Discriminately Treating Motion Components Evolves Joint Depth and Ego-Motion Learning

Mengtan Zhang, Zizhan Guo, Hongbo Zhao, Yi Feng, Zuyi Xiong, Yue Wang, Shaoyi Du, Hanli Wang, *Senior Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Unsupervised learning of depth and ego-motion, two fundamental tasks in 3D perception, has made significant strides in recent years. Nevertheless, most existing methods often treat ego-motion estimation as an auxiliary task, either indiscriminately mixing all motion types or excluding depth-independent rotational motions when generating supervisory signals. Such designs hinder the incorporation of sufficiently strong geometric constraints into the joint learning framework, thereby limiting the reliability and robustness of these models under diverse environmental conditions. This study introduces a discriminative treatment of motion components by leveraging the distinct geometric regularities inherent in their respective rigid flows, which simultaneously benefits both depth and ego-motion estimation. Given consecutive video frames, the network outputs are first employed to align the optical axes and imaging planes of the source and target cameras. Optical flows between the original frames are then transformed through these alignment processes, and their deviations are quantified to impose geometric constraints. These constraints are applied individually to each estimated ego-motion component, enabling more targeted and reliable refinement. These alignments further reformulate the joint learning process into coaxial and coplanar forms, where depth and each translation component can be mutually derived through closed-form geometric relationships, introducing complementary constraints that significantly improve the robustness of depth learning. DiMoDE, a general depth and ego-motion joint learning framework that incorporates all these innovative designs, achieves state-of-the-art performance in both tasks, as demonstrated by extensive experiments on multiple public datasets and a newly collected, diverse real-world dataset, particularly under challenging conditions where existing approaches often fail. Our source code will be publicly available at mias.group/DiMoDE upon publication.**

***Index Terms*—depth estimation, visual odometry, motion component, geometric constraint.**

(*Corresponding author: Rui Fan*)

Mengtan Zhang and Hongbo Zhao are with the Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 201210, China (e-mails: zmt200110@tongji.edu.cn, hongbozhao@tongji.edu.cn).

Zizhan Guo, Yi Feng, and Zuyi Xiong are with the College of Electronic & Information Engineering, Tongji University, Shanghai 201804, China (e-mails: 2431983@tongji.edu.cn, fengyi@ieee.org, 2153478@tongji.edu.cn).

Yue Wang is with the Department of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China (e-mail: wangyue@iipc.zju.edu.cn).

Shaoyi Du is with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, the National Engineering Research Center for Visual Information and Applications, and the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: dushaoyi@xjtu.edu.cn).

Hanli Wang is with the College of Electronic & Information Engineering, the School of Computer Science and Technology, and the Key Laboratory of Embedded System and Service Computing (Ministry of Education), Tongji University, Shanghai 201804, China (e-mail: hanliwang@tongji.edu.cn).

Rui Fan is with the College of Electronic & Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Autonomous Intelligent Unmanned Systems, the Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), and Shanghai Key Laboratory of Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China, as well as with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'An, Shaanxi 710049, China (e-mail: rui.fan@ieee.org).

I. Introduction

ACCURATE estimation of ego-motion and depth is crucial for 3D perception [1]. Recently, there has been a significant increase in methods that jointly learn ego-motion and depth from monocular video sequences in an unsupervised, end-to-end manner [2]–[4]. Compared to supervised depth estimation approaches [5]–[7], this paradigm eliminates the need for extensive, manually labeled annotations [2], [8], which are not always available in real-world scenes. Moreover, compared to traditional visual odometry methods that rely on 2D visual correspondences to explicitly solve the relative pose problem, such a joint learning paradigm can maintain robustness in scenarios where 2D correspondences are fairly sparse or unreliable [9], [10].

