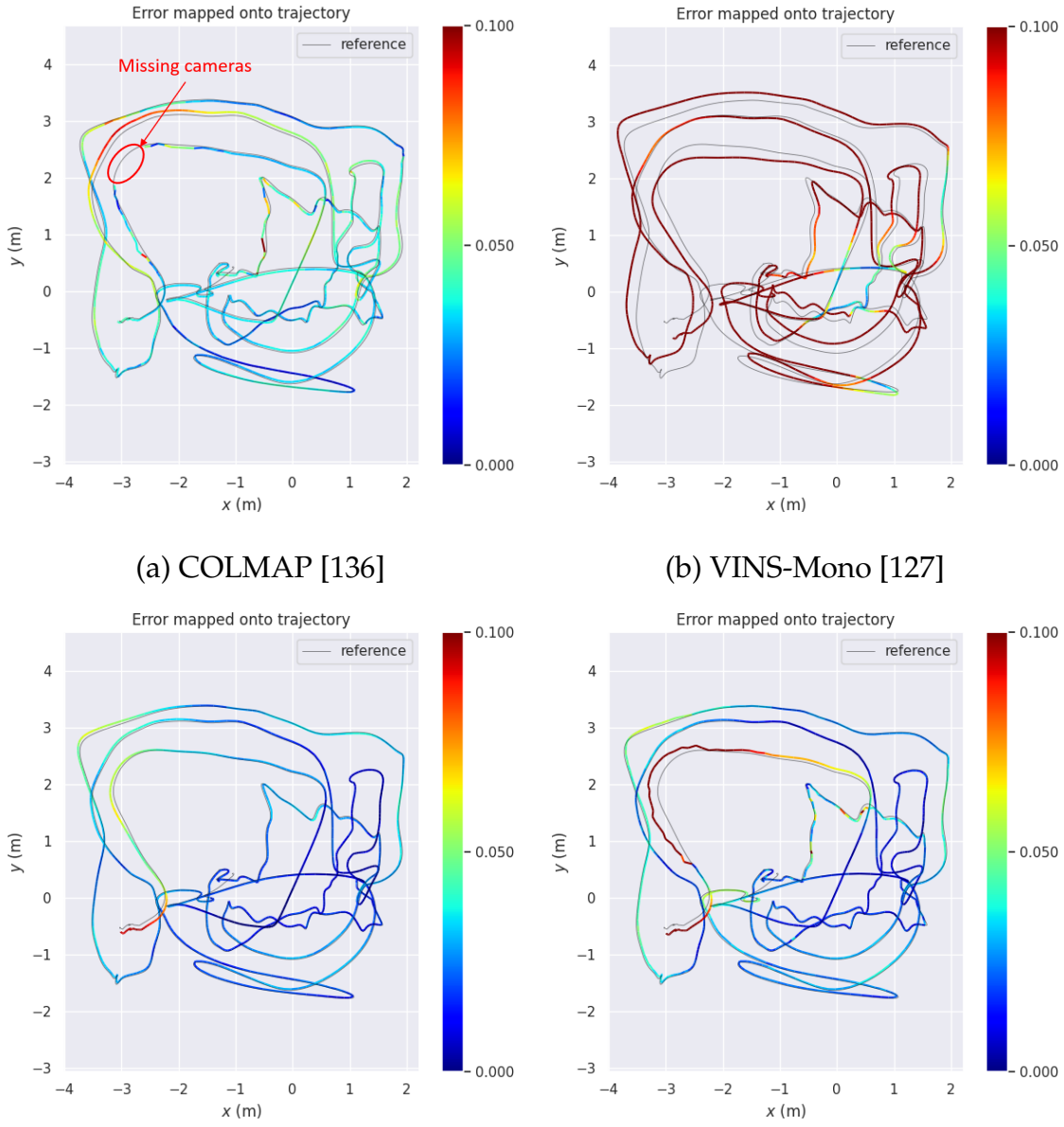
(a) (b)

Fig. 2.6 **Reconstruction results on the sequence V2_03_difficult.** Gray dots are the reconstructed 3D scene points and red cones denote the estimated cameras in the scene. (**a**) Reconstruction generated by COLMAP [136]. The upper right shows the images missed in the reconstruction. (**b**) Reconstruction generated by our approach.

Tab. 2.1 reports the quantitative comparisons of reconstructed trajectories. Our approach generally produces the smallest errors compared to ground truth in terms of both root mean square position error and median position error. Fig. 2.6 also shows the estimated trajectories and 3D points for a typical sequence on which COLMAP fails to estimate the camera poses due to strong motion blur and texture-less areas in the images (as shown in Fig. 2.5 (b)). On the other hand, our approach succeeds in registering all images based on VIO camera pose initialization. Also notice that our method shows a remarkable boost in accuracy compared to VINS-Mono, which confirms that our batched image registration and bundle adjustment with relative pose constraints can effectively deal with accumulated drifts of the initial camera poses and provide a high-precision 3D model.

Evaluation of Relative Pose Constraint

We evaluate the effectiveness of our proposed relative pose constraint and self-adaptive weighting factor in this section. The visualization results in Fig. 2.7 correspond to the qualitative evaluations on the sequence V2_03_difficult reported in Tab. 2.1. We highlight the missing cameras in the camera trajectory obtained by COLMAP [136] in Fig. 2.7 (a). Our method can generate a high-precision and complete camera trajectory (Fig 2.7 (c)) from the inaccurate prior camera poses obtained by the VIO method (Fig. 2.7 (b)). The self-adaptive weighting factor ensures that the visual constraints will be dominant during the optimization process



(a) COLMAP [136]

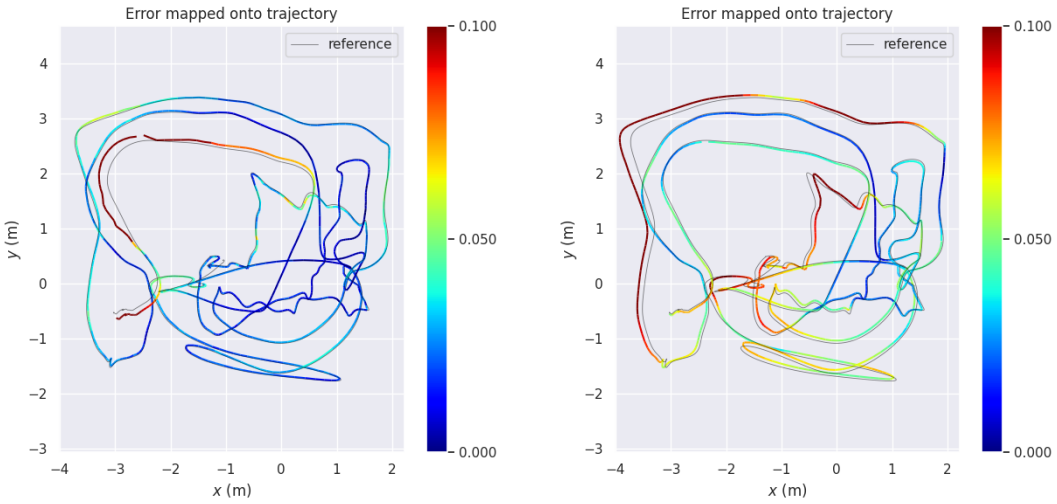
(b) VINS-Mono [127]

(c) Proposed method

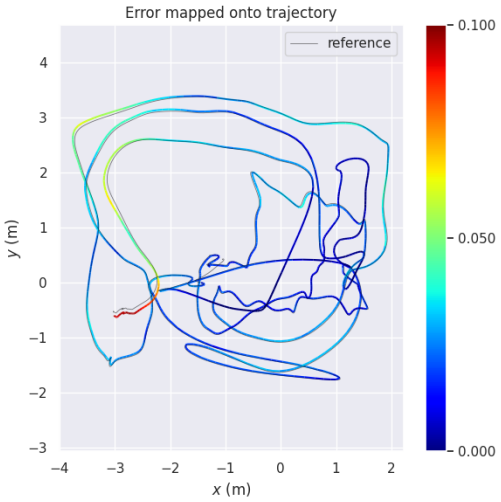
(d) Proposed method w/o RPC

Fig. 2.7 **Recovered camera trajectories on V2_03_difficult sequence from EuRoC Dataset [20].** (a) COLMAP [136], RMSE=0.041. (b) VINS-Mono [127], RMSE=0.191. (c) Proposed method, RMSE=0.029. (d) Proposed method without using relative pose constraints (RPC), RMSE=0.039.

when strong visual connectivity is detected, which prevents inaccurate relative pose constraints from dragging the solution either to a bad minimum or decreasing the accuracy of the solution. In Fig. 2.7 (d), we illustrate the recovered camera trajectory by our method without using relative pose constraints. Suffering from the



(a) $\alpha = 100$, RMSE=0.038 (b) $\alpha = 10000$, RMSE=0.062



(c) $\alpha = 1000$, RMSE=0.029

Fig. 2.8 **Recovered camera trajectories using different values of α.**

lack of visual constraints in texture-less areas, the recovered poses of cameras with weak visual connectivity to their neighboring cameras show large errors after the optimization. We also illustrate the impact of choosing different parameter values of α inEqn. 2.5 in Fig. 2.8. The best parameter value of α can be found by searching in a reasonable range, which is adopted during our experiments and determined to 1e3.

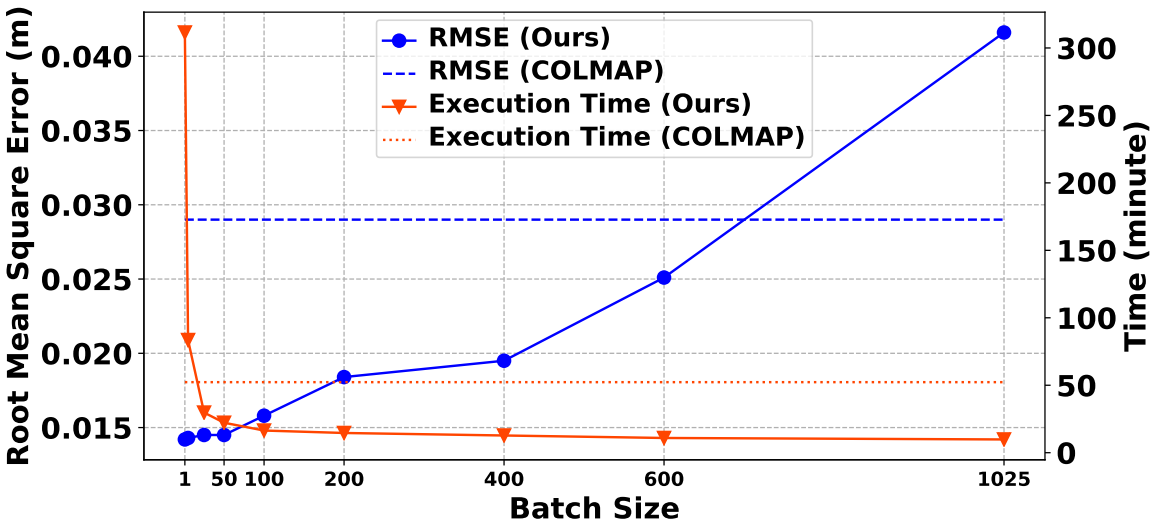
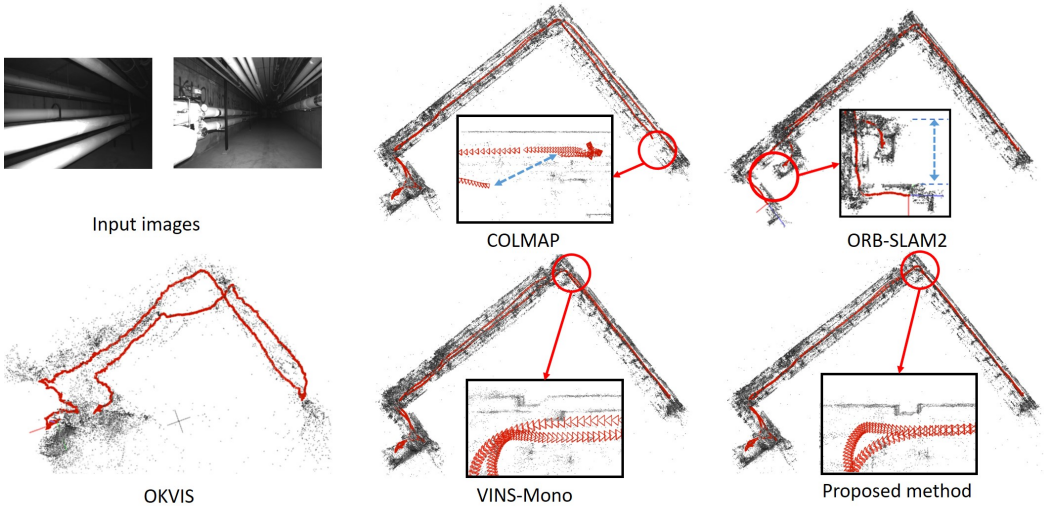


Fig. 2.9 **Trajectory RMSE and execution time using different batch sizes.** The trajectory RMSE and the execution time of COLMAP [136] are also shown by dotted lines for comparison.

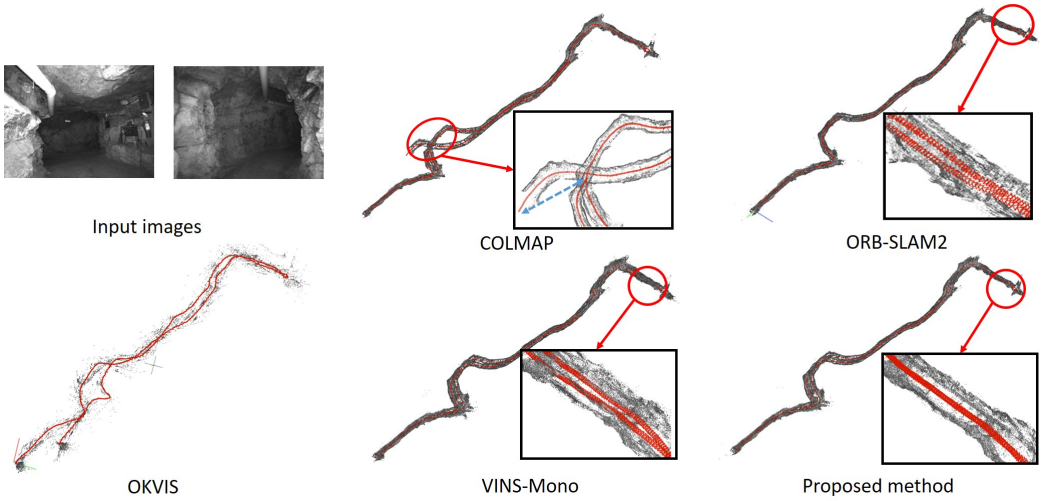
Influence of Batch Size

Next, we conduct experiments about the influence of batch size k in our method. The experiments are carried out on the V2_02_medium sequence from the EuRoC dataset, which contains 1025 images. We compare the recovered camera trajectories using different batch sizes. The relationship between the batch size and root mean square errors of the estimated camera trajectory is shown in Fig. 2.9. We also compare the execution time using different batch sizes. In general, a smaller batch size can bring a higher accuracy, but at the cost of computation time since the number of performing the global bundle adjustment increases. We find that when the batch size is smaller than a certain value, the accuracy does not change significantly. On the other hand, the execution time increases tremendously when the batch size is extremely small. This can be attributed to the computational costs for global bundle adjustment which runs in each batched reconstruction process. A smaller batch size results in more iterations and consequently produces further complexities in later iterations which refine a large number of registered cameras and points. Considering the trade-off between accuracy and efficiency, we finally selected 50 as the bath size in our experiments.

2.4.3 Qualitative Evaluation for Reconstructed 3D Models



(a) Tunnel sequence



(b) Mine sequence

Fig. 2.10 **Reconstruction results for two challenging environments: Tunnel and Mine from OIVIO dataset [81].** We compare our proposed method with COLMAP [136], ORB-SLAM2 [118], OKVIS [91] and VINS-Mono [127].

We finally provide some visual examples of the reconstructed cameras and 3D points for challenging image sequences from the OIVIO dataset [81]. Fig. 2.10 shows the reconstruction results obtained by our method, COLMAP [136], ORB-SLAM2 [118], OKVIS [91] and VINS-Mono [127]. Since OKVIS [91] and VINS-Mono [127] do not output 3D points, we additionally obtain 3D points for them by triangulating feature tracks approved by our geometric verification. We exclude

DSO [40] from this comparison because it fails to estimate complete trajectories for these sequences. For both of the scenes, COLMAP [136] fails to construct the continuous camera trajectories (pointed by blue arrows) due to the weak visual connectivity of the sequences. This especially happens in the Mine sequence, which has a longer trajectory, resulting in an inconsistent 3D model with duplicated structures (Fig. 2.10 (b)). On the other hand, we build a continuous camera trajectory by initializing camera poses using VIO estimation. Please also note that other visual SLAM or visual-inertial odometry methods, including ORB-SLAM2 [118], OKVIS [91] and VINS-Mono [127], suffer from inconsistent model reconstruction, *e.g.*, drifted and duplicated structures, because they estimate a locally consistent camera trajectory mainly based on feature matching only against to neighboring frames. In contrast, our system provides globally consistent 3D models by employing global bundle adjustment after registering each batched trajectory.

To sum up, we achieve the globally consistent 3D reconstruction that introduces the camera trajectory from VIO to an SfM pipeline. Experiments demonstrate that our method deals with challenging scenes providing less visual information while effectively utilizing the prior knowledge of camera poses from a locally consistent VIO estimation.

2.5 Conclusion

In this paper, we have proposed an SfM-based 3D reconstruction pipeline that effectively takes advantage of the camera pose information from a VIO. In contrast to existing SLAM-based visual-inertial reconstruction methods, we aim to construct a globally consistent and complete 3D model including camera poses and 3D points. Our method consists of a simple combination of VIO-aided camera pose initialization and SfM-based images-points reconstruction, but still gives a great margin in terms of the accuracy of the reconstructed model. Experiments on publicly available datasets demonstrate that our system can achieve an accurate and robust 3D reconstruction in challenging environments where the images provide less visual evidence for reconstruction. Also, we have shown that the computational time for the reconstruction can effectively be reduced by a batched incremental reconstruction process. One of the future works would be the detailed analysis for determining several parameters in our system that are highly relevant to its feasibility to noisy IMUs and consequent erroneous VIO estimation. Although we are currently tuning these parameters empirically, we believe that this work still suggests that the prior

knowledge of camera motion can benefit existing vision-based 3D reconstruction systems and implies there is significant room for improvement towards accurate 3D reconstruction.

Acknowledgement. This work is partly supported by JSPS KAKENHI Grant Number 17H00744.

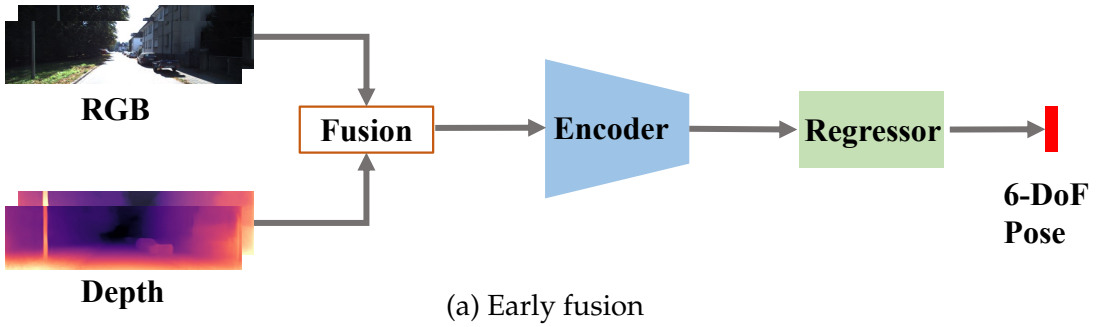
Chapter 3

Self-Supervised Ego-Motion Estimation Based on Multi-Layer Fusion of RGB and Inferred Depth

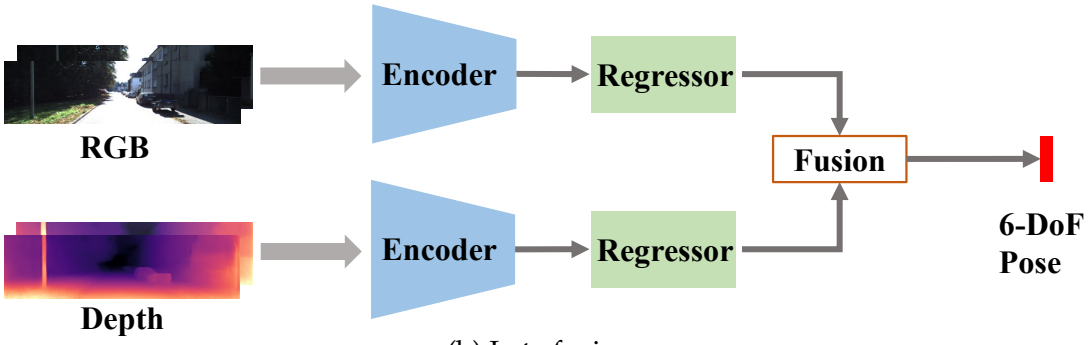
3.1 Introduction

Structure from Motion (SfM) and Simultaneous Localization and Mapping (SLAM) are popular and promising techniques in computer vision. One key component in them is to get an accurate ego-motion between two consecutive frames, which is often carried as camera pose estimation using RGB images. On top of the great success of classical methods based on 3D geometry and camera models, learning-based pose estimation methods [155, 190] have recently got increasing research interests for their good fits to training data and feasibility in typical severe situations, *e.g.*, poor lighting [92]. Most of these works treat the pose estimation problem as a regression from input color images, and design pose estimators based on the convolutional neural network (CNN) [155, 34, 190, 15, 53, 97, 5] or the recurrent neural network (RNN) [164, 163, 178, 197].