This unsupervised joint learning paradigm typically consists of a depth estimation network (hereafter known as *DepthNet*) and a pose estimation network (hereafter known as *PoseNet*) [4], [8]. DepthNet estimates a dense depth map. PoseNet regresses the following ego-motion transformation:

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^\top & 1 \end{pmatrix} \in \mathrm{SE}(3), \tag{1}$$

where $\boldsymbol{R} \in \mathrm{SO}(3)$ denotes the rotation matrix, $\boldsymbol{t} = (t_x, t_y, t_z)^\top \in \mathbb{R}^3$ denotes the translation vector, and $\boldsymbol{0}$ denotes a column vector of zeros. The estimated depth map and ego-motion are then utilized to calculate the rigid flow, a key component for warping the source image into the target view. By minimizing the photometric difference between the target and the warped images, these two networks can be trained simultaneously.

Fig. 1 presents the rigid flows resulting from various motion types, a topic that has received limited discussion in previous studies. As shown in Figs. 1(a) and 1(b), ego-motion that involves only translations yields fairly regular rigid flows. Specifically, when the camera moves perpendicularly to the optical axis, leading to a tangential translation vector $\boldsymbol{t}^{\mathrm{Tan}} = (t_x, t_y, 0)^\top$, the rigid flows of all pixels are parallel. When the camera moves along the optical axis, leading to a radial translation vector $\boldsymbol{t}^{\mathrm{Rad}} = (0, 0, t_z)^\top$, the rigid flows of all pixels move toward or away from the principal point.

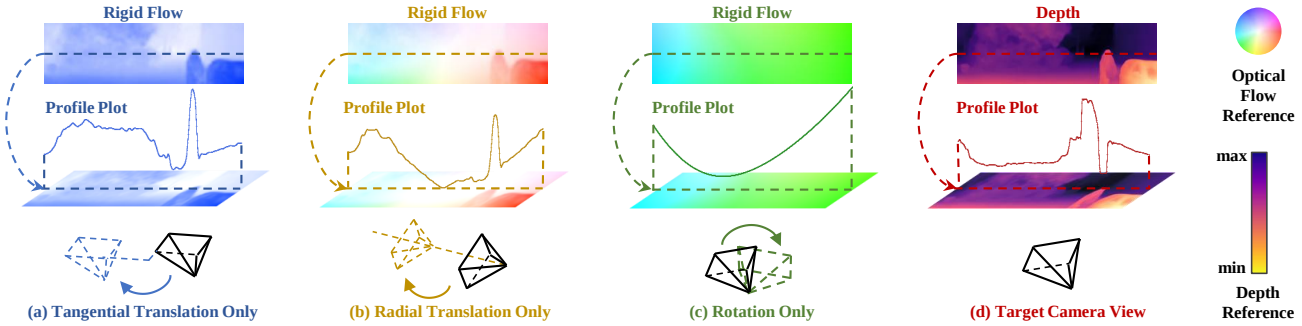

Fig. 1. Tangential and radial translations result in geometrically regular but distinct depth-dependent flows. Specifically, the rigid flow induced by tangential translation varies inversely with depth, while the one resulting from radial translation is not only depth-dependent but also subject to perspective scaling. In contrast, rotation results in irregular, depth-independent flows.

These regular rigid flows provide vital visual clues that can potentially improve the accuracy of estimated motion components. On the other hand, as illustrated in Fig. 1(c), ego-motion that involves only rotations results in fairly irregular rigid flows. Therefore, it is inappropriate to train PoseNet to indiscriminately regress rotation and translation components by mixing these two types of rigid flows for supervisory signal generation. However, there are currently few previous studies that focus specifically on this aspect.

In addition, we revisit Fig. 1 to explore potential improvements in DepthNet training. As demonstrated in the study [11], two images related by a pure-rotation transformation can be warped from one to the other using a homography matrix $\boldsymbol{H} = \boldsymbol{K}\boldsymbol{R}\boldsymbol{K}^{-1}$, where $\boldsymbol{K}$ denotes the camera's intrinsic matrix. Therefore, as shown in Fig. 1(c), the rigid flow produced solely by rotation is independent of depth and does not provide effective gradients necessary for the supervision of DepthNet training. In contrast, as illustrated in Fig. 1(a), pure tangential translation results in rigid flows characterized by superposed horizontal and vertical positional shifts. These shifts, commonly referred to as disparities in stereo vision, are inversely proportional to depth (see Fig. 1(d)) and can directly contribute to the training of DepthNet. Furthermore, as illustrated in Fig. 1(b), pure radial translation leads to perspective scaling in sequential images, which is also depth-dependent. Specifically, the scale change of an object in the image is inversely proportional to its depth. Thus, these rigid flows also provide valuable visual cues that can be effectively incorporated into the DepthNet training process. However, perspective scaling causes pixels near the image boundary to undergo larger displacements than those near the principal point. Consequently, mixing these two types of visual cues can lead to inconsistent gradients across the image, potentially impeding convergence and degrading depth estimation performance. Despite this limitation, the significant differences between tangential and radial translations have long been overlooked, with existing studies generally treating them indiscriminately.

These observations motivate us to discriminately treat the motion components in $\boldsymbol{T}$ by isolating the rotation and separately processing the tangential and radial translations. Specifically, the estimated rotation $\boldsymbol{R}$ is expected to align the camera orientation of the source view with that of the target view, thereby eliminating irregular rotational flows. The estimated tangential translation $\boldsymbol{t}^{\text{Tan}}$ and radial translation $\boldsymbol{t}^{\text{Rad}}$ are then expected to align the camera's optical axes and imaging planes in the source and target views, respectively, thereby generating two types of regular flows containing vital visual clues.

Building upon the aforementioned analyses, this study aims to reformulate the existing joint learning framework for depth and ego-motion estimation by introducing explicit geometric constraints from motion components. These constraints are imposed through the alignment of the optical axes and imaging planes between the source and target cameras. In practice, we utilize dense correspondences that adhere to the static scene hypothesis to objectively quantify the extent of both types of alignment. These correspondences, when transformed by performing both alignments based on the estimated depth and ego-motion, are expected to exhibit the aforementioned geometric regularities. Deviations from the expected geometric regularities are employed as supervisory signals to independently guide the learning of tangential and radial translations, while simultaneously constraining the learning of rotation. As the transformed correspondences progressively exhibit expected geometric regularities, depth-dependent visual cues arising from positional shifts and perspective scaling are effectively disentangled. By exploiting their distinct depth-dependent variations, these visual cues are processed separately to recover depth, thereby providing per-pixel constraints for depth estimation. The depth estimated under these constraints and the transformed correspondences are subsequently used to compute tangential and radial translations based on simplified geometric relationships. A geometric constraint is also incorporated to enforce consistency between these computed translations and those predicted by the PoseNet, thereby establishing a constraining cycle that enhances the robustness of depth learning. Incorporating the above-mentioned innovative components into a general depth and ego-motion joint learning framework yields **DiMoDE**, which discriminately treats motion components. Extensive experiments on multiple public datasets and our collected real-world video sequences demonstrate the state-of-the-art (SoTA) performance of our proposed DiMoDE in both monocular depth estimation and visual odometry, as well as the efficacy of discriminately treating motion components.

In summary, our main contributions are as follows:

- We conduct both phenomenological and theoretical analyses on the distinctions among rigid flows arising from different motion types and the long-overlooked limitations caused by indiscriminately mixing these flows when generating supervisory signals.
- We introduce optical axis and imaging plane alignment processes that reformulate joint learning from monocular video into coaxial and coplanar forms, thereby mitigating the effects of mixed motion types and enhancing the robustness and stability of depth learning.
- We propose a discriminative treatment of motion components that enables these alignments to be achieved using network outputs, not only eliminating the need for auxiliary pose estimation algorithms but also refining each component of the PoseNet outputs.
- We develop DiMoDE, a general framework compatible with various models, achieving notable improvements in network convergence and training robustness for both depth estimation and visual odometry across diverse challenging conditions.