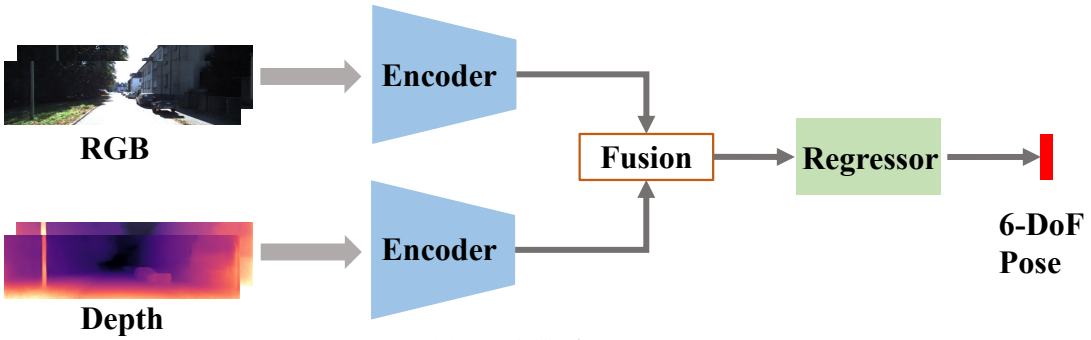
One common challenge of those learning-based pose estimators is the capability to generalize to different situations, *e.g.*, long-term scene changes including lighting changes [134]. Several works [22, 59, 170, 92] tackle this problem by adopting different modalities from additional sensors, such as LiDARs [92] and IMUs [22, 59, 170], along with RGB images. Alternatively, self-supervised training strategies for the



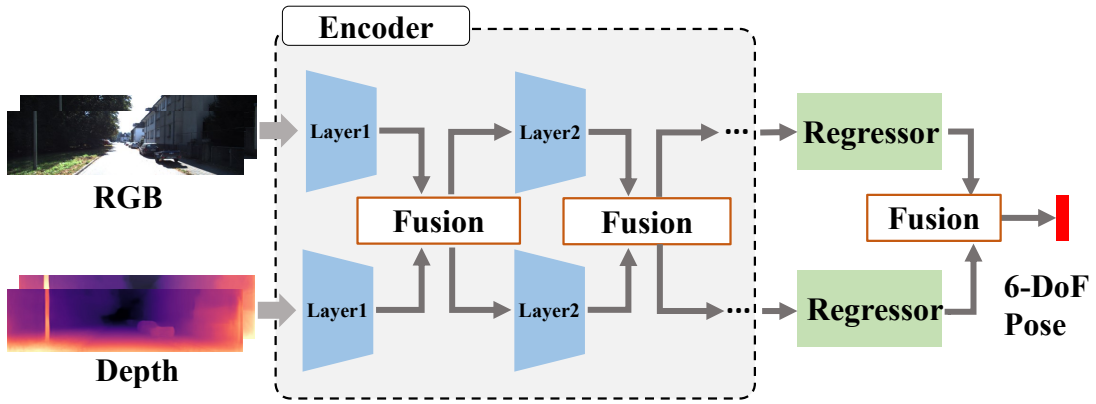
(a) Early fusion



(b) Late fusion



(c) Middle fusion



(d) Proposed multi-layer fusion (MLF-VO)

Fig. 3.1 Different fusion strategies depending on where to fuse RGB and depth modalities for pose estimation.

pose estimator have recently been studied [95, 183, 4] to utilize much more training data, which could consequently lead to a more generalized model. SfMLearner [190] simultaneously learns to estimate ego-motion and scene structure, *i.e.*, depth image, while measuring their consistency as supervision. Furthermore, recent works [163, 5, 197, 96] reported the fact that feeding a pose estimator not only with original images but also with depth images predicted from RGB images can improve the performance, as well as fusing different modalities from real sensors.

However, this approach raises one natural question: How can we effectively combine features from original color images and predicted *"pseudo"* depth images? Since the prediction via CNN involves multiple levels of features in intermediate layers, there are various design choices (cf. Fig. 3.1) to combine features from two streams, *e.g.*, simply feed a pose estimator with concatenated RGB-D images, or combine features after encoding two modalities independently? To the best of our knowledge, few works have conducted a thorough study to answer this question for the pose estimation task.

In this paper, **(1)** we design a baseline ego-motion estimation network for two paired images, which can be trained in a self-supervised manner. The network firstly predicts a depth image for each source image, then feeds all color and depth images to another pose estimator that predicts a relative pose between two cameras. **(2)** Second, we conduct a study on this pipeline to compare several design choices for fusing two modalities. **(3)** Based on the study, we propose an RGB-D-to-pose estimator that fuses two modalities in an incremental fashion. As shown in Fig. 3.1 (d), our pose estimator, MLF-VO (Visual Odometry using Multi-Layer Fusion), namely, consists of two streams for color and depth inputs and several fusion layers that combine intermediate features at multiple levels. Whereas the core architecture of the fusion layer originated from an existing work [166], we also propose a new regularization loss to effectively learn the model. **(4)** Finally, our overall pipeline achieves better pose accuracy than existing works on KITTI Odometry benchmark [47, 48], while requiring less computational time.

3.2 Related Work

3.2.1 Self-supervised learning of depth and ego-motion

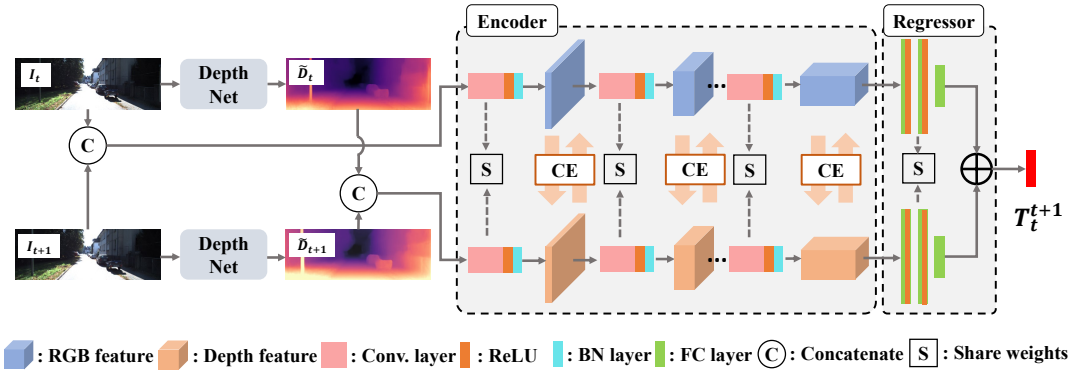
Self-supervised learning of depth and ego-motion from monocular video is originally offered in [190], which proposes to utilize two decoupled networks to estimate depth

and ego-motion independently and get supervisory signals by minimizing the photometric loss between the synthetic and original image [72]. Building upon this paradigm, recent works focusing on the improvement of self-supervised depth estimation [53] have achieved exciting progress, showing competitive performance compared to supervised methods.

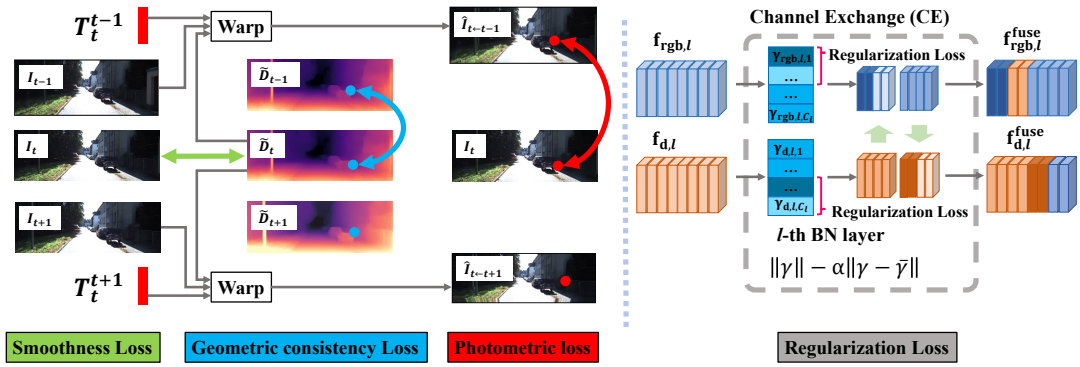
On the other hand, there are relatively few works focusing on the improvement of ego-motion estimation. [119] extends the standard ego-motion estimation network to incorporate feedback through iterative view synthesis. Instead of a direct regression model for ego-motion, [187] proposes to estimate an optical flow between two images and solves ego-motion as an optimal fundamental matrix. Besides these works which still perform ego-motion estimation purely in the RGB domain, some other works attempt to introduce data from other sensors, *e.g.*, LiDAR [92] and IMU [22, 170], as additional inputs to improve the ego-motion accuracy in the spirit of sensor fusion. Following the idea of performing motion estimation in a mixed domain, [5] further proposes a two-stream network that leverages the original RGB images and the internally inferred depth maps as inputs of the ego-motion network. Our method shares the idea with [5], but we further investigate how different fusion strategies affect the final performance and propose a new relative pose estimation network based on multi-layer fusion of RGB and inferred depth information.

3.2.2 Multi-modal fusion

Early studies of multi-modal fusion [139, 6, 19] categorize the fusion strategy into two broad types: early (raw-level) fusion and late (decision-level) fusion. Recent deep learning literature has also been studied mainly in either of these two branches [89, 128, 7]. Early fusion, which aggregates multiple modalities before making the decision, is often performed as a concatenation along the input channels [185, 181] or averaging [62], where the final decision is made by the subsequent single encoder and regressor. On the other hand, the late fusion strategy makes decisions from each modality solely. The final decision is obtained as an ensemble of multiple outputs, which can be carried out by averaging [5] or a learned meta-model [51, 166]. As an intersection of these two approaches, middle (feature-level) fusion has also been developed [128, 22, 170, 92]. This approach prepares a CNN or RNN-based encoder for each modality to obtain deeply encoded features. Features are then combined via a subsequent fusion layer, such as a self-attention mechanism [67, 156], and fed to the final regressor. We investigate all three fusion



(a) Forward pass of the proposed network (MLF-VO)



(b) Self-supervised training loss

Fig. 3.2 **The overview of the proposed MLF-VO.** (a) Our MLF-VO consists of two streams of CNN encoder and regressor fed with RGB and inferred depth images. After each BN layer of the encoder, intermediate features from two modalities are fused by CE operation. (b) Self-supervised training for our MLF-VO is based on the consistency of warped RGB (Photometric loss) and depth images (Geometric consistency loss). We additionally introduce a regularization loss that controls scaling factors of BN layers which are used in CE operation.

strategies in our ego-motion estimation pipeline and propose a new multi-layer fusion strategy.

3.3 Proposed Method

Fig. 3.2 illustrates the overview of our proposed ego-motion estimation framework (MLF-VO) and the training manner for the framework. In this section, we first introduce our baseline framework for ego-motion estimation composed of two independent CNN models for depth prediction and relative pose estimation (Sec. 3.3.1). Next, we detail our proposed relative pose estimator that fuses encoded features of

color and inferred depth images based on the channel exchange in multiple stages of the encoder (Sec. 3.3.2). Finally, Sec. 3.3.3 presents our self-supervised joint learning procedure for depth and relative pose estimators, along with a new regularization loss that encourages cross-modal information exchanges.

3.3.1 Baseline Ego-Motion Estimation Pipeline

Fig. 3.2 (a) illustrates our ego-motion estimation framework. Our framework consists of two independent prediction models: θ_{depth} for depth prediction and θ_{pose} for relative pose estimation. Assuming an image pair of consecutive video frames $\{I_t, I_{t+1}\}$, our pipeline firstly estimates depth images along with each input frame:

$$D_t = \theta_{\text{depth}}(I_t), \;\; D_{t+1} = \theta_{\text{depth}}(I_{t+1}), \tag{3.1}$$

where D_t is the pixel-aligned inverse depth map for I_t. We construct the depth prediction model in the same manner as in [53], which is based on the fully convolutional U-Net architecture obtaining depths at four scales. Next, the subsequent relative pose estimator θ_{pose} predicts a relative pose between consecutive frames as a final output of ego-motion:

$$T_t^{t+1} = \theta_{\text{pose}}(I_t \oplus I_{t+1}, D_t \oplus D_{t+1}), \tag{3.2}$$

where $\oplus$ denotes a concatenation along input channels. In contrast to several existing works regressing an ego-motion only from the original color image [190, 15, 53], we feed the model also with the inferred depth maps, which can provide additional information that is useful in some typical situations (cf. Fig. 3.3). Therefore, the fusion strategy for color and depth modalities plays an important role in estimating an accurate ego-motion.

3.3.2 Relative Pose Estimation Based on Multi-layer Fusion

As discussed in Sec. 3.2.2, there are various design choices to combine two different modalities into the final relative pose output. Based on a study of possible strategies applied in our baseline framework (Sec. 3.4.2), we finally decide to employ a *multi-layer fusion strategy* that fuses several features appearing in intermediate layers of the encoder (Fig. 3.1 (a)).

Our pose estimation model has a two-stream structure for encoding features from each RGB and inferred depth inputs. To effectively fuse multiple levels of features while remaining complementary characteristics of RGB and depth modalities, we exploit the Channel Exchange (CE) strategy [166] that swaps feature elements based on their importance. We employ the ResNet-18 architecture [63] as encoders for both streams, while sharing all weights except for Batch-Normalization (BN) layers so that we can ensure both features before BN are in the same latent space, *i.e.*, all feature elements can be replaced with the ones from the other modality. In every BN layer, CE evaluates the importance of each top (for color) or bottom (for depth) half of the feature channels as the BN scaling factor γ. We assume the channels with smaller γ than the threshold 0.02 become redundant to the final outputs and thus replace them with the same channels extracted from the other modality. After encoding each modality into the common feature space, we regress each feature into a 6-dimensional relative pose representation via a pose regressor composed of two fully-connected layers with ReLU activation and a final fully-connected layer. The final output of relative pose is obtained as a weighted sum of the outputs of each stream, where the weighting factor is jointly learned during the training.

Whereas the CE strategy originally has been tested on sensor fusion [166], we employ it for the fusion of color and inferred depth map. We also propose a new training loss for CE that prevents the model from reaching a singular solution, which will be introduced in the next subsection.

3.3.3 Self-Supervised Training

Self-supervised loss for depth and motion prediction

We train our MLF-VO in a self-supervised manner similar to [190, 53], which constructs a supervisory signal as the consistency of predicted motion and scene structure (Fig. 3.2 (b)). Assuming an input of target and source images $\{I_t, I_s\}$, the depth map D_t and the relative motion T_t^s, which are obtained from a forward pass, can present pixel-wise correspondences between two images, *i.e.*, we can generate a synthetic target view $\hat{I}_{t \leftarrow s}$ by projecting source image pixels onto the target image plane [72]. Therefore, the confidence of predictions can be evaluated by the photometric loss L_p that evaluates a photometric error [53] of the synthetic target view. Similarly, [15] introduces the geometric consistency loss L_{gc} that evaluates the consistency of predicted depth maps by projecting scene points defined from a depth

map instead of image pixels. We also adopt a smoothness loss for the depth map $L_s(I_t, D_t)$ in the same manner as in [52].

Regularization loss for CE

[166] originally imposes l_1 norm penalization on BN scaling factors to learn the redundant channels in CE. Although the l_1 norm loss is simple yet effective, there are still certain possibilities that the solution falls into a singular point, *i.e.*, all scaling factors converge to zeros thus all possible channels are exchanged. Therefore, we extend the original l_1 norm with the polarization regularizer proposed in [195] as:

$$L_r = \sum_m \sum_l \sum_i \|\gamma_{m,l,i}\| - \alpha\|\gamma_{m,l,i} - \hat{\gamma}_l\| \tag{3.3}$$

where $\hat{\gamma}_l$ denotes the mean of exchangeable scaling factors in the l-th BN layer. $\gamma_{m,l,i}$ denotes the scaling factor of the i-th exchangeable channel of the l-th BN layer of modality $m \in \{rgb, depth\}$.

End-to-end training

Motivated by per-pixel minimum loss strategy [53], which effectively tackles the problem of occluded regions in training, we prepare a set of training snippets $\{I_t, I_s, s \in \{t-1, t+1\}\}$ composed of one target image I_t and its consecutive source images $\{I_s\}$. The photometric loss $\bar{L_p}$ and geometric consistency loss $\bar{L_{gc}}$ are then re-formulated:

$$\begin{aligned} \bar{L_p}(I_t, I_s, T_t^s) &= \min_s L_p(I_t, \hat{I}_{t\leftarrow s}), \\ \bar{L_{gc}}(D_t, D_s, T_t^s) &= \min_s L_{gc}(D_t, D_s, T_t^s). \end{aligned} \tag{3.4}$$

Also exploiting the multi-scale predictions of the depth map, $\bar{L_p}$, $\bar{L_{gc}}$, and L_s are minimized over 4 output scales. Finally, the total self-supervised loss is formulated as:

$$\begin{aligned} L_{self} = \bar{L_p}(I_t, I_s, T_t^s) + \lambda_1 \bar{L_{gc}}(D_t, D_s, T_t^s) + \\ \lambda_2 L_s(I_t, D_t) + \lambda_3 L_r. \end{aligned} \tag{3.5}$$

Using this joint loss, we simultaneously learn both depth prediction and relative pose estimation model. We will provide details of our training setting on the KITTI Odometry dataset in Sec. 3.4.1.

3.4 Experiments

3.4.1 Experimental Setup

Datasets

We adopt the widely used KITTI Odometry benchmark [47, 48] to evaluate our method. The benchmark contains 22 urban and highway driving sequences, among which only 11 sequences (Sequence 00-10) have ground-truth trajectory labels from GPS/IMU readings. Following [15, 5, 197], we select Sequence 00-08 as training data and test the trained model on Sequence 09 and 10. In addition, we also report the results on the rest sequences of the KITTI Odometry benchmark in the same manner as in [197], which run the stereo version of ORB-SLAM2 to obtain reference trajectories.

Evaluation Metrics

We adopt the KITTI Odometry criterion, which reports the average translational error $T_{rel}(\%)$ and rotational errors $R_{rel}(^\circ/100m)$ of possible sub-sequences of length (100, 200, $\cdots$, 800) meters, as the main evaluation criteria. We also report the translational RMSE of the whole trajectory to evaluate the global trajectory accuracy. For self-supervised monocular methods, since the absolute scale is unknown, we scale and align the predicted trajectory to the ground-truth associated poses using [154] before evaluation.

Implementation Details

We implement our system based on the PyTorch framework [121]. Both the depth and pose networks receive input images of size 640×192 pixels. The batch size is set to 12 and the model is trained for 40 epochs. The learning rate is set to 1e-4 for the first 20 epochs and drops to 5e-5 for the remaining epochs. We train our full model using a single NVIDIA RTX 3090 and test the model using a single NVIDIA GTX 1080Ti. We empirically set the hyper-parameters as follows: λ_1=1e-2, λ_2=1e-3, λ_3=2e-5, α=1e-1.

3.4.2 Comparison of different fusion strategies

As mentioned in Sec. 3.3.2, we first evaluate various design choices of our relative pose estimators that leverage color and depth information. According to the stage of

Table 3.1 Comparison among different variants on sequences 09 and 10 of the KITTI Odometry dataset [47]. The best performance is in **bold**.

		Seq. 09			Seq. 10			Avg.		
Modality	Fusion	RMSE (m)	T_{rel} (%)	R_{rel} (deg/100m)	RMSE (m)	T_{rel} (%)	R_{rel} (deg/100m)	RMSE (m)	T_{rel} (%)	R_{rel} (deg/100m)
RGB	-	11.10	4.97	1.72	11.58	6.45	2.52	11.34	5.71	2.12
Depth	-	13.54	4.66	1.93	**6.04**	4.94	1.75	9.79	4.80	1.84
RGB+Depth	Early	13.77	5.22	1.79	10.42	5.56	2.13	12.10	5.39	1.96
RGB+Depth	Late	11.25	4.91	1.79	7.91	5.67	1.71	9.58	5.29	1.75
RGB+Depth	Middle	10.37	4.32	**1.37**	8.06	5.14	1.62	9.21	4.73	1.50
RGB+Depth	Multi-layer	**9.86**	**3.90**	1.41	7.36	**4.88**	**1.38**	**8.61**	**4.39**	**1.39**

fusions (cf. Fig. 3.1), we build 5 variants of the pose estimator and compare them with our proposed multi-layer fusion model on Sequence 09 and 10 of the KITTI Odometry dataset (Tab. 3.1). The first two variants are: 1) RGB: the ego-motion network only takes two RGB images as inputs. 2) Depth: the ego-motion network only takes two depth maps as inputs. The rest variants take both RGB and depth as inputs, but leverage different fusion strategies: 3) RGB+Depth (early fusion): the RGB and depth inputs are concatenated along channel dimension as an integrated input to a single ego-motion network. 4) RGB+Depth (late fusion): the RGB and depth streams are processed by two separate ego-motion networks and the outputs are combined using jointly learned weighting factors, which are normalized by an additional softmax layer [166]. 5) RGB+Depth (middle fusion): the RGB and depth features encoded from two separate encoders are fused before a single pose regressor using Soft Fusion [22]. 6) RGB+Depth (multi-layer fusion): our proposal as introduced in Sec. 3.3.2. For a justified comparison, we use the ResNet18 architecture as the encoder for all variants, and the regressors are all implemented as two fully-connected layers with ReLU activation and a final fully-connected layer. The first 5 variants are trained using the same loss function without regularization loss for CE.