The remainder of this article is organized as follows: Sect. II reviews existing monocular visual odometry and depth estimation methods. Sect. III introduces preliminaries of monocular depth and ego-motion joint learning. Sect. IV details the proposed discriminative treatment of motion components, and Sect. V presents the resulting DiMoDE framework, respectively. Sect. VI presents the experimental results across several public datasets and our collected real-world dataset. In Sect. VII, we discuss two limitations of the DiMoDE framework. Finally, Sect. VIII concludes this article and discusses possible directions for future work.

II. Related Work

A. Monocular Visual Odometry

Monocular visual odometry is a fundamental and extensively studied problem in the fields of robotics and computer vision [12]. Existing approaches can be broadly categorized into three main classes: geometry-based, learning-based, and hybrid.

Earlier geometry-based methods, such as VISO2 [13] and ORB-SLAM series [14]–[16], primarily rely on hand-crafted keypoints and descriptors for correspondence matching, and often perform poorly in texture-less environments [9], [12]. To address this limitation, methods like DSO [17] formulate direct image or feature warping as an energy minimization problem, thus avoiding the need for accurate correspondence matching. Despite their robustness in texture-less environments, these methods exhibit a lower accuracy ceiling compared to approaches based on correspondence matching [18].

In recent years, extensive learning-based methods [19], [20] have been proposed to tackle the above-mentioned challenges. These methods typically employ a PoseNet to regress camera poses from consecutive video frames in an end-to-end manner [8], [21]–[23]. By eliminating the need for explicit correspondence matching used to solve for relative pose, PoseNet remains robust in texture-less scenes and typically supports efficient inference. Among them, the study [8] proposes a joint depth and ego-motion learning paradigm that trains the PoseNet with only unlabeled monocular video, thereby alleviating the need for expensive manual annotation. Subsequent studies have focused on addressing issues such as scale inconsistency and limited generalizability [23], [24], while also incorporating temporal and spatial cues into PoseNet to improve their performance [10], [25], [26]. Despite these advancements, the fundamental challenges of generating more reliable supervisory signals in an unsupervised manner to ensure accurate and robust ego-motion learning remain unsolved [12], [18].

More recently, hybrid approaches [12], [18], [27], [28] have introduced a new paradigm that integrates neural network predictions, such as depth and correspondence matching, with back-end optimization for relative pose estimation. Despite the incorporation of geometric constraints during inference, enabling these hybrid methods to achieve superior performance over fully learning-based approaches, the resulting high storage demands and computational complexity significantly hinder their applicability in real-world, resource-constrained environments.

This study incorporates explicit geometric constraints from each ego-motion component by enforcing the alignment of optical axes and imaging planes within the joint learning framework. The proposed DiMoDE framework enables a simple ResNet-based PoseNet [29] to achieve performance comparable to SoTA geometric-based and hybrid approaches, with significantly lower storage and computational overhead.

B. Monocular Depth Estimation

Monocular depth estimation, which infers depth information from a single RGB image, has found widespread applications in areas such as autonomous driving and embodied artificial intelligence. Early approaches typically adopted a supervised learning paradigm that requires extensive depth ground truth for model training [5]. Building on this foundation, subsequent studies [30]–[33] improved depth estimation performance by reformulating the depth regression problem as either a per-pixel classification task that predicts discretized depth intervals or a classification-regression task that predicts depth as a weighted combination of bin centers. More recently, the studies [7], [34], [35] have leveraged vision foundation models to achieve highly accurate depth estimation.

To alleviate the need for extensive ground-truth data in monocular depth estimation, growing interest has been directed toward unsupervised learning of depth from monocular video, a paradigm first introduced in the study [8]. To improve depth estimation performance within this paradigm, the study [2] introduced a per-pixel minimum reprojection loss to enhance the photometric supervisory signals, while studies [4], [36], [37] explored more sophisticated network architectures. Subsequent studies [3], [38]–[41] have sought to preserve the effectiveness of this training paradigm in scenarios with occlusions and dynamic objects by introducing additional visual cues. However, these methods rely solely on per-pixel supervisory consistency, which limits their robustness under adverse conditions, such as nighttime or poor weather, where such signals often become unreliable.