One interesting observation from the validation results is that the early and late fusion methods sometimes show worse performance than the methods leveraging only a single modality. This result implies the importance of selecting an appropriate fusion strategy. Middle fusion provides rather better results than the early and late fusion methods while performing fusion for encoded features that include higher semantic information than the final decision. Multi-layer fusion strategy gives the best performance among these comparisons. We attribute the results to the fact that the CE operation in multiple layers evaluates encoded features from both coarse (more semantic) and fine (more structural) layers, which further exploit the

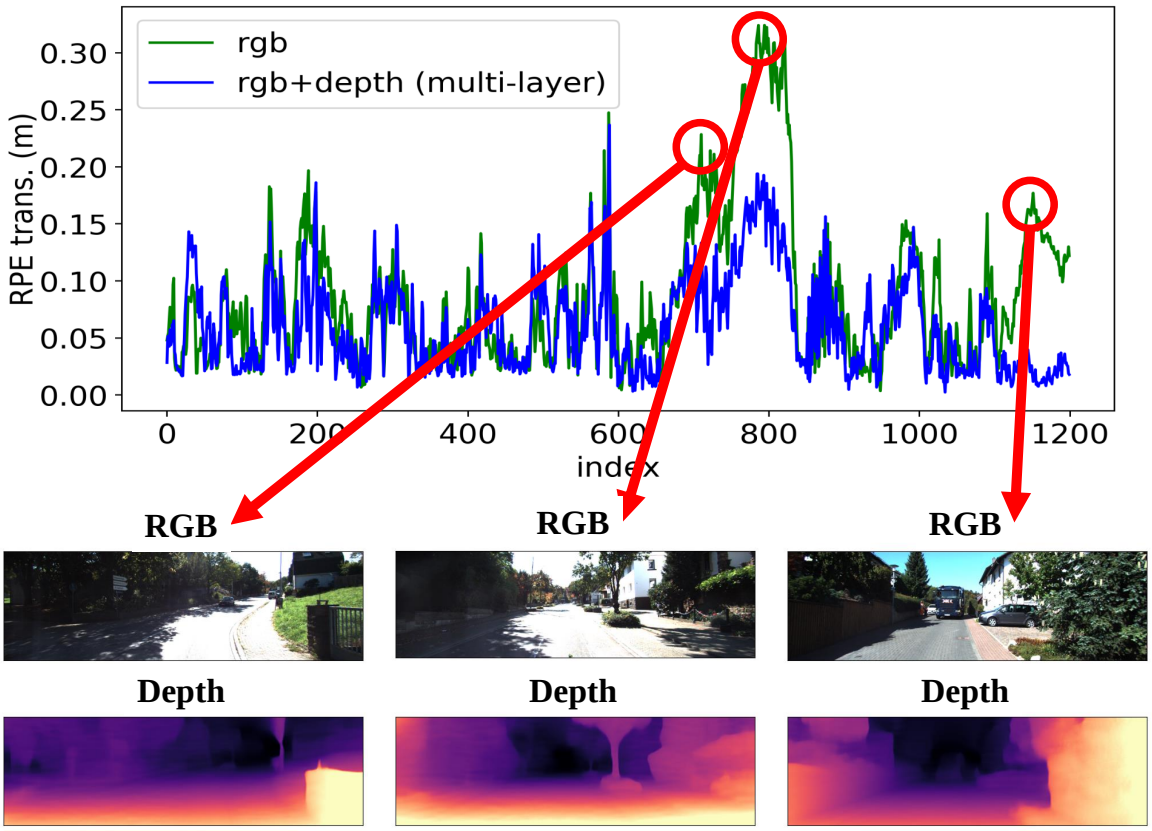


Fig. 3.3 Visual examples where fusing depth modality can help to obtain more accurate ego-motion. The top is the curve of relative translational errors of consecutive frames on Sequence 10 of the KITTI Odometry benchmark. The bottom lists RGB images and inferred depth map where our model performs explicitly better than the RGB-only model.

complementary properties between RGB and depth information. Based on these observations, we finally decided to employ the CE-based multi-layer fusion strategy for our relative pose estimation model. In the following comparisons, we refer to the proposed multi-layer fusion model as "MLF-VO".

Fig. 3.3 shows visual examples where fusing depth modality can help to obtain more accurate ego-motion. In the top of Fig. 3.3, we plot the curve of relative translational pose errors of all consecutive frames on Sequence 10. At the bottom, we list test scenes where our multi-level fusion network produces explicit smaller relative translational pose errors than the RGB-only network. We note the first two scenes show stronger lighting than the training data, and the third scene includes a dynamic object while the ego-motion is small. Ego-motion estimated using only RGB images results in relatively large pose errors in these specific sections. We

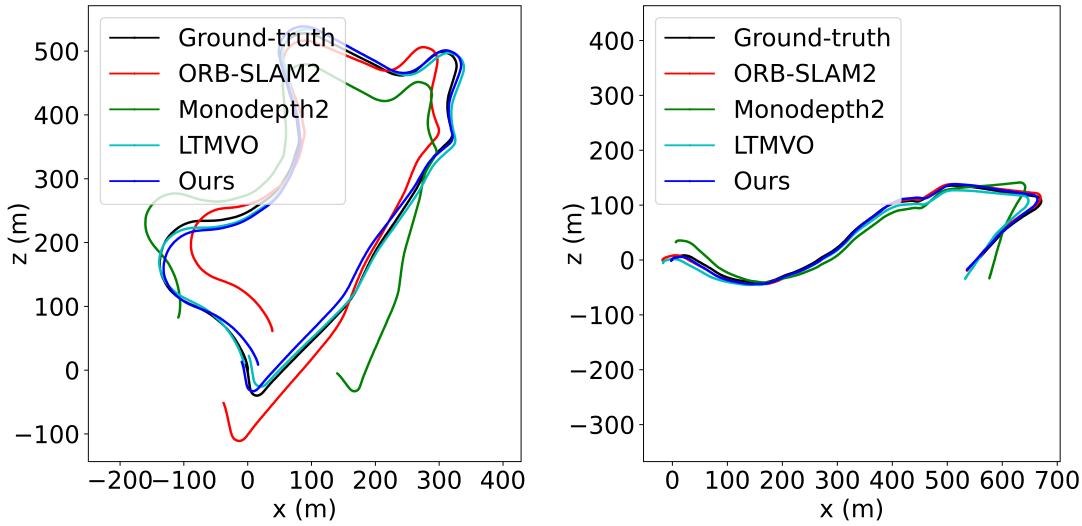
Table 3.2 Odometry results compared with the state-of-the-art methods. The best results of each block are highlighted by **bold** style. In the bottom block (Self Sup.), the best and the second best results are highlighted by **red** and blue characters, respectively.

	Method	Seq. 09			Seq.10			Avg.		
		RMSE (m)	T_{rel} (%)	R_{rel} (deg/100m)	RMSE (m)	T_{rel} (%)	R_{rel} (deg/100m)	RMSE (m)	T_{rel} (%)	R_{rel} (deg/100m)
Geo.	ORB-SLAM2-M (w/o LC) [117]	41.75	10.03	**0.29**	7.74	3.64	**0.32**	24.75	6.84	**0.31**
	ORB-SLAM2-M (w LC) [117]	**9.84**	**3.48**	0.39	**7.10**	**3.46**	0.38	**8.47**	**3.47**	0.39
Sup.	DeepVO [164]	-	-	-	-	8.11	8.83	-	-	-
	BeyondTracking [178]	-	-	-	-	**3.94**	**1.72**	-	-	-
Self Sup.	PoseGraph [97]	-	8.10	2.81	-	12.90	3.17	-	10.50	2.99
	Monodepth2 [53]	76.42	17.22	3.86	20.47	11.72	5.35	48.45	14.47	4.61
	SC-SfMLearner [15]	-	11.20	3.35	-	10.10	4.96	-	10.65	4.16
	TSN [5]	-	6.72	1.69	-	9.52	1.59	-	8.12	1.64
	LTMVO [197]	11.30	**3.49**	**1.00**	11.80	5.81	1.80	11.55	4.65	1.40
	MLF-VO (Ours)	**9.86**	3.90	1.41	**7.36**	**4.88**	**1.38**	**8.61**	**4.39**	**1.39**

attribute the failures to the lighting conditions of the sections, which produce heavily saturated regions in the input RGB image. On the other hand, an inferred depth map can still provide the structural information of these scenes. Therefore, our ego-motion estimation utilizing both of RGB and depth images largely improves the accuracy of estimated motion. These results also demonstrate that our multi-layer fusion in the relative pose estimator can retain complementary information extracted from different modalities for pose estimation.

3.4.3 Comparison with other methods

In Tab. 3.2, we compare our best model with several existing methods, including the monocular version of ORB-SLAM2 [117] which employs a classic geometry-based estimation (Geo.), supervised learning methods (Sup.) [164, 178] and self-supervised learning methods (Self Sup.) [97, 53, 15, 5, 197]. For ORB-SLAM2, we notice that it sometimes fails in the initialization stage, thus we run it 5 times on both test sequences and choose the medians of them as final results. For Monodepth2 [53], we use the pre-trained model provided by the authors. For the remaining methods, we take the results reported in their paper. '-' means the results are not available from that paper. Note that all supervised methods are trained on Sequence 00, 02, 08, 09 of the KITTI Odometry dataset, while the self-supervised methods are trained using the same data split as ours. Our MLF-VO achieves the best average performance among other self-supervised methods. Compared to Monodepth2 [53]



(a) Seq.09 (b) Seq.10

Fig. 3.4 Qualitative evaluation on Sequence 09 and 10 of KITTI Odometry benchmark.

and SC-SfMLearner [15], which are trained by a similar loss to ours, our final model outperforms them by a significant margin in the task of ego-motion estimation. TSN [5] consists of a two-stream network for RGB and depth which is similar to ours. Our method still shows superior results demonstrating that the fusion strategy plays an important role in the design of the network utilizing two or more modalities as inputs. LTMVO [197] introduces LSTM structures to capture temporal information from color and depth images. Compared to this state-of-the-art ego-motion estimation pipeline, the proposed MLF-VO also provides comparable accuracy of estimated motions, while requiring less computation (shown below). Furthermore, the proposed method is also comparable with the geometric method and supervised methods.

In Fig. 3.4, we illustrate our global trajectories with other methods on Sequence 09 and 10. ORB-SLAM2 [117] shows large scale-drift errors on Sequence 09 while our method and LTMVO [197] show superior global trajectories. Our global RMSEs of trajectories are smaller than all other self-supervised methods, which indicates that the global consistency is well preserved in our method.

To show the relevance of results, we further report our results on Sequence 11-21 in Tab. 3.3. Our method achieves competitive results among the learning-based methods. In terms of the RMSE and T_{rel} error, our method outperforms the other learning-based methods with a significant margin, and is comparable even with ORB-SLAM2-M with loop closure.

Table 3.3 Average results on Sequence 11-21 of KITTI Odometry benchmark.

Method		RMSE (m)	T_{rel} (%)	R_{rel} (deg/100m)
Geo.	ORB-SLAM2-M (w/o LC) [117]	81.20	19.60	0.94
	ORB-SLAM2-M (w LC) [117]	**44.09**	**12.96**	**0.71**
Self Sup.	Monodepth2 [53]	99.36	11.42	3.38
	SC-SfMLearner [15]	156.66	19.04	5.77
	LTMVO [197]	73.18	7.08	**1.56**
	MLF-VO (Ours)	**49.96**	**6.44**	2.07

Table 3.4 Single-view depth estimation results on Eigen test split of KITTI raw dataset [48].

Method	Error metric ↓				Accuracy metric ↑		
	Abs Rel	Sq Rel	RMSE	RMSE log	< 1.25	$< 1.25^2$	$< 1.25^3$
Monodepth2 [53]	0.115	0.903	4.863	0.193	**0.877**	0.959	0.981
SC-SfMlearner [15]	0.137	1.089	5.439	0.217	0.830	0.942	0.975
TSN [5]	0.139	1.063	5.349	0.221	0.817	-	-
LTMVO [197]	0.115	0.871	**4.778**	**0.191**	0.874	**0.961**	**0.982**
Ours	**0.114**	**0.849**	4.847	0.194	0.871	0.957	0.981

Evaluation of Monocular Depth Estimation

Although our main purpose is to improve the accuracy of ego-motion estimation, we also report our depth estimation results for completeness as a self-supervised joint depth and ego-motion estimation framework. When evaluating depth estimation, we train our model on the Eigen split of the KITTI raw dataset and cap the maximum depth to 80m, keeping the same configuration as other methods. In Tab. 3.4, we report our system's depth accuracy on the Eigen test split with other methods. Our method achieves similar performance to LTMVO [197] and Monodepth2 [53].

Computation time analysis

In addition to the accuracy evaluation, we also compare the inference time of our method with other methods. We obtain the runtime of ORB-SLAM2 from their official report on the KITTI Odometry benchmark, and the runtime of LTMVO from their public paper. Tab. 3.5 reports the comparison results with the image resolution and device. Given similar computational resources, our method is much faster than LTMVO [197]. We think this is because our network structure is simpler than LTMVO which employs an LSTM network.

Table 3.5 Pose inference time per image pair.

Method	Resolution	Device	Time (ms)
ORB-SLAM2-M [117]	376×1241	2-core CPU	60
LTMVO [197]	192×640	GTX TitanXP	70
MLF-VO (Ours)	192×640	GTX 1080Ti	**27**

3.5 Conclusion

In this paper, we have evaluated various strategies to fuse the RGB image and inferred depth image for ego-motion estimation. Through the validation of our baseline pipeline, we found several important observations for effectively fusing information from different modalities. We finally incorporate them into our pipeline and build a relative pose estimator that fuses modalities in multiple stages of the feature encoder and achieves state-of-the-art performance among self-supervised ego-motion estimation methods.

Chapter 4

EMR-MSF: Self-Supervised Recurrent Monocular Scene Flow Exploiting Ego-Motion Rigidity

4.1 Introduction

Scene flow estimation, which involves estimating both 3D structure and 3D motion of a dynamic scene from its two consecutive observations, has been receiving increasing attention due to its significance in areas such as robotics [32], augmented reality [73], and autonomous vehicles [112]. Recently, deep learning has demonstrated remarkable progress in the domain of scene flow estimation based on various input modalities, including stereo images [13, 74, 108, 135, 160, 133], RGB-D pairs [107, 126, 151, 109], or Lidar points [101, 57, 168, 177, 125, 171, 35, 24, 33, 161]. These methods, however, either require strict sensor calibrations (*e.g.*, stereo-based), or expensive devices (*e.g.*, RGB-D or Lidar-based) to achieve satisfactory performance, which restricts their widespread applications.

On the other hand, monocular scene flow estimation methods [18, 179, 180, 196, 104, 77, 69, 70, 12] which only require a monocular camera for obtaining both 3D structure and 3D motion, have been presented as an economical yet effective solution for dynamic 3D perception. The methods [18, 179] combined with supervised learning have yielded promising results, yet the primary challenge facing them has been the limited availability of ground-truth training data. To address this limitation, several multi-task methods [180, 196, 104, 165, 77] have been proposed to jointly learn the depth, 2D optical flow and camera ego-motion networks from monocular

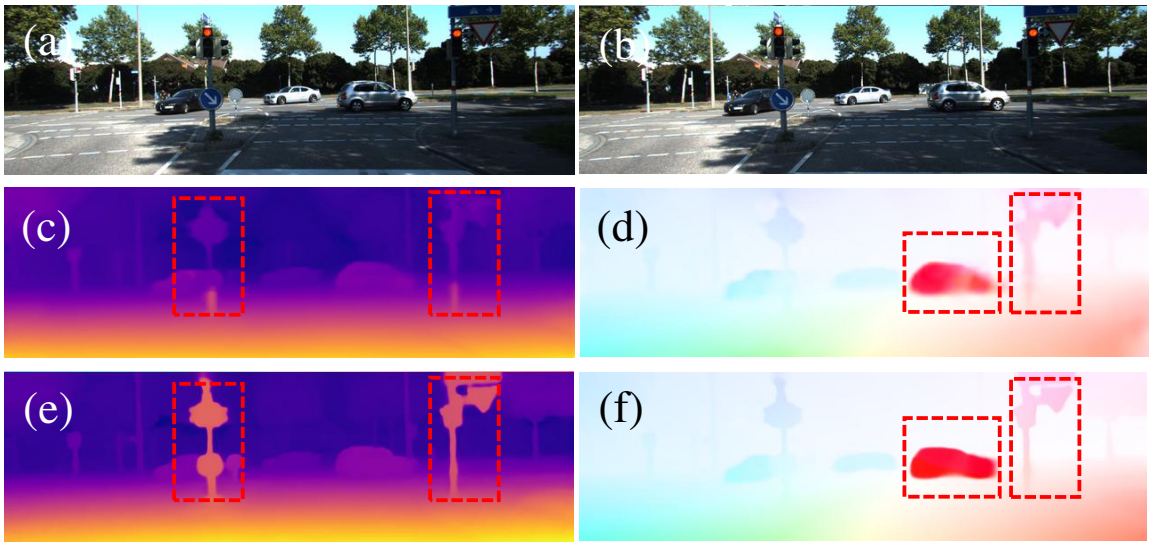


Fig. 4.1 **Comparison between our method and [69].** (a) input first frame, (b) input second frame, (c) depth of first frame from [69], (d) synthesized optical flow from [69], (e) depth of first frame from our method, (f) synthesized optical flow from our method. Our method generates more regularized and detailed predictions as shown in red boxes.

sequences in a self-supervised manner, and the scene flow can be calculated from the outputs. Recently, [69, 70, 12] have shown it feasible to train a single network to directly estimate both depth and 3D scene flow from two monocular images and outperform the previous multi-task methods. These methods typically build upon a standard optical flow pipeline (*e.g.*, PWC-Net [143] or RAFT [149]) as a basis and adapt it for monocular scene flow. Despite the notable progress achieved by these methods, their accuracy still lags behind the supervised monocular methods by a large margin.

In this paper, we propose a novel approach for self-supervised monocular scene flow estimation, which outperforms the previous methods significantly as shown in Fig.4.1. To introduce explicit 3D geometry-oriented property, we follow the network architecture proposed in the supervised RGB-D method RAFT-3D [151] that iteratively refines a dense SE3 motion field for scene flow estimation. This improvement of architecture compared to previous methods directly improves the performance to a new level, but we argue that it still lacks the usage of *Ego-Motion Rigidity (EMR)*, an important prior that pixels in static regions should have the same SE3 motion as the ego-motion. A novel module named ego-motion aggregation (EMA) is thus proposed to jointly estimate ego-motion as well as a rigidity soft

mask from the dense SE3 motion field. A new motion consistency loss is elaborately designed for constraining motion estimations in static areas represented by the rigidity soft mask. However, we notice that the network is inclined to select only a small subset of static regions which leads to a rigidity soft mask of low quality. To mitigate this problem, we adopt an efficient mask regularization loss to encourage the network to locate as many static regions as possible. Further performance improvement is attributed to our proposed training strategies including a gradient detachment technique and an improved view synthesis process.

Our main contributions are summarized as follows:

- We propose a novel self-supervised monocular scene flow estimation by incorporating 3D geometry-oriented network architecture property and exploiting ego-motion rigidity (EMR-MSF). To the best of our knowledge, we are the first method capable of jointly estimating depth, dense SE3 motion field and ego-motion from monocular images, as well as full scene flow derived from them.
- We introduce a novel ego-motion aggregation (EMA) module accompanied by a rigidity soft mask to precisely locate static regions for robust and accurate ego-motion estimation.
- We propose two new training losses to constrain the motion estimations in static regions, along with two effective training strategies to enhance the accuracy as explained in Sec. 4.3.3.
- We conduct extensive experiments to verify the effectiveness of our proposed method, resulting in a 44% accuracy boost in the SF-all metric compared to the previous state-of-the-art method on the task of monocular scene flow estimation, as well as superior results in monocular depth and visual odometry.

4.2 Related Work

4.2.1 Scene flow

As first introduced in [157], scene flow estimation is defined as the task of jointly estimating 3D structures and 3D motions for each scene point. The early studies [9, 68, 158–160] are based on stereo inputs and approach the scene flow estimation as an energy minimization problem. Recently, deep learning has demonstrated

powerful capabilities in end-to-end learning of scene flow estimation from stereo inputs [74, 108, 135]. Additionally, approaches that leverage pre-existing 3D structure through inputs of RGB-D sequences [107, 126, 151, 109] or Lidar points [101, 177, 125, 171, 35, 33, 161] have also been proposed for various scenarios.

4.2.2 Monocular scene flow

The advancement of deep learning techniques has facilitated the acquisition of scene flow solely from monocular images, with early methods relying on supervised learning [18, 179]. To exploit vast amounts of unlabeled data, a multitude of self-supervised multi-task approaches [180, 165, 196, 104, 77, 100] have been introduced that jointly predict depth, 2D optical flow, and camera motion from monocular sequences. While the recovery of scene flow is possible using the aforementioned outputs, the accuracy of such estimations is notably inadequate in temporally occluded areas. Hur et al. [69] first present a novel self-supervised model capable of inferring depth and 3D motion field from monocular sequences, which surpasses the performance of previous multi-task methods. Subsequent studies extend their method into a multi-frame model [70], or employ a recurrent network architecture [12] for better accuracy.