Recently, several studies [42], [43] have focused specifically on enhancing the robustness of depth estimation under adverse conditions. These methods typically adopt domain adaptation strategies, wherein networks are initially trained on clear daytime scenes and subsequently adapted to more challenging environments such as nighttime or fog. Consequently, their reliance on fixed data splits hinders the full utilization of large-scale data collected under diverse and dynamically changing conditions, limiting the continuous evolution of DepthNet.

In contrast, this study focuses on the discriminative treatment of motion components during the original supervisory signal generation, enabling a robust constraint cycle for depth estimation. To the best of our knowledge, DiMoDE is the first approach to effectively address the challenges posed by adverse environments within a unified and general joint learning framework.

III. Preliminaries

Given a monocular camera with the intrinsic matrix

$$\boldsymbol{K} = \begin{pmatrix} f_u & 0 & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{2}$$

where f_u and f_v denote the focal lengths in the horizontal and vertical directions, respectively, and $\boldsymbol{p}_0 = (u_0, v_0)^\top$ represents the principal point, a pair of RGB images, $\boldsymbol{I}_t$ and $\boldsymbol{I}_s$, captured from the target and source views, respectively, are fed into the DepthNet to generate their corresponding depth maps, $\boldsymbol{D}_t$ and $\boldsymbol{D}_s$. Let $\boldsymbol{p}_t = (u_t, v_t)^\top$ and $\boldsymbol{p}_s = (u_s, v_s)^\top$ be a pair of corresponding pixels in the target and source images, respectively. The depth values at these pixels are denoted as $\hat{z}_t = \boldsymbol{D}_t(\boldsymbol{p}_t)$ and $\hat{z}_s = \boldsymbol{D}_s(\boldsymbol{p}_s)$. Their respective homogeneous coordinates are denoted as $\tilde{\boldsymbol{p}}_t$ and $\tilde{\boldsymbol{p}}_s$. Meanwhile, $\boldsymbol{I}_t$ and $\boldsymbol{I}_s$ are concatenated and fed into the PoseNet to estimate the ego-motion $\hat{\boldsymbol{T}}$ from the source view to the target view. The estimated depth and ego-motion are then used to generate a rigid flow map $\boldsymbol{F}^{\text{Rig}}_{t\to s}$ in the target view based on the following relation:

$$\tilde{\boldsymbol{p}}_s = \tilde{\boldsymbol{p}}_t + \begin{pmatrix} \boldsymbol{f}^{\text{Rig}}_{t\to s} \\ 0 \end{pmatrix} = \frac{1}{\hat{z}_s} \boldsymbol{K} \left(\hat{z}_t \hat{\boldsymbol{R}} \boldsymbol{K}^{-1} \tilde{\boldsymbol{p}}_t + \hat{\boldsymbol{t}} \right), \tag{3}$$

where $\boldsymbol{f}^{\text{Rig}}_{t\to s} = \boldsymbol{F}^{\text{Rig}}_{t\to s}(\boldsymbol{p}_t)$ denotes the rigid flow at pixel $\boldsymbol{p}_t$ and is expected to equal the optical flow $\boldsymbol{f}^{\text{Opt}}_{t\to s} = \boldsymbol{F}^{\text{Opt}}_{t\to s}(\boldsymbol{p}_t) = \boldsymbol{p}_s - \boldsymbol{p}_t$ in static regions. $\boldsymbol{I}_s$ is then warped into the target view based on the rigid flow map, resulting in a synthesized image $\boldsymbol{I}'_t$ expressed as follows [8]:

$$\boldsymbol{I}'_t = \mathcal{W}\left(\boldsymbol{I}_s, \boldsymbol{F}^{\text{Rig}}_{t\to s} \right), \tag{4}$$

where $\mathcal{W}(\cdot)$ denotes the image warping function. To quantify the appearance discrepancy between $\boldsymbol{I}'_t$ and $\boldsymbol{I}_t$, the following photometric reprojection loss is computed:

$$\mathcal{L}_{\text{pho}} = \alpha \frac{1 - \text{SSIM}(\boldsymbol{I}'_t, \boldsymbol{I}_t)}{2} + (1-\alpha) \left\| \boldsymbol{I}'_t - \boldsymbol{I}_t \right\|_1, \tag{5}$$

where SSIM denotes the pixel-wise structural similarity index operation [44], and the weight α is empirically set to 0.85, following the study [2]. (5) serves as the supervisory signal for the training of both DepthNet and PoseNet [23].