4.2.3 Rigidity in Scene Flow

Scene flow estimation can benefit from prior knowledge about rigidity, which assumes that pixels belonging to the same rigid object should undergo the same rigid transformation. To leverage the rigidity information in the scene, object detection or segmentation networks are commonly used to identify rigid instances and incorporated in scene flow estimation methods [108, 21, 133, 13] for better performance. Teed et al. [151] first propose the rigid-motion embeddings which softly and differentiably group pixels into rigid objects to exploit object-level rigidity. On the other hand, ego-motion rigidity, where the motion of pixels in static regions is constrained by the camera ego-motion, is widely used in self-supervised multi-task methods [180, 165, 196, 104, 77, 100] but often in a hard and non-differentiable way. In contrast, our proposed method jointly reasons ego-motion and rigidity soft mask in a fully differentiable manner, providing more robust and accurate scene flow estimation.

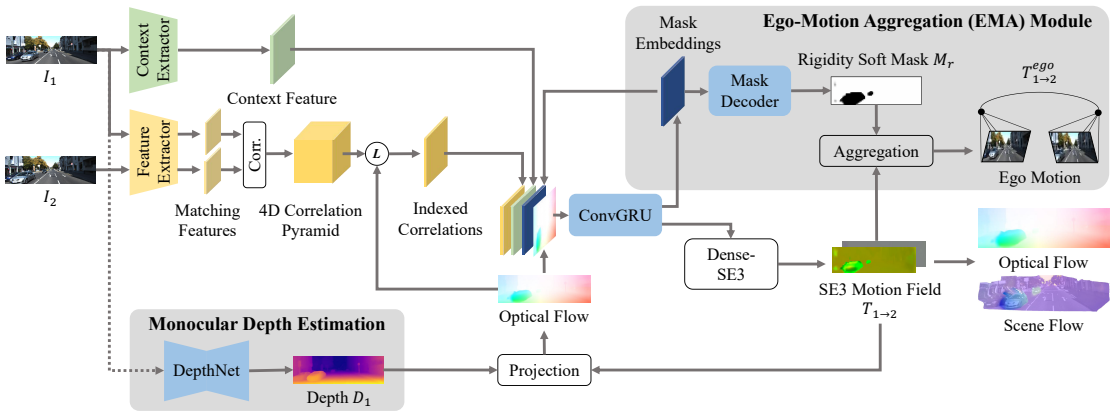


Fig. 4.2 **Proposed network architecture**. We highlight the different parts from RAFT-3D [151] with the shaded boxes, including 1) end-to-end trainable monocular depth estimation that substitutes the estimated depths for fixed input depths in the original structure, 2) an ego-motion aggregation (EMA) module for inferring ego-motion along with a learnable rigidity soft mask for locating static regions.

4.3 Proposed Method

Given two temporally consecutive monocular images $\{I_1, I_2\} \in \mathbb{R}^{H\times W\times 3}$, our method aims to recover 1) the corresponding depth maps $D_1, D_2 \in \mathbb{R}^{H\times W\times 1}$, 2) the dense SE3 motion field $T_{1\to2} \in SE(3)^{H\times W}$ that assigns a rigid transformation to each pixel of I_1 to I_2, and 3) the ego-motion $T^{ego}_{1\to2} \in SE(3)$ from I_1 to I_2. The optical flow $F_{1\to2} \in \mathbb{R}^{H\times W\times 2}$ and scene flow $S_{1\to2} \in \mathbb{R}^{H\times W\times 3}$ from I_1 to I_2 can be further recovered from the estimated D_1 and $T_{1\to2}$. In the following sections, we will begin by providing an overview of the proposed network architecture which incorporates effective designs for 3D estimations from a supervised method (Sec. 4.3.1). Afterward, we provide a detailed description of the proposed ego-motion aggregation (EMA) module that we utilize for estimating ego-motion, as well as a learnable rigidity soft mask for effectively locating static regions (Sec. 4.3.2). Finally, we elaborate on our self-supervised training in Sec. 4.3.3, which includes novel loss functions designed to fully exploit ego-motion rigidity, as well as improved training strategies.

4.3.1 Network Overview

Fig. 4.2 demonstrates the overview of our network. We highlight the different parts of our network compared to RAFT-3D [151], which is the basis of our network architecture, inside the shaded boxes. Our network consists of five stages: 1) monocular depth estimation, 2) feature extraction 3) correlation computing, 4)

iterative refinement, and 5) ego-motion aggregation. We first employ a monocular depth network to estimate the depth maps of input images instead of the fixed depths used in the original RAFT-3D structure. We adopt SDFA-Net [191] for depth estimation for its superior performance, which infers disparity from a monocular image under the assumption of a fixed baseline, and further converts the disparity into depth using pre-known focal length and baseline values. For feature extraction, correlation computing, and iterative refinement, we utilize the designs of RAFT-3D, which include the construction of a 4D all-pairs correlation pyramid from extracted features of input images and the use of a ConvGRU unit followed by a Dense-SE3 layer for iterative residual refinement of the SE3 field estimate. The ego-motion aggregation (EMA) module is employed to further infer ego-motion from the estimated SE3 motion field, which is elaborated on in the next section. The 3D scene flow and 2D optical flow can be synthesized from the estimated depth and SE3 motion field for various applications.

4.3.2 Ego-Motion Aggregation

As demonstrated in our ablation study 4.4.3, the joint learning of the depth and dense SE3 motion field in the self-supervised scenario can lead to significant ambiguities between the estimations of structure and motion, where the estimated SE3 motions of pixels belonging to the same rigid object, *e.g.*, the static regions, may be inconsistent. To mitigate such ambiguities, we incorporate ego-motion estimation into the joint learning to provide additional constraints in static regions. We propose to aggregate the ego-motion from the estimated SE3 motion field in contrast to previous multi-task methods [165, 180, 196], which utilize a separate network to regress ego-motion from input images. Furthermore, to handle the dynamic regions which are non-relevant to ego-motion, we introduce a learnable rigidity soft mask to predict per-pixel rigidity, thus locating static regions for stable ego-motion estimation.

Our ego-motion aggregation module proceeds in three steps, as shown in the upper-right corner of Fig. 4.2. We first incorporate the mask embeddings, a 16-channel feature map initialized to zero values, as new inputs and outputs to the convGRU unit, which is iteratively updated alongside the SE3 motion field. Next, we decode the mask embeddings using a mask decoder consisting of two convolutional layers and a sigmoid activation layer to obtain the rigidity soft mask M_r. The rigidity soft mask assigns a probability to each pixel, indicating the probability of it belonging to the static region. In the final step, we derive the ego-motion as an aggregation of the estimated SE3 motion field based on the learned rigidity soft mask, which is

formulated as:

$$T^{ego}_{1\to 2} = \mathrm{Exp}(\frac{\sum M_r \mathrm{Log}(T_{1\to 2})}{\sum M_r}), \tag{4.1}$$

where Log(·) maps SE(3) components to the Lie algebra, and Exp(·) performs the inverse operation.

As the ego-motion is differentiably computed from the SE3 motion field, the learning of ego-motion will implicitly impose constraints on the estimation of the SE3 motion field. In the next section, we further combine the self-supervised losses with two new losses utilizing the ego-motion estimation and learned rigidity soft mask to explicitly regularize the motion estimations in static regions.

4.3.3 Self-supervised Training

Self-supervised Loss

To enable self-supervised training, the estimated depth D_1 of the first image and the SE3 motion field $T_{1\to 2}$ are first converted into the scene flow representation $(u, v, \Delta D)$ with known camera intrinsics [109], where (u, v) denotes the standard optical flow $F_{1\to 2}$, and ΔD denotes the depth change registered to the first frame I_1. We denote $\overline{D}_1 = D_1 + \Delta D$, which represents the transformed depth map registered to the first frame. We obtain the 2D rigid flow $F^{ego}_{1\to 2}$ in the same manner by replacing $T_{1\to 2}$ with $T^{ego}_{1\to 2}$. The losses for our joint self-supervised learning are introduced as follows:

Temporal Photometric loss. We minimize the photometric differences between the original image and the synthesized images from flow field $F_{1\to 2}$ and $F^{ego}_{1\to 2}$, formulated by

$$L_p = \frac{1}{HW} \sum M_{noc} \odot pe(I_1, w(I_2, F_{1\to 2})), \tag{4.2}$$

$$L^{ego}_p = \frac{1}{HW} \sum M_{ol} \odot M_{noc} \odot pe(I_1, w(I_2, F^{ego}_{1\to 2})), \tag{4.3}$$

$$pe(I_a, I_b) = \frac{\alpha}{2}(1 - \mathrm{SSIM}(I_a, I_b)) + (1 - \alpha)\,|I_a - I_b|\,, \tag{4.4}$$

where $\frac{1}{HW}\sum$ is used for the notation of the mean over all pixels and $\odot$ means element-wise multiplication. $w(\cdot,\cdot)$ is the view synthesis function with the flow field and $pe(\cdot,\cdot)$ measures the photometric difference between two images. The occlusion mask M_{noc} is derived from the forward-backward consistency check [111] using $F_{1\to 2}$ and $F_{2\to 1}$. We additionally use an outlier mask M_{ol} [75] for calculating L^{ego}_p, which masks out pixels with either large photometric errors mainly resulting from possible

occluded or moving regions, or very small photometric errors mainly resulting from textureless regions. Note that M_r is not leveraged here for two reasons: 1) M_{ol} performs more stable than the learned mask M_r at the beginning of training. 2) M_{ol} can better locate pixels which are informative for learning ego-motion estimation.

Spatial Photometric Loss. To address scale ambiguity in monocular scene flow learning, we utilize stereo samples during training as proposed in previous works [69, 70, 12]. We use the stereoscopic image synthesis loss utilized in [191] to regularize depth estimation on an absolute scale and denote it as L_d in our method.

Geometric loss. To constrain the estimated motion field in 3D space, we exploit the geometric consistency between the transformed depth map $\overline{D}_1$ and estimated D_2:

$$L_g = \frac{1}{HW} \sum M_{noc} \odot ge(\overline{D}_1, w(D_2, F_{1\to2})), \tag{4.5}$$

$$ge(D_a, D_b) = \frac{|D_a - D_b|}{D_a + D_b}, \tag{4.6}$$

where $ge(\cdot,\cdot)$ measures the normalized difference [15] between two depth maps.

Smoothness loss. The k-th order edge-aware smoothness loss function is defined as:

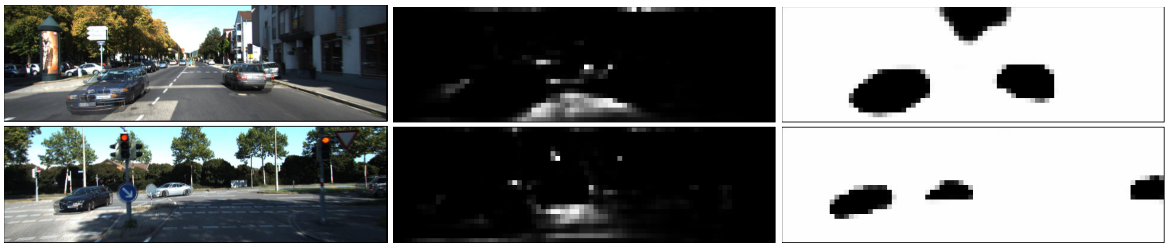
$$L_s(O) = \frac{1}{HW} \sum \left|\frac{\partial^k O}{\partial x^k}\right| e^{-\beta\left|\frac{\partial I_1}{\partial x}\right|} + \left|\frac{\partial^k O}{\partial y^k}\right| e^{-\beta\left|\frac{\partial I_1}{\partial y}\right|}, \tag{4.7}$$

where O is a dense prediction, which can be $\mathrm{Log}(T_{1\to2})$, D_1 and $F_{1\to2}$ in our case. We apply first-order edge-aware smoothness loss to $\mathrm{Log}(T_{1\to2})$ and D_1, denoted as $L_{s,t}$ and $L_{s,d}$ separately, and apply second-order edge-aware smoothness loss to $F_{1\to2}$ as $L_{s,f}$. The total smoothness loss is calculated as $L_s = \lambda_{st}L_{s,t} + \lambda_{sd}L_{s,d} + \lambda_{sf}L_{s,f}$.

Motion Consistency Loss. To further regularize the SE3 motion field in static regions, we propose to explicitly constrain the motion estimations in these regions to be consistent with the estimated ego-motion, formulated as:

$$L_c = \frac{1}{HW} \sum M_r \odot |\mathrm{Log}(T_{1\to2}) - \mathrm{Log}(T^{ego}_{1\to2})|, \tag{4.8}$$

Mask Regularization loss. We observe that the estimated rigidity soft mask tends to degenerate during training. This is intuitively reasonable since theoretically the ego-motion can be represented as the SE3 motion of any single pixel in static regions, thus the rigidity soft mask is inclined to select only a small subset of static regions due to L_c. To address this problem, we propose a mask regularization loss to encourage the rigidity soft mask to locate static regions as many as possible for fully exploiting



Overlayed inputs W/o L_m W L_m

Fig. 4.3 **Visualization of estimated rigidity soft masks**. The middle column shows the degeneration cases of the estimated rigidity soft mask, which is solved by introducing L_m during training.

ego-motion rigidity in static regions, which is formulated as:

$$L_m = \frac{1}{HW} \sum \frac{1 - M_r}{\gamma + M_r}, \tag{4.9}$$

where γ is a hyper-parameter. We provide a visual comparison of the estimated rigidity soft mask without and with L_m in Fig. 4.3.

Total Loss. We calculate losses for both the final and intermediate estimations from our recurrent structure. We use an upper-right index $(\cdot)^i$ to denote the losses related to the i-th iteration. The total loss of our method can be summarized as:

$$\begin{aligned} L_{total} = L_d + \sum_{i=1}^{N} \zeta^{N-i} \Big(& L_p^i + L_p^{ego,i} \\ & \lambda_g L_g^i + \lambda_s L_s^i + \lambda_c L_c^i + \lambda_m L_m^i \Big), \end{aligned} \tag{4.10}$$

where N is the iteration number, ζ is the weight decay factor, and $\lambda = [\lambda_g, \lambda_s, \lambda_c, \lambda_m]$ is the set of hyper-parameters balancing different losses.

Improved Training Strategies

Gradient Detachment. Our loss functions except L_d are calculated for both the final and intermediate estimations of motion field and ego-motion for preventing divergence of training. However, joint learning of depth and coarse motion estimations from early iterations can hinder the learning of the depth network. To address this issue, we propose to detach the gradients of depth estimations when calculating losses using intermediate motion estimations, which ensures that joint learning only occurs when the finest motion estimations are utilized.

Improved view synthesis process. We leverage the full-image warping technique proposed in [140] to provide better supervisory signals at image boundaries during the calculation of photometric loss, which uses cropped images as inputs to the network, but refers to the uncropped images when performing view synthesis. We further leverage this idea during the calculation of geometric loss in Eqn. 4.5, where we refer to the estimated depths of uncropped images for depth synthesis.

4.4 Experimental Results

Our proposed method is evaluated on various tasks including scene flow, monocular depth, and visual odometry.

4.4.1 Implementation Details

We implement our network with Pytorch [122]. All components of our network are trained from scratch, except the encoder in the depth network and the context extractor, which use ImageNet [30] pretrained weights. We use the Adam optimizer [102] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ to train our network. During training, the images are first resized into the resolution of 800×240, and cropped off the top, bottom, left and right 10% pixels to obtain the input images of 640×192 to leverage the improved view synthesis process. During the test, the images are resized into 640×192 for processing and the results are bilinearly rescaled back to the original size for evaluation. We use a two-staged training process for better stability of our method. During the first stage, we separately train the depth network using spatial photometric loss L_d and depth smoothness loss $L_{s,d}$. Then, we train our full network using the total loss L_{total} for the rest epochs. The training is carried on for 50 epochs total, 20 epochs for the first stage, and 30 epochs for the second stage. The initial learning rate is set to 1e-4, and downgraded by half at epoch 20, 25, 30, and 40. The hyper-parameters of our method are set as: $[\alpha, \beta, \gamma, \zeta] = [0.15, 10, 1, 0.9]$, $[\lambda_{s,t}, \lambda_{s,d}, \lambda_{s,f}] = [0.001, 1, 1]$, $[\lambda_g, \lambda_s, \lambda_c, \lambda_m] = [0.1, 0.1, 0.1, 0.1]$, $N = 12$. For data augmentation, we employ random color augmentation, random horizontal flipping and random time order switching. We use the LieTorch [152] library to perform backpropagation of the SE3 motion field.

Our network is trained using a batch size of 8 on a machine equipped with 4 GTX 3090 GPUs for all experiments. The training takes about two days when the iteration number is equal to 12. Color augmentation, horizontal flipping

augmentation, and time-order switching augmentation are applied with a probability of 50% for each during the experiment. For color augmentation, we adopt random gamma adjustments (uniformly sampled from $[0.8, 1.2]$), brightness adjustments (with a multiplication factor uniformly sampled from $[0.5, 2.0]$) and color channel adjustments (with a multiplication factor uniformly sampled from $[0.8, 1.2]$ for each color channel). To ensure stable initialization of the full network during the second stage of training, we disable the use of the non-occlusion mask M_{noc} when calculating the losses L_p and L_g, and remove the mask regularization loss L_m for the first 3k iterations during the second-stage training.

4.4.2 Datasets and Evaluation Metrics

Datasets

For the scene flow task, we use the same data setting as previous self-supervised monocular scene flow methods [69, 70, 12], which use KITTI Scene Flow Training and Testing as two test sets, and spilt the remaining data into 25801 samples for training and 1684 samples for validation. For comparison in the task of monocular depth estimation, we follow the data split used in [191], but remove the samples which are the last images of sequences, which gives us 22568 samples for training and 1774 for validation. The depth evaluation is conducted on the Eigen Test split [38], which contains 697 images with ground-truth labels. For the task of visual odometry, we use the official odometry data split, which uses Seq. 00-08 for training and Seq. 09-10 for testing, as done in [165, 197, 76].

Metrics

We follow the evaluation metric of KITTI Scene Flow benchmark [112] for scene flow estimation, which evaluates the outlier rate of the disparity for the reference frame (D1-all) and for the target image mapped into the reference frame (D2-all), as well as of the optical flow (F1-all). The outlier rate of the scene flow (SF-all) is obtained by checking if a pixel is an outlier on either of them. For monocular depth evaluation, we use the publicly used metrics, including Abs Rel, Sq Rel, RMSE, logRMSE, A1= $\delta < 1.25$, A2= $\delta < 1.25^2$ and A3= $\delta < 1.25^3$. For visual odometry evaluation, we adopt the KITTI odometry criterion, which reports the average translational error T_{rel} and rotational error R_{rel} of possible sub-sequences of length (100, 200, 800) meters as the main criteria.

Table 4.1 **Quantitative ablation study of key components.** EMR: Ego-Motion Rigidity, MRL: Mask Regularization Loss, GD: Gradient Detachment, IVS: Improved View Synthesis. All components effectively improve the performance, especially the EMR component.

EMR	MRL	GD	IVS	D1-all ↓	D2-all ↓	F1-all ↓	SF-all ↓	EPE-noc ↓	EPE-occ ↓	EPE-all ↓
-	-	-	-	13.20	21.90	14.16	28.15	2.78	12.57	4.83
√	-	-	-	9.99	18.25	13.55	24.21	2.68	10.85	4.44
√	√	-	-	9.83	17.25	13.65	23.07	2.69	10.17	4.23
√	√	√	-	9.39	16.90	13.51	22.86	2.65	10.04	4.21
-	-	√	√	11.73	18.76	12.54	24.80	2.74	7.63	3.81
√	√	√	√	**9.03**	**15.42**	**11.93**	**21.17**	**2.53**	**7.07**	**3.56**

Table 4.2 **Ablation study of the iteration number.** More iterations give better performance up to about 12, but with a slower speed.

Iter. Num.	D1-all ↓	D2-all ↓	F1-all ↓	SF-all ↓	Runtime
2	9.03	15.42	11.93	21.17	127 ms
4	8.65	13.93	**11.36**	19.05	151 ms
8	8.38	13.14	11.76	18.31	204 ms
12	**8.37**	**12.86**	11.58	**18.11**	250 ms

4.4.3 Ablation Studies

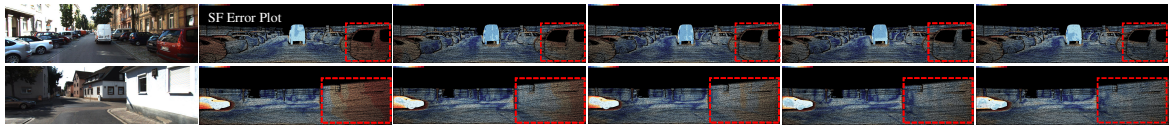
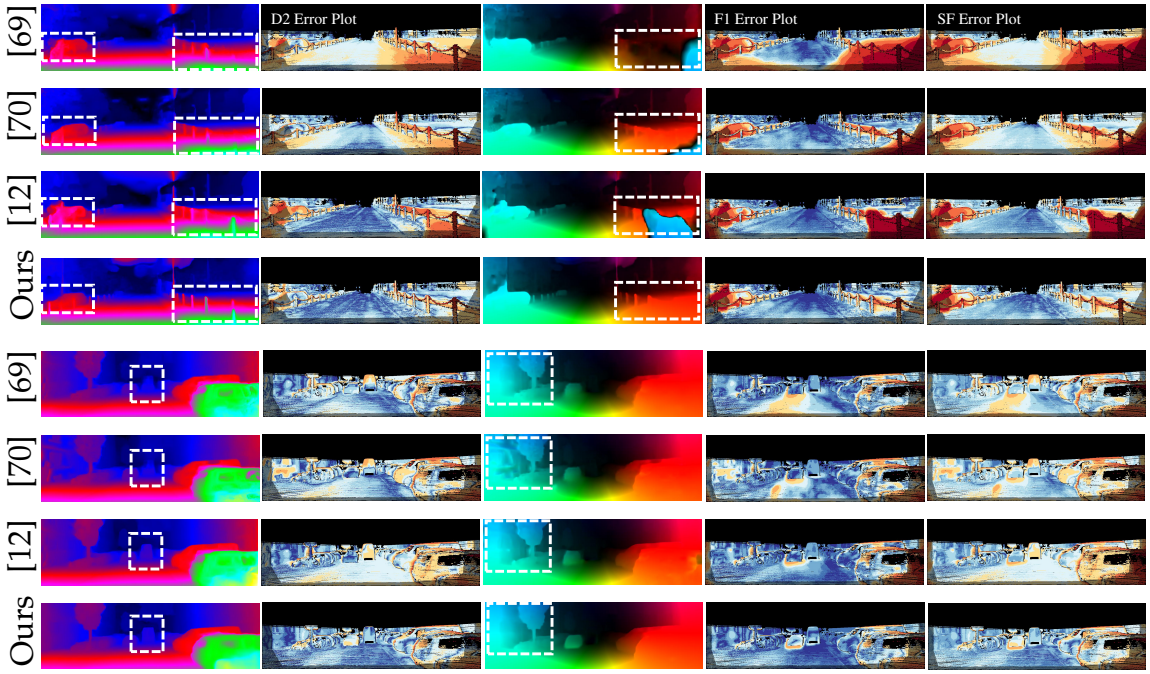


Fig. 4.4 **Qualitative ablation study of proposed components.** The erroneous predictions are gradually reduced by incorporating proposed components as shown in the red boxes.

We first conduct ablation studies to verify the effectiveness of each proposed component of our method on the task of scene flow estimation, including 1) ego-motion rigidity (EMR), which includes the ego-motion aggregation module and losses for L_p^{ego} and L_c, 2) mask regularization loss L_r (MRL), 3) gradient detachment technique (GD), and 4) improved view synthesis (IVS). For efficiency, the ablation studies are conducted using iteration number equal to 2. We report both the scene flow metrics and end-point-error (EPE) of synthesized optical flow in Tab. 4.1. Each proposed component proves to be effective in improving the overall scene flow accuracy. The largest performance gain is obtained by exploiting the ego-motion

Fig. 4.5 **Qualitative evaluation on KITTI Scene Flow Testing set.** We compare our method with Self-Mono-SF [69], Multi-Mono-SF [70] and RAFT-MSF [12] for two scenes using the visualizations provided by the KITTI benchmark [112]. From left to right: disparity visualization of I_t, $D2$ error plot, optical flow visualization, corresponding $F1$ error plot and combined SF error plot.



rigidity, which is in line with our expectation that ego-motion rigidity is an important prior in the task of scene flow estimation. Fig. 4.4 gives a visualization of the achieved error reduction on SF-all error plots from each component. The erroneous estimations in static regions and image boundaries are largely reduced by incorporating our contributions. The ablation study on the iteration number is reported in Tab. 4.2. The performance is about to reach convergence when the iteration number is 12. We also report the runtime for efficiency comparison, which is tested on a single GTX 3090 device for each model. For the following experiments, we always set the iteration number to 12.

Table 4.3 **Quantitative evaluation of the scene flow on the KITTI Scene Flow Training set and Testing set.** The best results are in **bold**.

Method	KITTI Scene Flow Training Set				KITTI Scene Flow Testing Set			
	D1-all ↓	D2-all ↓	F1-all ↓	SF-all ↓	D1-all ↓	D2-all ↓	F1-all ↓	SF-all ↓
Mono-SF [18]	16.72	18.97	11.85	21.60	16.32	19.59	12.77	23.08
GeoNet [180]	49.54	58.17	37.83	71.32	-	-	-	-
DF-Net [196]	46.50	61.54	27.47	73.30	-	-	-	-
EPC++ [104]	23.84	60.32	19.64	-	-	-	-	-
Self-Mono-SF [69]	31.25	34.86	23.49	47.05	34.02	36.34	23.54	49.54
Multi-Mono-SF [70]	27.33	30.44	18.92	39.82	30.78	34.41	19.54	44.04
RAFT-MSF [12]	18.34	23.65	17.51	30.97	21.21	27.51	18.37	34.98
EMR-MSF (Ours)	**8.37**	**12.86**	**11.58**	**18.11**	**9.70**	**14.51**	**11.93**	**19.74**

4.4.4 Comparison with State-of-the-art Methods

Scene Flow Evaluation

We compare our method with other state-of-the-art monocular scene flow methods on both the KTTII Scene Flow Training set and Testing Set as shown in Tab. 4.3. Our method achieves the best performance among all methods based on self-supervised learning, and even outperforms Mono-SF [18], which is a hybrid method based on the combination of supervised monocular depth estimation and energy minimization. In Fig. 4.5, we visualize the estimations and error maps of our method and other methods on samples from the KITTI Scene Flow Testing set. In the highlighted regions, our method shows better regularized and detailed estimations compared to other methods which give no consideration to exploit ego-motion rigidity. The error maps of various metrics are provided for better visualization.

Optical Flow Evaluation

Table 4.4 presents the quantitative comparison of optical flow estimation results of our method with additional self-supervised multi-task methods on the KITTI Scene Flow Training set and Testing set. The training settings of our method are the same as those used in experiments for scene flow evaluation. Our method outperforms all other compared methods on the KITTI Scene Flow Training set. On the KITTI Scene Flow Testing set, our method is slightly surpassed by [100] which requires stereo images during testing, whereas our method only relies on monocular images for testing.

Table 4.4 **Quantitative evaluation of the optical flow on the KITTI Scene Flow Training set and Testing set.** The best results are in **bold**. Methods marked with (*) use stereo images for estimation.

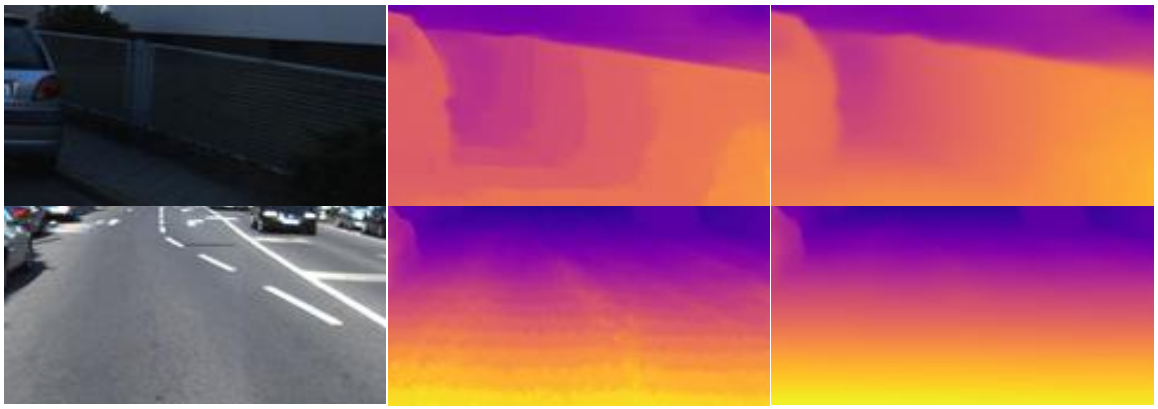
Method	Training set		Testing set
	EPE	F1-all	F1-all
GeoNet [180]	10.81	-	-
DF-Net [196]	8.98	26.01	25.70
Self-Mono-SF [69]	7.51	23.49	23.54
Multi-Mono-SF [70]	-	18.92	19.54
CC-uft [131]	5.66	20.93	25.27
UnOS [165]	5.58	-	18.00
EPC++ [104]	5.43	19.64	20.52
RAFT-MSF [12]	-	17.51	18.37
UnRigidFlow* [100]	5.19	14.68	**11.66**
EffiScene* [77]	4.20	14.31	13.08
EMR-MSF (Ours)	**3.46**	**11.58**	11.93

Table 4.5 **Quantitative evaluation of the monocular depth on the KITTI Eigen split.** M: trained on monocular videos, S: trained on stereo pairs. MS: trained on stereo videos. The best results are in **bold**.

Method	Sup.	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	A1 ↑	A2 ↑	A3 ↑
Monodepth2 [53]	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
PackNet-SfM [58]	M	0.111	0.785	4.601	0.189	0.878	0.960	0.982
DIFFNet [189]	M	0.102	0.764	4.483	0.180	0.896	0.965	0.983
RA-Depth [65]	M	**0.096**	**0.632**	**4.216**	**0.171**	**0.903**	**0.968**	**0.985**
Monodepth2 [53]	S	0.109	0.873	4.960	0.209	0.864	0.948	0.975
FAL-Net [55]	S	0.097	0.590	3.991	0.177	0.893	0.966	0.984
PLADE-Net [54]	S	0.092	0.626	4.046	0.175	0.896	0.965	0.984
SDFA-Net [191]	S	**0.090**	**0.538**	**3.896**	**0.169**	**0.906**	**0.969**	**0.985**
EPC++ [104]	MS	0.127	0.936	5.008	0.209	0.841	0.946	0.979
Self-Mono-SF [69]	MS	0.125	0.978	4.877	0.208	0.851	0.950	0.978
Monodepth2 [53]	MS	0.106	0.818	4.750	0.196	0.874	0.957	0.979
DIFFNet [189]	MS	0.101	0.749	4.445	0.179	0.898	0.965	0.983
RAFT-MSF [12]	MS	0.093	0.781	4.321	0.186	0.901	0.960	0.981
EMR-MSF (Ours)	MS	**0.088**	**0.552**	**3.946**	**0.169**	**0.905**	**0.970**	**0.986**

Monocular Depth Evaluation

We compare our method trained on the KITTI Eigen split with other state-of-the-art monocular depth methods as shown in Tab. 4.5. We use the 697 test samples for



RGB SDFA-Net [191] EMR-MSF (Ours)

Fig. 4.6 **Visualization of estimated depth.** We compare our results with SDFA-Net [191].

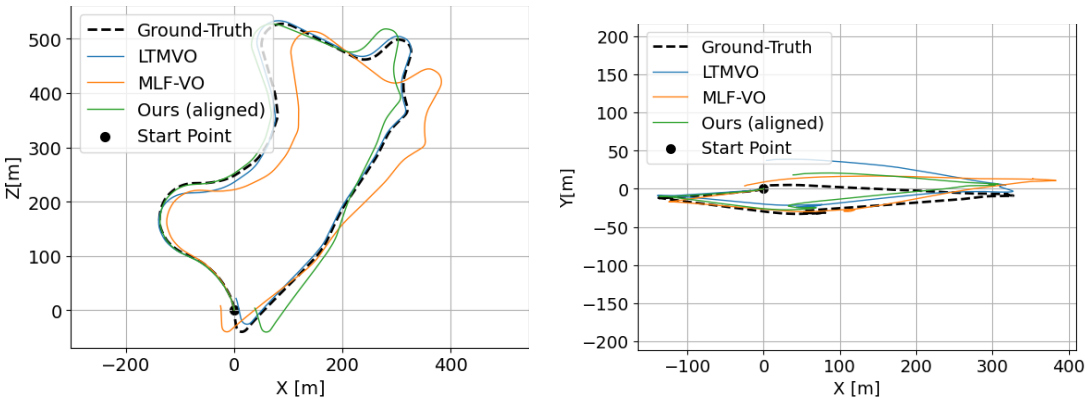
evaluation. We split the compared methods into methods using monocular videos for training, methods using stereo pairs for training, and methods using stereo videos for training. Our method achieves the best performance in 4 metrics among all compared methods and second best in the left 3 metrics. A visual comparison between our results and [191] is given in Fig. 4.6. Our method produces smoother depth estimations than [191], which we attribute to the joint learning of depth and motion. for which we think the main reason is the imperfect motion estimation and occlusion handling when calculating temporal losses. On the other hand, incorporating temporal losses produces smoother estimations in planes than [191].

Visual Odometry Evaluation

Finally, we compare the performance of our method trained on the KITTI Odometry split with other monocular methods in the task of visual odometry, including ORB-SLAM2 [117], a traditional method, as well as other self-supervised learning-based methods. We provide both results of ORB-SLAM2 with and without loop closure. For evaluating monocular methods, we perform the scale alignment to align the predicted up-to-scale trajectories to the ground-truth associated poses using [154]. Since our method leverages stereo samples during training, it is possible for our method to predict trajectories on a real scale. For a fair comparison, we provide both aligned and not aligned trajectories of our method in the table. As shown in Table 4.6, our method outperforms the previous self-supervised learning-based methods in all metrics, and even achieves better accuracy than traditional methods with loop closure in terms of the t_{err} metric. This demonstrates the effectiveness of our

Table 4.6 **Quantitative evaluation of the visual odometry.** The best results are highlighted by **bold** style.

Method	Seq.09		Seq.10	
	t_{err} (%) ↓	r_{err} (°/100) ↓	t_{err} (%) ↓	r_{err} (°/100) ↓
ORB-SLAM2 (w/o LC) [117]	10.03	0.29	3.64	0.32
ORB-SLAM2 (w LC) [117]	3.48	0.39	3.46	0.38
GeoNet [180]	39.43	14.30	28.99	8.85
Monodepth2 [53]	17.22	3.86	11.72	5.35
EPC++ [104]	8.84	3.34	8.86	3.18
LTMVO [197]	3.49	1.00	5.81	1.80
MLF-VO [76]	3.90	1.41	4.88	1.38
EMR-MSF (Ours)	3.49	0.78	3.11	1.04
EMR-MSF (Ours, aligned)	**3.30**	**0.78**	**2.35**	**1.04**



(a) Top view (b) Front view

Fig. 4.7 **Trajectories on Sequence 09 of KITTI Odometry benchmark**. Both the top view and front view are provided for better visualization.

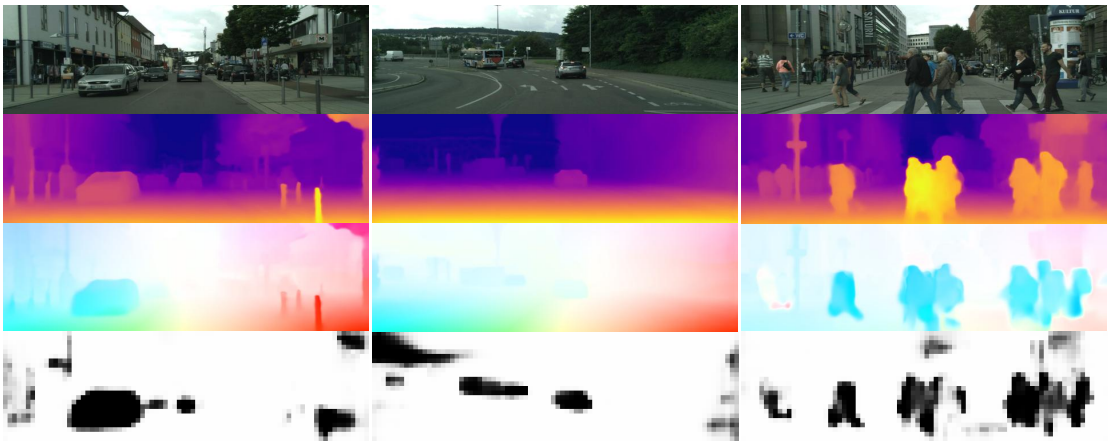


Fig. 4.8 **Generalization test on Cityscapes [26].** From top to bottom: input first frame, estimated depth of first frame, synthesized optical flow, estimated rigidity soft mask.

ego-motion aggregation module in improving the accuracy of visual odometry. We also provide a qualitative comparison of the estimated trajectories from our method, LTMVO [197], and MLF-VO [76] in Fig. 4.7. Our method yields trajectories with overall smaller drifts than the other methods.

4.4.5 Generalization Ability

We use the Cityscapes dataset [26] to test the generalization ability of our model trained on the KITTI dataset [48]. Several visual samples are provided in Fig. 4.8. Our method remarkably generalizes to unseen data, including some significantly dynamic scenes which are rarely present in the training data, such as the presence of numerous pedestrians crossing before the vehicle.

4.4.6 Visualization of Predictions

In Fig. 4.9, we provide visualizations of the predictions obtained by our method. We visualize the estimated SE3 motion field $T_{1\to2}$ as the translation field $\tau_{1\to2} \in \mathbb{R}^{H\times W\times 3}$ and rotation field $\phi_{1\to2} \in \mathbb{R}^{H\times W\times 3}$, where $(\tau_{1\to2}, \phi_{1\to2}) = \mathrm{Log}(T_{1\to2})$. We normalize the values in the translation field and rotation field into the range $[0, 1]$ as a color image. We can observe that our method is capable of estimating a constant SE3 motion for pixels in static regions. We attribute this to the effective exploitation of ego-motion rigidity in our method.

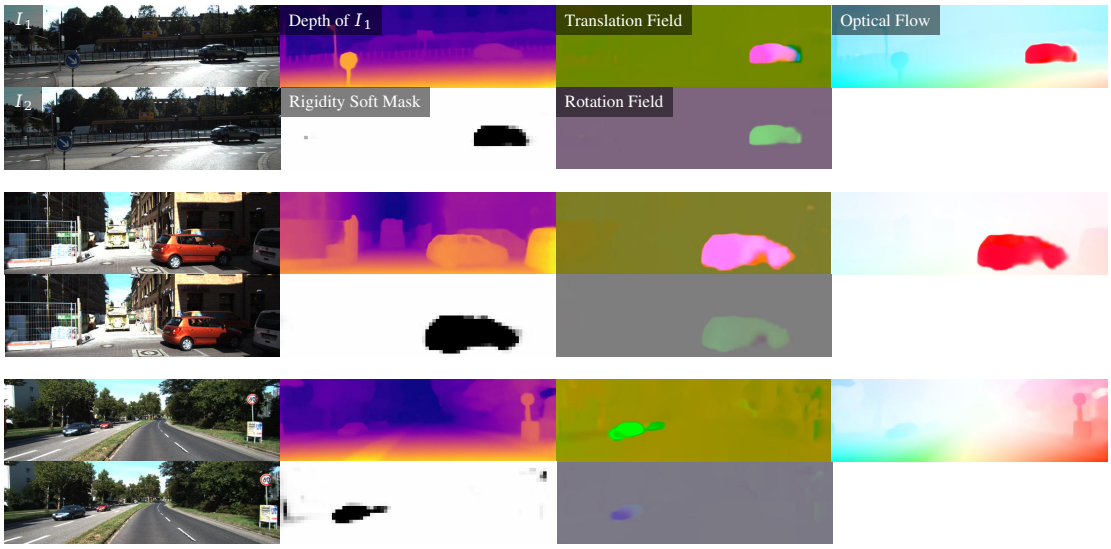


Fig. 4.9 **Visualization of predictions by our method on the KITTI Scene Flow Testing set.** We visualize the estimated SE3 motion field as the translation field and the rotation field. Pixels with the same color have the same translation/rotation.

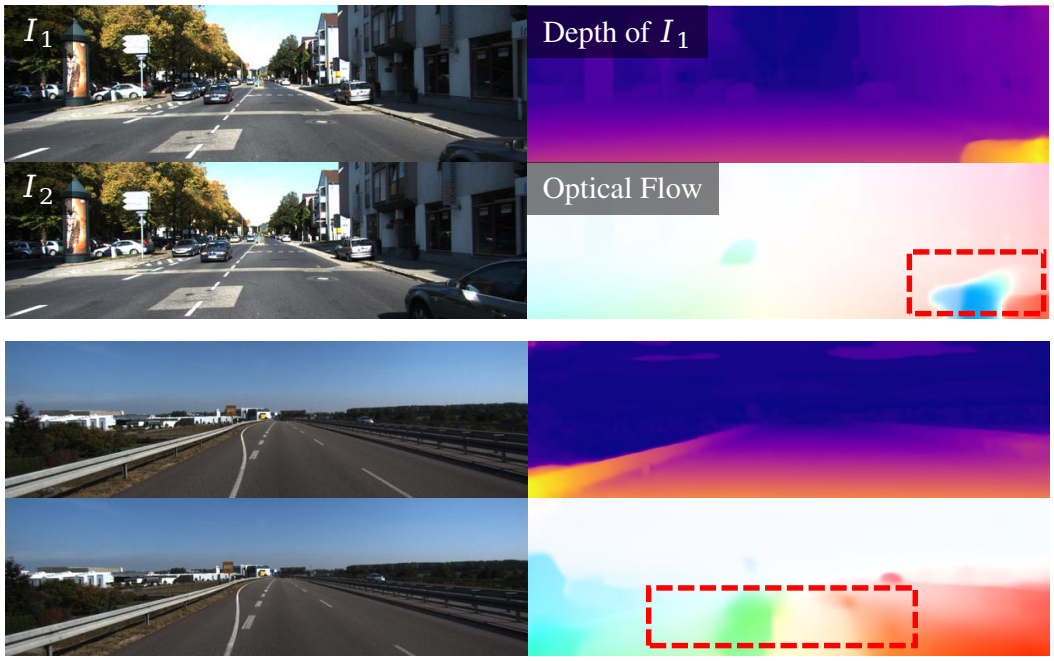


Fig. 4.10 **Failure cases of our method.** The erroneous estimations are highlighted in red boxes.

4.4.7 Faliure Cases

Fig. 4.10 shows some failure cases of our method. Significant estimation errors may still occur in our method for moving objects at the edges of images or textureless regions accompanied by significant motion. Improving the accuracy of estimates in these situations could be a future work for us.

4.4.8 Additional Qualitative Comparisons

We provide additional qualitative comparison results of scene flow estimation in Fig. 4.11 and Fig. 4.12.

4.4.9 Additional Generalization Examples

In Fig. 4.13, we present additional generalization results of our model originally trained on the KITTI [48] dataset, to the Cityscapes [26] dataset. Moreover, we compare the visual results of our model with those of the model trained on the same data from Self-Mono-SF [69]. Our model exhibits superior generalization capabilities, particularly in static regions such as planar roads and walls.

4.5 Conclusion

In this paper, we have proposed a novel self-supervised monocular method named EMR-MSF for scene flow estimation. Our methods incorporates a 3D geometry-oriented network architecture with novel designs to exploit ego-motion rigidity, which results in well-regularized scene flow estimations from solely monocular images. Our proposed approach demonstrates promising potential for monocular dynamic 3D perception and is capable of various computer tasks including scene flow, optical flow, depth, and ego-motion estimation.

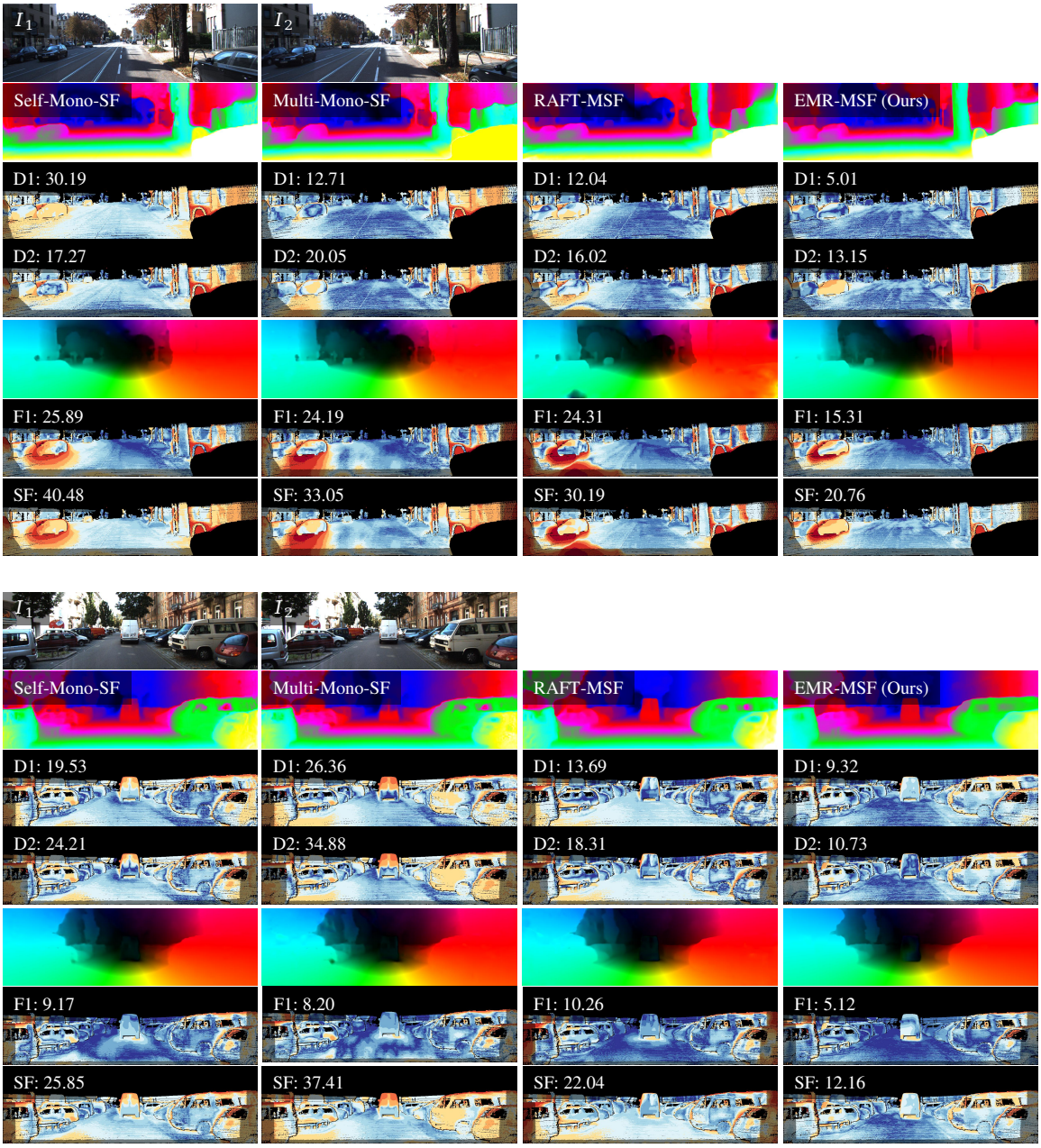


Fig. 4.11 **Qualitative evaluation on KITTI Scene Flow Testing set (1).** We compare our method with Self-Mono-SF [69], Multi-Mono-SF [70] and RAFT-MSF [12] for two scenes using the visualizations provided by the KITTI benchmark [112]. From top to bottom: input images, disparity visualization of I_t, $D1$ error plot, $D2$ error plot, optical flow visualization, corresponding $F1$ error plot and combined SF error plot. The outlier rates are shown on each error plot.

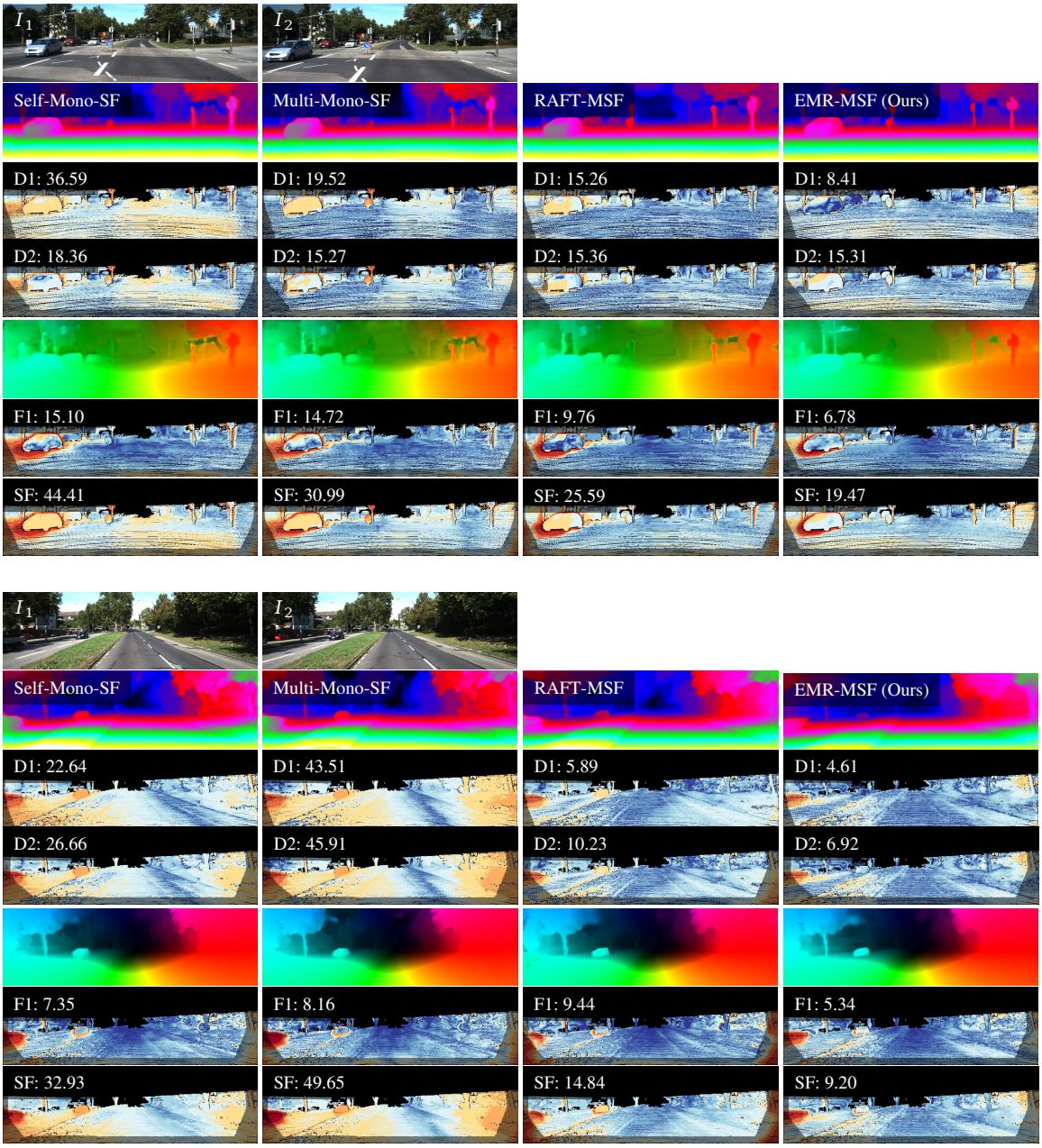


Fig. 4.12 **Qualitative evaluation on KITTI Scene Flow Testing set (2).** We compare our method with Self-Mono-SF [69], Multi-Mono-SF [70] and RAFT-MSF [12] for two scenes using the visualizations provided by the KITTI benchmark [112]. From top to bottom: input images, disparity visualization of I_t, $D1$ error plot, $D2$ error plot, optical flow visualization, corresponding $F1$ error plot and combined SF error plot. The outlier rates on shown on each error plot.

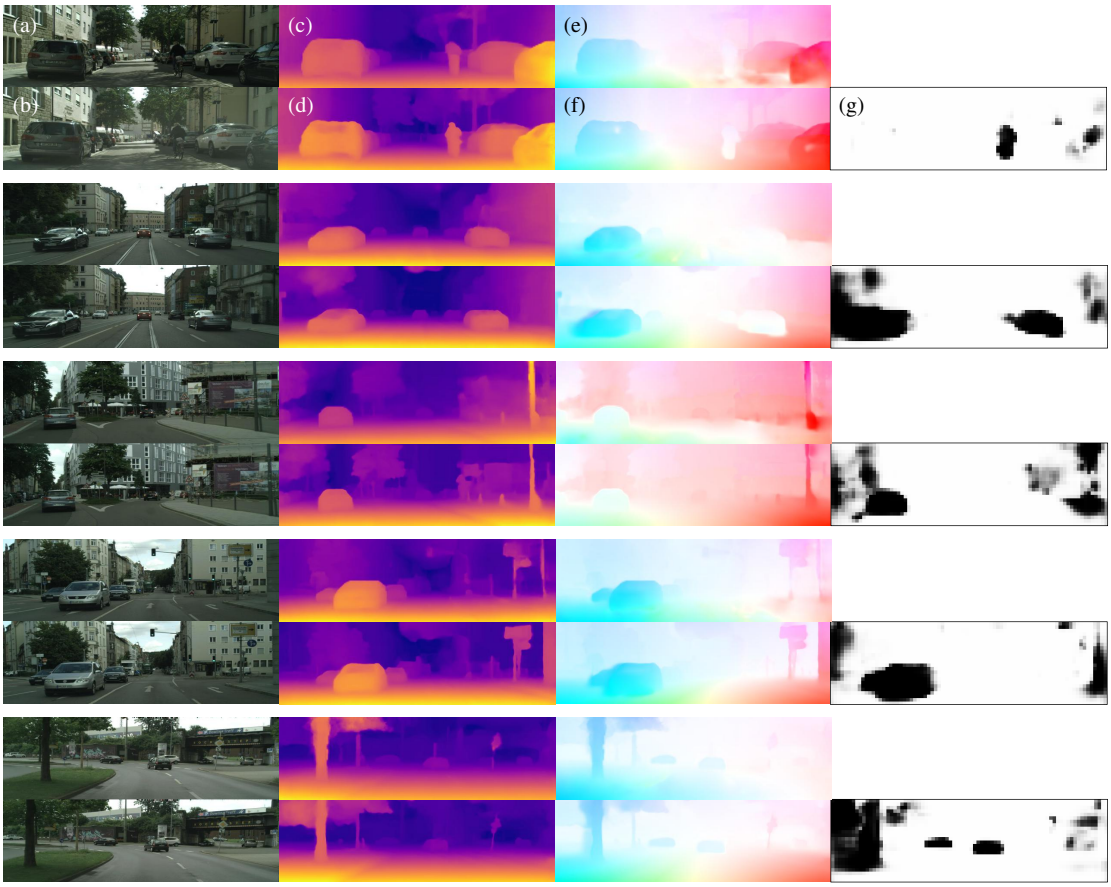


Fig. 4.13 **Comparison of generalization ability between our method and [69] on Cityscapes dataset [26].** (a) input first frame, (b) input second frame, (c) predicted depth of the first frame by [69], (d) predicted depth of the first frame by our method, (e) synthesized optical flow by [69], (f) synthesized optical flow by our method, (g) predicted rigidity soft mask by our method. Our method shows a better generalization ability than [69], especially for the predictions in static regions.

Chapter 5

Geometry-aided Neural Radiance Fields for Novel View Synthesis in Monocular Gastroscopy

5.1 Introduction

Gastroscopy plays a crucial role in minimally invasive diagnostic applications. It captures rich 2D RGB information within the patient's gastric cavity, which offers the practitioners valuable assistance in clinical diagnosis and intervention for various pathological conditions. However, a notable limitation in gastroscopic examinations lies in the constrained viewpoints that practitioners have inside the stomach, which are determined by the trajectory of the gastric endoscope. This typically impedes the practitioners from obtaining adjustable and comprehensive observations within the gastric cavity.

A common solution to enable generating free-viewpoint observations, *i.e.*, novel view synthesis, within the stomach involves reconstructing a 3D representation of the stomach based on the pre-captured gastroscopic images. Structure-from-motion (SfM) [138, 124, 136] is a general technique used to recover camera poses and generate a 3D point cloud representation of the captured scene from image collections, which has been widely applied in several studies [115, 144, 106, 46, 2, 173] to reconstruct the 3D model of a target organ from an endoscope video. [173] successfully reconstructed camera poses and the 3D model of the entire stomach from a standard monocular gastroscopic video by investigating the combined effect of chromo-endoscopy and color channel selection on SfM. For the enhanced visualization of the reconstructed

3D model, they further utilized Poisson surface reconstruction [83] to generate textured meshes from the 3D point cloud acquired through SfM. Although novel view synthesis can be achieved from their reconstructed textured 3D meshes, the 3D model typically exhibits noise and incompleteness due to the existence of low-texture and non-Lambertian regions in gastroscopic images, which consequently leads to low-quality image synthesis.

Recently, neural radiance fields (NeRF) [114, 8] have shown significant progress in the tasks of novel view synthesis and 3D reconstruction from posed images. In contrast to traditional 3D reconstruction methods producing explicit and discrete representations such as point clouds and meshes, NeRF learns the implicit and continuous representation of the scene appearance and geometry, which is encoded in the parameters of multi-layer perceptrons (MLP). The MLP network takes a 3D point position and a 2D camera viewing direction as inputs, predicting the corresponding RGB color and density information. The observation of the 3D scene from an arbitrary viewpoint can finally be obtained through the integration of color and density information along cast camera rays using volume rendering [80]. The emerging NeRF technique has been applied to diverse medical domains, such as 3D reconstruction of deformable tissues from single-viewpoint stereo endoscopy [167, 182] or monocular endoscopy [10], computed tomography [132, 27] and magnetic resonance imaging [176, 137], while it is still not fully explored and evaluated in the context of novel view synthesis based on monocular gastroscopy.

In this paper, we primarily explore the application of neural radiance fields to monocular gastroscopic data, with the aim of achieving high-quality results in novel view synthesis. We observe that directly applying the state-of-the-art general NeRF method [8] trained with color-based losses to monocular gastroscopic data results in broken geometry and blurry image rendering. One main reason for the performance degradation is attributed to the view sparsity in local regions present in gastroscopic data, which results in the insufficiency of color-based loss to address the shape-radiance ambiguity during the training of neural radiance fields. To enhance the training results on monocular gastroscopic data, we incorporate the color-based loss with the additional geometry-based supervision signals exploited from the point clouds reconstructed by SfM, sharing a similar idea with DS-NeRF [31]. Notably, our proposed geometry-based supervision differs from the one in DS-NeRF [31] mainly in two aspects: 1) In addition to utilizing the sparse depth maps obtained from point clouds for supervision, we also incorporate a depth smoothness loss based on the shape priors of the stomach. 2) DS-NeRF [31] imposes depth supervision solely to

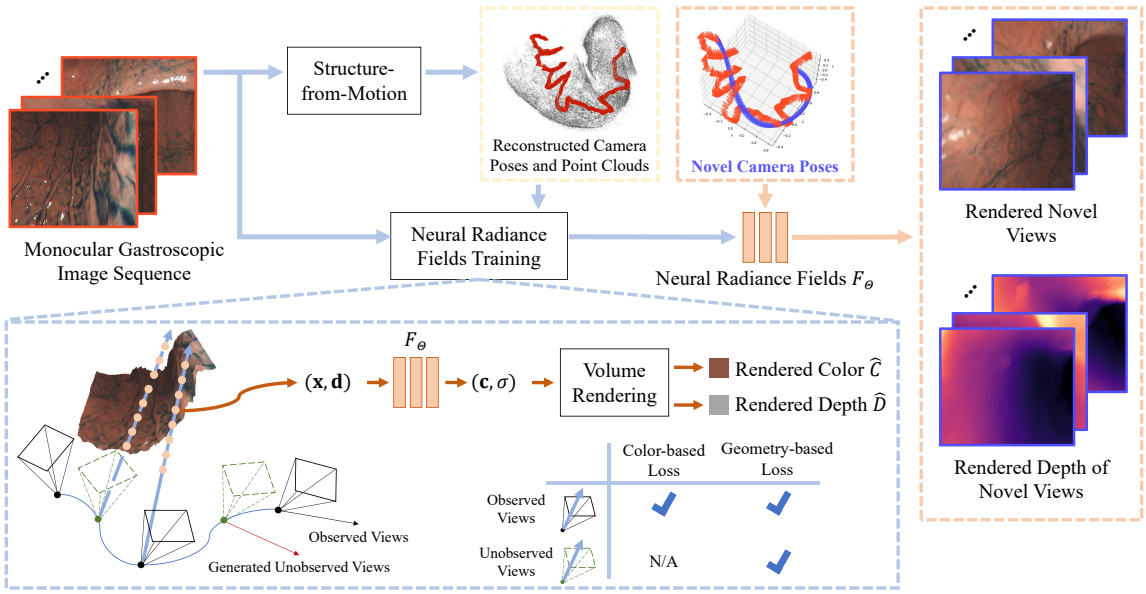


Fig. 5.1 The pipeline of our proposed method. Our method aims to recover the camera poses and learn the neural radiance fields representation F_Θ of the stomach from a monocular gastroscopic image sequence. The key insight in our training of neural radiance fields is to apply the geometry-based loss not only on the observed views but also on the generated unobserved views, which effectively exploits the geometry obtained from the reconstructed point clouds to constrain the learned geometry of neural radiance fields. As an application of the learned neural radiance fields F_Θ, both RGB observations and depth perceptions from new perspectives within the stomach can be generated through volume rendering [80].

pre-captured observed views, whereas our method imposes depth supervision to both pre-captured observed views and unobserved views randomly interpolated from observed views, which better regularizes the learned geometry of neural radiance fields. Our experimental results showcase the efficacy of our proposed geometry-based supervision in enhancing both rendering quality and recovered geometry compared to the baseline method [8].

5.2 Methodology

The pipeline of our proposed method is illustrated in Fig. 5.1. Given a monocular gastroscopic image sequence, our method aims to recover the camera poses and learn the neural radiance fields representation of the stomach. We initially follow the structure-from-motion steps proposed in [173] to reconstruct the camera poses and 3D point clouds of the stomach. Random pose interpolations between the recovered

consecutive camera poses are performed subsequently to generate unobserved views (*c.f.*, Sec.5.2.1). During the following training of neural radiance fields, we conduct separate ray sampling on the observed views and the unobserved views, and employ volume rendering to obtain the rendered color and depth of each cast ray (*c.f.*, Sec.5.2.2). At the end of processing each training iteration, both the color-based loss and the improved geometry-based loss are computed to achieve optimal training results (*c.f.*, Sec.5.2.3).

5.2.1 Unobserved View Interpolation

After recovering the camera poses of all observed views through SfM, we generate k unobserved views between each consecutive observed view pair. We parameterize the camera pose T by $T = (t, \hat{q})$, where $t \in \mathbb{R}^3$ is the 3D position and $\hat{q}$ is the unit quaternion representing rotation. Given a pair of observed views with camera poses T_1 and T_2, we interpolate the 3D position and rotation separately to generate the camera pose $T_u = (t_u, \hat{q}_u)$ of the unobserved view by

$$\begin{aligned} t_u &= (1-\alpha)t_1 + \alpha t_2 \\ q_u &= q_1(q_1^{-1}q_2)^{\alpha}, \end{aligned} \tag{5.1}$$

where α is randomly sampled from 0 to 1. To maximize the utilization of additional geometric constraints brought by unobserved views (*c.f.*, Sec.5.2.3), we regenerate new unobserved views every 2000 training iterations during our experiments.

5.2.2 Ray Sampling and Volume Rendering

In each iteration of NeRF training, we sample two types of camera rays, denoted as the color ray and depth ray. The color ray passes through the camera origin and the pixels on the image plane, with its ground-truth color value being the color of the corresponding pixel. The depth ray passes through the camera origin and the reconstructed 3D points visible in the camera, with its reference depth value being the transformed depth of the corresponding 3D point in the camera coordinate. We sample both color rays and depth rays on the observed views, denoted as $\mathcal{R}_c^{ob}$ and $\mathcal{R}_d^{ob}$ separately, and only sample depth rays on the unobserved views, denoted as $\mathcal{R}_d^{nv}$.

Given a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ from the camera origin $\mathbf{o}$ with the viewing direction $\mathbf{d}$, we compute the color of this camera ray using volume rendering as

described in [114]:

$$\hat{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\mathrm{d}t, \tag{5.2}$$

where $T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))\mathrm{d}s$, t_n and t_f are the near and far bounds of the traveled distance t. The density σ and color $\mathbf{c}$ are predicted by the MLP network $F_\Theta : (\mathbf{r}(t), \mathbf{d}) \rightarrow (\sigma, \mathbf{c})$. Similarly, the depth of this camera ray can be rendered as follows:

$$\hat{D}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))t\mathrm{d}t. \tag{5.3}$$

For the computing of training loss, we render both the color and depth for camera rays in $\mathcal{R}_c$ and only render the depth for camera rays in $\mathcal{R}_d^{ob} \cup \mathcal{R}_d^{nv}$.

5.2.3 Training Loss

Given sampled sets of camera rays $\mathcal{R}_c^{ob}$, $\mathcal{R}_d^{ob}$ and $\mathcal{R}_d^{nv}$, and their rendered results, the training loss of our method is computed as follows:

Color-based loss.

The color-based loss is computed from camera rays in $\mathcal{R}_c^{ob}$ and is formulated as:

$$L_{color}^{ob} = \sum_{\mathbf{r}\in\mathcal{R}_c^{ob}} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|_2^2, \tag{5.4}$$

where $\hat{C}(\mathbf{r})$ is the rendered color of the camera ray in $\mathcal{R}_c^{ob}$ and $C(\mathbf{r})$ is the corresponding ground-truth color.

Geometry-based loss for observed views.

For each camera ray $\mathbf{r}$ in $\mathcal{R}_d^{ob}$, we first compute the difference between the rendered depth $\hat{D}(\mathbf{r})$ and its reference depth value $D_{pc}(\mathbf{r})$ obtained from the reconstructed 3D point points:

$$l_d(\mathbf{r}) = \|\hat{D}(\mathbf{r}) - D_{pc}(\mathbf{r})\|_2^2. \tag{5.5}$$

We further introduce the smoothness loss to enforce local smoothness in the rendered depth of camera rays either in $\mathcal{R}_c^{ob}$ or $\mathcal{R}_d^{ob}$, considering that the internal structure of

the stomach is typically smooth and continuous:

$$l_s(\mathbf{r}) = |\partial_u \hat{D}(\mathbf{r})| + |\partial_v \hat{D}(\mathbf{r})|, \tag{5.6}$$

where (u, v) is the corresponding pixel coordinate of the camera ray $\mathbf{r}$ on the image plane. We also combine the KL divergence loss $l_{KL}(\mathbf{r})$ introduced in [31] to constrain the ray distribution to be unimodal. Please refer to [31] for further details on $l_{KL}(\mathbf{r})$. Our final geometry-based loss for the observed views is formulated as:

$$\begin{aligned} L^{ob}_{depth} = \sum_{\mathbf{d}\in\mathcal{R}^{ob}_d} \lambda_d \|\hat{D}(\mathbf{r}) - D_{pc}(\mathbf{r})\|^2_2 + \lambda_{KL} l_{KL}(\mathbf{r}) \\ + \sum_{\mathbf{d}\in\mathcal{R}^{ob}_c \cup \mathcal{R}^{ob}_d} \lambda_s l_s(\mathbf{r}), \end{aligned} \tag{5.7}$$

where λ_d, λ_{KL} and λ_s are hyperparameters.

Geometry-based loss for unobserved views.

Applying geometry-based loss solely on the observed views may sometimes be ineffective in constraining the learned 3D geometry of neural radiance fields due to the sparse depth supervision signals obtained from the reconstructed 3D point clouds. Thus, we propose to integrate additional geometry-based supervision on generated unobserved views to further regularize the learned geometry, which is formulated as:

$$\begin{aligned} L^{nv}_{depth} = \sum_{\mathbf{d}\in\mathcal{R}^{ob}_d} \lambda_d \|\hat{D}(\mathbf{r}) - D_{pc}(\mathbf{r})\|^2_2 \\ + \lambda_{KL} l_{KL}(\mathbf{r}) + \lambda_s I_s(\mathbf{r}). \end{aligned} \tag{5.8}$$

The final training loss of our method is concluded as: $L_{total} = L^{ob}_{color} + L^{ob}_{depth} + L^{nv}_{depth}$.

5.3 Results

5.3.1 Datasets and Implementation Details

In our experiments, we evaluate our method using two calibrated monocular gastroscopic videos obtained from [173], identified as Seq. A and Seq. B. For both sequences, we reduce the frame rate to one-fourth and then reserve every second

frame for the quantitative evaluation of novel view synthesis, with the remaining frames utilized for the training of neural radiance fields.

We implement our method building upon Zip-NeRF [8], by incorporating the novel geometry-based supervision introduced in the methodology section. During training, We regenerate $k = 2$ unobserved views every 2000 training iterations. We sample 8192 camera rays for $\mathcal{R}_c$, and 4096 camera rays for both $\mathcal{R}_d$ and $\mathcal{R}_d^{nv}$. The hyperparameters in training loss are set as: $\lambda_d = 10, \lambda_{KL} = 0.1, \lambda_s = 10$. The number of training iterations is set to 25,000 and other hyperparameters are set the same as Zip-NeRF [8]. All experiments are performed on a single Nvidia RTX-3090 GPU.

5.3.2 Results of Novel View Synthesis

Table 5.1 Quantitative evaluation of novel view synthesis. The best results are highlighted using **bold** formatting.

	Seq. A		Seq. B	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DS-NeRF [31]	22.49	0.838	20.35	0.694
Zip-NeRF [8]	25.07	0.856	22.78	0.751
Ours w/o L_{depth}^{nv}	25.47	0.861	23.17	0.762
Ours	**26.73**	**0.870**	**23.37**	**0.767**

Tab. 5.1 presents a quantitative comparison between our proposed method and other approaches. Across both sequences, our approach attains consistent superior results in synthesizing high-quality views from novel perspectives. The improvement relative to Zip-NeRF [8] shows the positive impact of the proposed geometry-based loss on the results of neural radiance fields trained for monocular gastroendoscopy. Incorporating L_{depth}^{nv} into our approach yields further performance gains, demonstrating that applying additional geometry-based constraints to the unobserved views can further enhance the learned geometry of neural radiance fields, consequently contributing to better image rendering results. Fig. 5.2 provides qualitative results of the novel view synthesis presented by our method and other approaches. Our full method produces more photorealistic synthesized images from novel viewpoints compared to other approaches, including more image details such as sharp edges between foreground and background, and realistic specular regions.

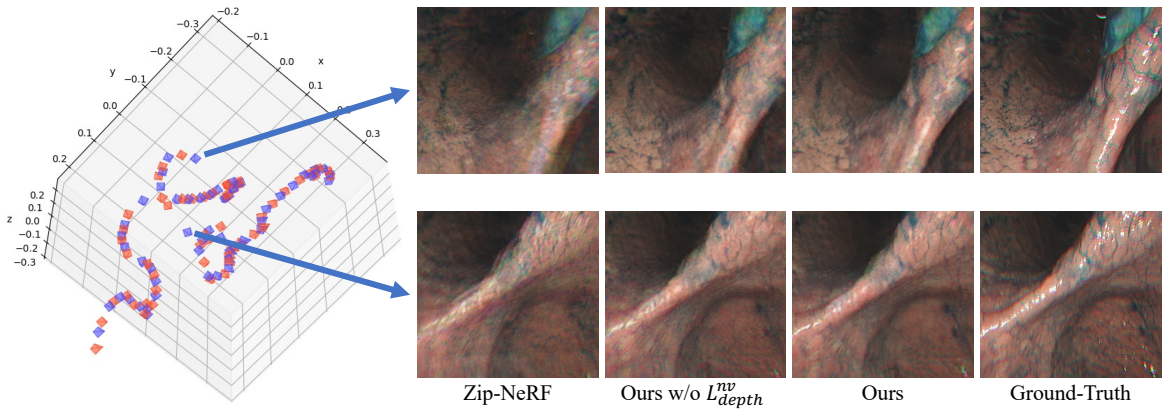


Fig. 5.2 The qualitative results of novel view synthesis. The left side presents the camera poses of images used for training (marked in red) and testing (marked in blue) in Seq.B. The right side showcases the image rendering results of different methods given the camera poses of two selected testing images. Our full method produces high-fidelity image renderings with more details compared to other methods.

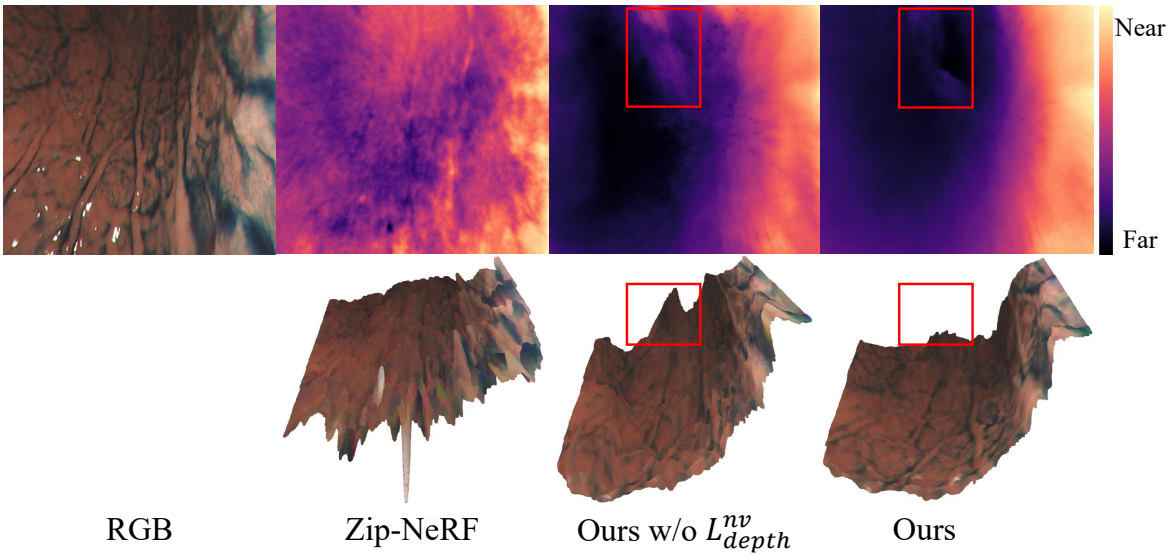


Fig. 5.3 The qualitative results of learned geometry. For each method, we present both the rendered depth (first row) and the corresponding unprojected point clouds (second row) for better visualization. We highlight the differences in red boxes.

5.3.3 Results of Learned Geometry

We further evaluate the learned geometry across different methods by comparing the rendered depth maps. Since the ground-truth depth maps are not available, only the qualitative evaluation of the learned geometry is provided as shown in Fig. 5.3. For better visualization, both the rendered depth maps and their corresponding unprojected 3D point clouds are presented. The depth map produced

by Zip-NeRF [8] exhibits significant errors, which shows that it is challenging for general neural radiance field methods to accurately learn geometry from monocular gastroendoscopy with sole RGB color supervision. 'Ours w/o L_{depth}^{nv}', which incorporates geometry-based supervision only on the observed training views, produces a more plausible depth map rendering. However, artifacts are still present, as observed in the highlighted region within the red box in Fig. 5.3. Our full method produces the visually best results, by incorporating geometry-based supervision on both the observed and unobserved training views.

5.4 Conclusion

In this paper, we employ the technique of neural radiance fields for the task of synthesizing free-viewpoint views within the stomach of patients. To enhance the performance of the neural radiance fields approach for high-quality novel view synthesis on monocular gastroscopic data, we augment the original color-based training loss with a geometry-based loss. This augmentation enables effective utilization of the point clouds reconstructed from images to constrain the implicit geometry of the neural radiance fields. Additionally, we propose to apply the geometry-based supervision to randomly generated unobserved views during the training phase, which further regularizes the learned geometry within the neural radiance field, and contributes to performance enhancements in novel view synthesis. The experimental results demonstrate that our approach is capable of both high-quality novel view synthesis and geometry recovery based on monocular gastroscopy data. One limitation of our approach lies in its dependence on the precise estimation of camera poses from the SfM, which is not guaranteed in the context of gastroscopic data. In the future, we plan to incorporate the refinement of camera poses into the training of neural radiance fields to alleviate the reliance on accurately pre-computed camera poses.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

Over the past decade, significant progress has been made in methods related to recovering camera motion and scene structure from monocular image sequences. From the early stages of geometry-based traditional approaches towards maturity to the middle and later stages where learning-based data-driven methods gradually demonstrated potential, and more recently, the emergent approaches based on neural implicit representation, these advancements have laid a solid theoretical foundation for the future development of this field. Despite these achievements, practical applications still pose a series of challenges. This paper aims to address these challenges and propose effective solutions, encompassing the following aspects:

In Chapter 2, we introduced a 3D reconstruction pipeline based on Structure-from-Motion that efficiently utilizes camera pose information obtained from Visual-Inertial Odometry. The camera pose information is seamlessly integrated into each step of the reconstruction process to address challenges posed by diverse image conditions. Extensive experimental results show that our pipeline performs better than the state-of-the-art SfM approaches in terms of reconstruction accuracy and robustness for challenging sequential image collections.

In Chapter 3, We evaluate different approaches for integrating RGB and inferred depth images in the context of self-supervised ego-motion estimation, which reveals key insights into the effective fusion of information from diverse modalities. We systematically incorporated these findings into our pipeline, constructing a relative pose estimator that integrates modalities at multiple stages of the feature encoder. The resulting system achieves state-of-the-art performance among self-supervised ego-motion estimation methods.

In Chapter 4, We introduced EMR-MSF, a novel self-supervised monocular method tailored for joint estimation of camera motion, depth and scene flow. Our approach integrates a 3D geometry-oriented network architecture with innovative designs aimed at leveraging ego-motion rigidity, leading to well-regularized scene flow estimations from monocular images alone. Our proposed approach proves effective in multiple computer vision tasks, encompassing scene flow, optical flow, depth, and ego-motion estimation, demonstrating the promising potential for monocular dynamic 3D perception.

In Chapter 5, We utilize neural radiance fields trained from monocular gastroscopic images to generate free-viewpoint perspectives within the stomach of a patient. To improve training results on monocular gastroscopic data, we integrate color-based loss along with additional geometry-based supervision signals derived from the point clouds reconstructed through SfM. Our experimental results showcase the efficacy of our proposed geometry-based supervision in enhancing both rendering quality and recovered geometry compared to the baseline method.

6.2 Future works

Although we have already discussed some improvements in the studies of recovering camera motion and 3D structures from sequential monocular images, there are still problems to be addressed. In this section, we discuss possible future works as below:

Extension of EMR-MSF to indoor dynamic datasets

The proposed EMR-MSF has demonstrated promising results on outdoor autonomous driving datasets. However, its performance on indoor dynamic datasets, which may involve more intricate and dynamic motions, remains a research topic that requires further exploration.

Utilization of neural implicit representation in jointly recovering camera pose and 3D structures for non-Lambertian scenes

Current geometry-based traditional methods and self-supervised data-driven methods rely on the assumption of multi-view consistency, where the scene surface is considered Lambertian, and its color does not change with varying viewpoints. However, real-world scenes often include non-Lambertian surfaces, posing a challenge for existing methods. Methods based on neural implicit representation, with the ability to model view-dependent color, hold the potential for camera pose estimation and non-Lambertian scene recovery in more challenging scenarios.

References

[1] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. http://ceres-solver.org.

[2] Pablo F Alcantarilla, Adrien Bartoli, François Chadebecq, Christophe Tilmant, and Vincent Lepilliez. Enhanced imaging colonoscopy facilitates dense motion-based 3d reconstruction. In *Proc. EMBC*, 2013.

[3] Hadi AliAkbarpour, Kannappan Palaniappan, and Guna Seetharaman. Fast Structure from Motion for Sequential and Wide Area Motion Imagery. In *Proc. ICCV Workshop*, 2015.

[4] Yasin Almalioglu, Muhamad Risqi U Saputra, Pedro PB de Gusmao, Andrew Markham, and Niki Trigoni. GANVO: Unsupervised Deep Monocular Visual Odometry and Depth Estimation with Generative Adversarial Networks. In *Proc. Intl. Conf. on Robotics and Automation*, 2019.

[5] Rares Ambrus, Vitor Guizilini, Jie Li, and Sudeep Pillai Adrien Gaidon. Two Stream Networks for Self-Supervised Ego-Motion Estimation. In *Proc. CoRL*, 2020.

[6] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[7] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE PAMI*, 41(2):423–443, 2018.

[8] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proc. ICCV*, 2023.

[9] Tali Basha, Yael Moses, and Nahum Kiryati. Multi-view scene flow estimation: A view centered variational approach. *IJCV*, 101:6–21, 2013.

[10] Víctor M Batlle, José MM Montiel, Pascal Fua, and Juan D Tardós. Lightneus: Neural surface reconstruction in endoscopy using illumination decline. In *Proc. MICCAI*, 2023.

[11] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Proc. ECCV*, 2006.

[12] Bayram Bayramli, Junhwa Hur, and Hongtao Lu. Raft-msf: Self-supervised monocular scene flow using recurrent optimizer. *arXiv preprint arXiv:2205.01568*, 2022.

[13] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *Proc. CVPR*, 2017.

[14] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[15] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video. *Proc. NeurIPS*, 2019.

[16] Michael Bloesch, Michael Burri, Sammy Omari, Marco Hutter, and Roland Siegwart. Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback. *Intl. J. of Robotics Research*, 36(10):1053–1072, 2017.

[17] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[18] Fabian Brickwedde, Steffen Abraham, and Rudolf Mester. Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. In *Proc. ICCV*, 2019.

[19] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal Distributional Semantics. *Journal of artificial intelligence research*, 49:1–47, 2014.

[20] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W. Achtelik, and Roland Siegwart. The EuRoC micro aerial vehicle datasets. *Intl. J. of Robotics Research*, 35(10):1157–1163, 2016.

[21] Zhe Cao, Abhishek Kar, Christian Hane, and Jitendra Malik. Learning independent object motion from unlabelled stereoscopic videos. In *Proc. CVPR*, 2019.

[22] Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni. Selective Sensor Fusion for Neural Visual-Inertial Odometry. In *Proc. CVPR*, 2019.

[23] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. CVPR*, 2019.

[24] Wencan Cheng and Jong Hwan Ko. Bi-pointflownet: Bidirectional learning for point cloud based scene flow estimation. In *Proc. ECCV*, 2022.

[25] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proc. CVPR*, 2020.

[26] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. CVPR*, 2016.

[27] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert PH Shum, and Chris G Willcocks. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In *Proc. EMBC*, 2022.

[28] Hainan Cui, Shuhan Shen, Wei Gao, and Zhanyi Hu. Efficient large-scale structure from motion by fusing auxiliary imaging information. *IEEE Transactions on Image Processing*, 24(11):3561–3573, 2015.

[29] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. MonoSLAM: Real-Time Single Camera SLAM. *IEEE PAMI*, 29(6):1052–1067, 2007.

[30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[31] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proc. CVPR*, 2022.

[32] Ayush Dewan, Tim Caselitz, Gian Diego Tipaldi, and Wolfram Burgard. Rigid scene flow for 3d lidar scans. In *Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems*, 2016.

[33] Lihe Ding, Shaocong Dong, Tingfa Xu, Xinli Xu, Jie Wang, and Jianan Li. Fh-net: A fast hierarchical network for scene flow estimation on real-world point clouds. In *Proc. ECCV*, 2022.

[34] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. CamNet: Coarse-to-Fine Retrieval for Camera Re-Localization. In *Proc. ICCV*, 2019.

[35] Guanting Dong, Yueyi Zhang, Hanlin Li, Xiaoyan Sun, and Zhiwei Xiong. Exploiting rigidity constraints for lidar scene flow estimation. In *Proc. CVPR*, 2022.

[36] Jingming Dong and Stefano Soatto. Domain-size pooling in local descriptors: DSP-SIFT. In *Proc. CVPR*, 2015.

[37] Tue-Cuong Dong-Si and Anastasios I. Mourikis. Consistency analysis for sliding-window visual odometry. In *Proc. Intl. Conf. on Robotics and Automation*, 2012.

[38] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Proc. NeurIPS*, 2014.

[39] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-Scale Direct monocular SLAM. In *Proc. ECCV*, 2014.

[40] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct Sparse Odometry. *IEEE PAMI*, 40(3):611–625, 2017.

[41] Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Comm. ACM*, 24(6):381–395, 1981.

[42] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems. *IEEE Transactions on Robotics*, 33(2):249–265, 2016.

[43] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building Rome on a Cloudless Day. In *Proc. ECCV*, 2010.

[44] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proc. CVPR*, 2018.

[45] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Proc. NeurIPS*, 2022.

[46] Ryo Furukawa, Hiroki Morinaga, Yoji Sanomura, Shinji Tanaka, Shigeto Yoshida, and Hiroshi Kawasaki. Shape acquisition and registration for 3d endoscope based on grid pattern projection. In *Proc. ECCV*, 2016.

[47] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proc. CVPR*, 2012.

[48] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Intl. J. of Robotics Research*, 32(11):1231–1237, 2013.

[49] Patrick Geneva, James Maley, and Guoquan Huang. An Efficient Schmidt-EKF for 3D Visual-Inertial SLAM. In *Proc. CVPR*, 2019.

[50] Ross Girshick. Fast r-cnn. In *Proc. ICCV*, 2015.

[51] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. Multiple Classifier Systems for the Classification of Audio-Visual Emotional States. In *International Conference on Affective Computing and Intelligent Interaction*, 2011.

[52] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *Proc. CVPR*, 2017.

[53] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *Proc. ICCV*, 2019.

[54] Juan Luis Gonzalez and Munchurl Kim. Plade-net: towards pixel-level accuracy for self-supervised single-view depth estimation with neural positional encoding and distilled matting loss. In *Proc. CVPR*, 2021.

[55] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *Proc. NeurIPS*, 2020.

[56] Michael Grupp. evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo, 2017.

[57] Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, and Panqu Wang. Hplflownet: Hierarchical permutohedral lattice flownet for scene flow estimation on large-scale point clouds. In *Proc. CVPR*, 2019.

[58] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proc. CVPR*, 2020.

[59] Liming Han, Yimin Lin, Guoguang Du, and Shiguo Lian. DeepVIO: Self-supervised Deep Learning of Monocular Visual Inertial Odometry using 3D Geometric Constraints. In *Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems*, 2019.

[60] Richard Hartley and Peter Sturm. Triangulation. *CVIU*, 68(2):146–157, 1997.

[61] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[62] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture. In *Proc. ACCV*, 2016.

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. CVPR*, 2016.

[64] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. ICCV*, 2017.

[65] Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie, and Jian Yang. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *Proc. ECCV*, 2022.

[66] Joel A. Hesch and Stergios I. Roumeliotis. A Direct Least-Squares (DLS) Method for PnP. In *Proc. ICCV*, 2011.

[67] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-Based Multimodal Fusion for Video Description. In *Proc. ICCV*, 2017.

[68] Frédéric Huguet and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. In *Proc. ICCV*, 2007.

[69] Junhwa Hur and Stefan Roth. Self-supervised monocular scene flow estimation. In *Proc. CVPR*, 2020.

[70] Junhwa Hur and Stefan Roth. Self-supervised multi-frame monocular scene flow. In *Proc. CVPR*, 2021.

[71] Arnold Irschara, Christof Hoppe, Horst Bischof, and Stefan Kluckner. Efficient structure from motion with weak position and orientation priors. In *Proc. CVPR Workshop*, 2011.

[72] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial Transformer Networks. *Proc. NeurIPS*, 2015.

[73] Mariano Jaimez, Mohamed Souiai, Jörg Stückler, Javier Gonzalez-Jimenez, and Daniel Cremers. Motion cooperation: Smooth piece-wise rigid scene flow from rgb-d images. In *Proc. 3DV*, 2015.

[74] Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In *Proc. ICCV*, 2019.

[75] Hualie Jiang, Laiyan Ding, Zhenglong Sun, and Rui Huang. Dipe: Deeper into photometric errors for unsupervised learning of depth and ego-motion from monocular videos. In *Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems*, 2020.

[76] Zijie Jiang, Hajime Taira, Naoyuki Miyashita, and Masatoshi Okutomi. Self-supervised ego-motion estimation based on multi-layer fusion of rgb and inferred depth. In *Proc. Intl. Conf. on Robotics and Automation*, 2022.

[77] Yang Jiao, Trac D Tran, and Guangming Shi. Effiscene: Efficient per-pixel rigidity inference for unsupervised joint learning of optical flow, depth, camera pose and motion segmentation. In *Proc. CVPR*, 2021.

[78] Eagle S. Jones and Stefano Soatto. Visual-Inertial Navigation, Mapping and Localization: A Scalable Real-Time Causal Approach. *Intl. J. of Robotics Research*, 30(4):407–430, 2011.

[79] Sho Kagami, Hajime Taira, Naoyuki Miyashita, Akihiko Torii, and Masatoshi Okutomi. 3D Pipe Network Reconstruction Based on Structure from Motion with Incremental Conic Shape Detection and Cylindrical Constraint. In *Proc. IEEE International Symposium on Industrial Electronics*, 2020.

[80] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984.

[81] Mike Kasper, Steve McGuire, and Christoffer Heckman. A Benchmark for Visual-Inertial Odometry Systems Employing Onboard Illumination. In *Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems*, 2019.

[82] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057*, 2020.

[83] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graphics*, 32(3):1–13, 2013.

[84] Jonathan Kelly and Gaurav S Sukhatme. Visual-Inertial Sensor Fusion: Localization, Mapping and Sensor-to-Sensor Self-Calibration. *Intl. J. of Robotics Research*, 30(1):56–79, 2011.

[85] Kurt Konolige and Motilal Agrawal. FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping. *IEEE Trans. Robotics*, 24(5):1066–1077, 2008.

[86] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Proc. NeurIPS*, 2012.

[87] Pierre-Yves Lajoie, Siyi Hu, Giovanni Beltrame, and Luca Carlone. Modeling Perceptual Aliasing in SLAM via Discrete–Continuous Graphical Models. *IEEE RA-L*, 4(2):1232–1239, 2019.

[88] Zakaria Laskar, Sami Huttunen, Daniel Herrera, Esa Rahtu, and Juho Kannala. Robust Loop Closures for Scene Reconstruction by Combining Odometry and Visual Correspondences. In *Intl. Conf. Image Proc.*, 2016.

[89] Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Annual Meeting of the Association for Computational Linguistics*, 2014.

[90] Vincent Lepetit, Francesc Moreno-Noguer, and P Fua. EPnP: Efficient Perspective-n-Point Camera Pose Estimation. *IJCV*, 81(2):155–166, 2009.

[91] Stefan Leutenegger, Paul Furgale, Vincent Rabaud, Margarita Chli, Kurt Konolige, and Roland Siegwart. Keyframe-Based Visual-Inertial SLAM using Nonlinear Optimization. *Proc. RSS*, 2013.

[92] Bin Li, Mu Hu, Shuling Wang, Lianghao Wang, and Xiaojin Gong. Self-Supervised Visual-LiDAR Odometry With Flip Consistency. In *Proc. WACV*, 2021.

[93] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes. In *Proc. CoRL*, 2021.

[94] Heng Li, Xiaodong Gu, Weihao Yuan, Luwei Yang, Zilong Dong, and Ping Tan. Dense rgb slam with neural implicit maps. In *Proc. ICLR*, 2023.

[95] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. UnDeepVO: Monocular Visual Odometry through Unsupervised Deep Learning. In *Proc. Intl. Conf. on Robotics and Automation*, 2018.

[96] Shunkai Li, Xin Wang, Yingdian Cao, Fei Xue, Zike Yan, and Hongbin Zha. Self-Supervised Deep Visual Odometry with Online Adaptation. In *Proc. CVPR*, 2020.

[97] Yang Li, Yoshitaka Ushiku, and Tatsuya Harada. Pose Graph optimization for Unsupervised Monocular Visual Odometry. In *Proc. Intl. Conf. on Robotics and Automation*, 2019.

[98] Yunpeng Li, Noah Snavely, Daniel P Huttenlocher, and Pascal Fua. Worldwide Pose Estimation using 3D Point Clouds. In *Proc. ECCV*, 2012.

[99] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proc. ICCV*, 2021.

[100] Liang Liu, Guangyao Zhai, Wenlong Ye, and Yong Liu. Unsupervised learning of scene flow estimation fusing with local rigidity. In *Proc. IJCAI*, 2019.

[101] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proc. CVPR*, 2019.

[102] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[103] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004.

[104] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE PAMI*, 42(10):2624–2641, 2019.

[105] Quan-Tuan Luong and Olivier D Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *IJCV*, 17(1):43–75, 1996.

[106] Kristen L Lurie, Roland Angst, Dimitar V Zlatev, Joseph C Liao, and Audrey K Ellerbee Bowden. 3d reconstruction of cystoscopy videos for comprehensive bladder records. *Biomedical optics express*, 8(4):2106–2123, 2017.

[107] Zhaoyang Lv, Kihwan Kim, Alejandro Troccoli, Deqing Sun, James M Rehg, and Jan Kautz. Learning rigidity in dynamic scenes with a moving camera for 3d motion field estimation. In *Proc. ECCV*, 2018.

[108] Wei-Chiu Ma, Shenlong Wang, Rui Hu, Yuwen Xiong, and Raquel Urtasun. Deep rigid instance scene flow. In *Proc. CVPR*, 2019.

[109] Lukas Mehl, Azin Jahedi, Jenny Schmalfuss, and Andrés Bruhn. M-fuse: Multi-frame fusion for scene flow estimation. In *Proc. WACV*, 2023.

[110] Christopher Mei, Gabe Sibley, Mark Cummins, Paul Newman, and Ian Reid. A Constant-Time Efficient Stereo SLAM system. In *Proc. BMVC.*, 2009.

[111] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proc. AAAI*, 2018.

[112] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. CVPR*, 2015.

[113] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. CVPR*, 2019.

[114] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020.

[115] Steven Mills, Lech Szymanski, and Reuben Johnson. Hierarchical structure from motion from endoscopic video. In *Proc. of Int. Conf. on Image and Vision Computing New Zealand (IVCNZ)*, 2014.

[116] Anastasios I. Mourikis and Stergios I. Roumeliotis. A Multi-State Constraint Kalman Filter for Vision-Aided Inertial Navigation. In *Proc. Intl. Conf. on Robotics and Automation*, 2007.

[117] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on robotics*, 33(5):1255–1262, 2017.

[118] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on robotics*, 31(5):1147–1163, 2015.

[119] Seyed Shahabeddin Nabavi, Mehrdad Hosseinzadeh, Ramin Fahimi, and Yang Wang. Unsupervised Learning of Camera Pose with Compositional Re-estimation. In *Proc. WACV*, 2020.

[120] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. CVPR*, 2019.

[121] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch, 2017.

[122] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Proc. NeurIPS*, 2019.

[123] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. ECCV*, 2020.

[124] Marc Pollefeys, David Nistér, J-M Frahm, Amir Akbarzadeh, Philippos Mordohai, Brian Clipp, Chris Engels, David Gallup, S-J Kim, Paul Merrell, et al. Detailed real-time urban 3d reconstruction from video. *IJCV*, 78:143–167, 2008.

[125] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Flot: Scene flow on point clouds guided by optimal transport. In *Proc. ECCV*, 2020.

[126] Yi-Ling Qiao, Lin Gao, Yu-Kun Lai, Fang-Lue Zhang, Ming-Ze Yuan, and Shihong Xia. Sf-net: Learning scene flow from rgb-d images with cnns. In *Proc. BMVC.*, 2018.

[127] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Trans. Robotics*, 34(4):1004–1020, 2018.

[128] Dhanesh Ramachandram and Graham W Taylor. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE signal processing magazine*, 34 (6):96–108, 2017.

[129] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *Proc. ICCV*, 2021.

[130] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE PAMI*, 44(3), 2022.

[131] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proc. CVPR*, 2019.

[132] Albert W Reed, Hyojin Kim, Rushil Anirudh, K Aditya Mohan, Kyle Champley, Jingu Kang, and Suren Jayasuriya. Dynamic ct reconstruction from limited views with implicit neural representations and parametric motion fields. In *Proc. ICCV*, 2021.

[133] Zhile Ren, Deqing Sun, Jan Kautz, and Erik Sudderth. Cascaded scene flow prediction using semantic segmentation. In *Proc. 3DV*, 2017.

[134] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *Proc. CVPR*, 2018.

[135] René Schuster, Oliver Wasenmuller, Georg Kuschk, Christian Bailer, and Didier Stricker. Sceneflowfields: Dense interpolation of sparse scene flow correspondences. In *Proc. WACV*, 2018.

[136] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proc. CVPR*, 2016.

[137] Liyue Shen, John Pauly, and Lei Xing. Nerp: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[138] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo Tourism: Exploring Photo Collections in 3D. In *Proc. ACM SIGGRAPH*, 2006.

[139] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *ACM international conference on Multimedia*, 2005.

[140] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. Smurf: Self-teaching multi-frame unsupervised raft with full-image warping. In *Proc. CVPR*, 2021.

[141] Hauke Strasdat, J Montiel, and Andrew J. Davison. Scale Drift-Aware Large Scale Monocular SLAM. *Robotics: Science and Systems VI*, 2(3):7, 2010.

[142] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proc. ICCV*, 2021.

[143] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proc. CVPR*, 2018.

[144] Deyu Sun, Jiquan Liu, Cristian A Linte, Huilong Duan, and Richard A Robb. Surface reconstruction from tracked endoscopic video using the structure from motion approach. In *Proc. of Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions (AE-CAI)*, 2013.

[145] Niko Sünderhauf and Peter Protzel. Switchable Constraints for Robust Pose Graph SLAM. In *Proc. IEEE/RSJ Conf. on Intelligent Robots and Systems*, pages 1879–1884, 2012.

[146] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *arXiv preprint arXiv:1806.04807*, 2018.

[147] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proc. CVPR*, 2017.

[148] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018.

[149] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. ECCV*, 2020.

[150] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Proc. NeurIPS*, 2021.

[151] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proc. CVPR*, 2021.

[152] Zachary Teed and Jia Deng. Tangent space backpropagation for 3d transformation groups. In *Proc. CVPR*, 2021.

[153] Bill Triggs, Philip F. McLauchlan, Richard Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment—A Modern Synthesis. In *International workshop on vision algorithms*, 1999.

[154] Shinji Umeyama. Least-Squares Estimation of Transformation Parameters between Two Point Patterns. *IEEE PAMI*, 13(4):376–380, 1991.

[155] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMoN: Depth and Motion Network for Learning Monocular Stereo. In *Proc. CVPR*, 2017.

[156] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proc. NeurIPS*, 2017.

[157] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *Proc. ICCV*, 1999.

[158] Christoph Vogel, Konrad Schindler, and Stefan Roth. Piecewise rigid scene flow. In *Proc. ICCV*, 2013.

[159] Christoph Vogel, Stefan Roth, and Konrad Schindler. View-consistent 3d scene flow estimation over multiple frames. In *Proc. ECCV*, 2014.

[160] Christoph Vogel, Konrad Schindler, and Stefan Roth. 3d scene flow estimation with a piecewise rigid scene model. *IJCV*, 115(1):1–28, 2015.

[161] Guangming Wang, Yunzhe Hu, Zhe Liu, Yiyang Zhou, Masayoshi Tomizuka, Wei Zhan, and Hesheng Wang. What matters for 3d scene flow network. In *Proc. ECCV*, 2022.

[162] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[163] Rui Wang, Stephen M Pizer, and Jan-Michael Frahm. Recurrent Neural Network for (Un-)supervised Learning of Monocular VideoVisual Odometry and Depth. In *Proc. CVPR*, 2019.

[164] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks. In *Proc. Intl. Conf. on Robotics and Automation*, 2017.

[165] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proc. CVPR*, 2019.

[166] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep Multimodal Fusion by Channel Exchanging. *Proc. NeurIPS*, 2020.

[167] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *Proc. MICCAI*, 2022.

[168] Zirui Wang, Shuda Li, Henry Howard-Jenkins, Victor Prisacariu, and Min Chen. Flownet3d++: Geometric losses for deep scene flow estimation. In *Proc. WACV*, 2020.

[169] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

[170] Peng Wei, Guoliang Hua, Weibo Huang, Fanyang Meng, and Hong Liu. Unsupervised Monocular Visual-inertial Odometry Network. In *Proc. IJCAI*, 2020.

[171] Yi Wei, Ziyi Wang, Yongming Rao, Jiwen Lu, and Jie Zhou. Pv-raft: point-voxel correlation fields for scene flow estimation of point clouds. In *Proc. CVPR*, 2021.

[172] Tomás Werner and Tomás Pajdla. Cheirality in epipolar geometry. In *Proc. ICCV*, 2001.

[173] Aji Resindra Widya, Yusuke Monno, Kosuke Imahori, Masatoshi Okutomi, Sho Suzuki, Takuji Gotoda, and Kenji Miki. 3d reconstruction of whole stomach from endoscope video using structure-from-motion. In *Proc. EMBC*, 2019.

[174] Aji Resindra Widya, Yusuke Monno, Masatoshi Okutomi, Sho Suzuki, Takuji Gotoda, and Kenji Miki. Stomach 3D Reconstruction Based on Virtual Chromoendoscopic Image Generation. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2020.

[175] Changchang Wu. Towards Linear-Time Incremental Structure from Motion. In *Proc. 3DV*, 2013.

[176] Qing Wu, Yuwei Li, Yawen Sun, Yan Zhou, Hongjiang Wei, Jingyi Yu, and Yuyao Zhang. An arbitrary scale super-resolution approach for 3d mr images via implicit neural representation. *IEEE Journal of Biomedical and Health Informatics*, 27(2):1004–1015, 2022.

[177] Wenxuan Wu, Zhi Yuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. Pointpwc-net: Cost volume on point clouds for (self-) supervised scene flow estimation. In *Proc. ECCV*, 2020.

[178] Fei Xue, Xin Wang, Shunkai Li, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Beyond Tracking: Selecting Memory and Refining Poses for Deep Visual Odometry. In *Proc. CVPR*, 2019.

[179] Gengshan Yang and Deva Ramanan. Upgrading optical flow to 3d scene flow through optical expansion. In *Proc. ECCV*, 2020.

[180] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proc. CVPR*, 2018.

[181] Jin Zeng, Yanfeng Tong, Yunmu Huang, Qiong Yan, Wenxiu Sun, Jing Chen, and Yongtian Wang. Deep Surface Normal Estimation With Hierarchical RGB-D Fusion. In *Proc. CVPR*, 2019.

[182] Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *Proc. MICCAI*, 2023.

[183] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction. In *Proc. CVPR*, 2018.

[184] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proc. CVPR*, 2023.

[185] Yinda Zhang and Thomas Funkhouser. Deep Depth Completion of a Single RGB-D Image. In *Proc. CVPR*, 2018.

[186] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *Proc. 3DV*, 2022.

[187] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards Better Generalization: Joint Depth-Pose Learning Without PoseNet. In *Proc. CVPR*, 2020.

[188] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. *arXiv preprint arXiv:2110.09482*, 2021.

[189] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *Proc. BMVC.*, 2021.

[190] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised Learning of Depth and Ego-Motion From Video. In *Proc. CVPR*, 2017.

[191] Zhengming Zhou and Qiulei Dong. Self-distilled feature aggregation for self-supervised monocular depth estimation. In *Proc. ECCV*, 2022.

[192] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In *Proc. ICCV*, 2021.

[193] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proc. CVPR*, 2022.

[194] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. *arXiv preprint arXiv:2302.03594*, 2023.

[195] Tao Zhuang, Zhixuan Zhang, Yuheng Huang, Xiaoyi Zeng, Kai Shuang, and Xiang Li. Neuron-level Structured Pruning using Polarization Regularizer. In *Proc. NeurIPS*, 2020.

[196] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proc. ECCV*, 2018.

[197] Yuliang Zou, Pan Ji, Quoc-Huy Tran, Jia-Bin Huang, and Manmohan Chandraker. Learning Monocular Visual Odometry via Self-Supervised Long-Term Modeling. In *Proc. ECCV*, 2020.

Related Publications by the Author

International Conferences

1. Zijie Jiang, Hajime Taira, Naoyuki Miyashita and Masatoshi Okutomi, "VIO-Aided Structure from Motion Under Challenging Environments," Proceedings of IEEE International Conference on Industrial Technology (ICIT2021), pp.950-957, Mar. 2021.

2. Zijie Jiang, Hajime Taira, Naoyuki Miyashita and Masatoshi Okutomi, "Self-Supervised Ego-Motion Estimation Based on Multi-Layer Fusion of RGB and Inferred Depth," Proceedings of IEEE International Conference on Robotics and Automation (ICRA2022), pp.7605-7611, May 2022.

3. Zijie Jiang and Masatoshi Okutomi, "EMR-MSF: Self-Supervised Recurrent Monocular Scene Flow Exploiting Ego-Motion Rigidity," Proceedings of IEEE International Conference on Computer Vision (ICCV2023), pp.69-78, Oct. 2023.

4. Zijie Jiang, Yusuke Monno and Masatoshi Okutomi, "Neural Radiance Fields for Novel View Synthesis in Monocular Gastroscopy," Submitted to Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society, July 2024.

Domestic Conferences

1. 蒋子傑, 田平創, 宮下尚之, 奥富正敏, "カラー画像と推定深度画像の Multi-Layer Fusion に基づく自己教師あり学習による Ego-Motion 推定", 第28回画像センシングシンポジウム(SSII2022), June 2022.