IV. Discriminative Treatment of Motion Components

Ego-motion estimation plays a role equally important to that of depth estimation, as the supervisory signal is jointly determined by the transformation $\boldsymbol{T}$ predicted by PoseNet and the depth map $\boldsymbol{D}_t$ generated by DepthNet. Unfortunately, ego-motion estimation has long been regarded as an auxiliary task [2], [4], with limited attention given to thoroughly analyzing the outputs produced by PoseNet. As discussed in Sect. I, $\boldsymbol{T}$ is a mixture of three types of transformations as follows:

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^\top & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{I} & \boldsymbol{t}^{\text{Rad}} \\ \boldsymbol{0}^\top & 1 \end{pmatrix}}_{\boldsymbol{T}^{\text{Rad}}} \underbrace{\begin{pmatrix} \boldsymbol{I} & \boldsymbol{t}^{\text{Tan}} \\ \boldsymbol{0}^\top & 1 \end{pmatrix}}_{\boldsymbol{T}^{\text{Tan}}} \underbrace{\begin{pmatrix} \boldsymbol{R} & \boldsymbol{0} \\ \boldsymbol{0}^\top & 1 \end{pmatrix}}_{\boldsymbol{T}^{\text{Rot}}}, \tag{6}$$

where $\boldsymbol{T}^{\text{Rot}}$, $\boldsymbol{T}^{\text{Tan}}$, and $\boldsymbol{T}^{\text{Rad}}$ represent pure rotation, pure tangential translation, and pure radial translation transformations, respectively. The prior art [11] claims that $\boldsymbol{T}^{\text{Rot}}$ is irrelevant to the training of DepthNet, as in the pure rotation case, one image can be warped to the other using $\boldsymbol{T}^{\text{Rot}}$ alone, without the need for depth information [45]. More importantly, even slight rotational errors in $\boldsymbol{T}^{\text{Rot}}$ can be mistakenly interpreted as translational motions, leading to erroneous estimations of $\boldsymbol{T}^{\text{Tan}}$ and $\boldsymbol{T}^{\text{Rad}}$. The resulting supervisory signals can, in turn, mislead the training of DepthNet. Therefore, the authors proposed to exclude rotational motions from DepthNet training within the joint learning framework. Specifically, the original target and source images, $\boldsymbol{I}_t$ and $\boldsymbol{I}_s$, are respectively warped based on $\hat{\boldsymbol{T}}^{\text{Rot}}$, which is estimated using either the five-point algorithm [46] or an auxiliary network [29], to align the orientations of source and target cameras. The warped image pair is then fed into the joint learning framework, enabling both PoseNet and DepthNet to be trained under the assumption that only translational motions are involved. Nonetheless, rotation estimation performed by an independent algorithm or network remains decoupled from the original joint learning framework, which introduces considerable computational and storage overhead without directly benefiting PoseNet training. Moreover, rotational errors introduced by decoupled algorithms are no longer considered for minimization during training, allowing them to propagate through the pipeline and degrade overall framework performance. More critically, all existing methods overlook the distinct regularities inherent in the two types of translational motions, which are closely tied to depth estimation.

Therefore, this study investigates the discriminative treatment of three distinct motion components to simultaneously improve both depth and pose estimation performance within the joint learning framework. Specifically, when the ego-motion corresponds to a pure rotation transformation $\boldsymbol{T}^{\text{Rot}}$, the pixel correspondences $\boldsymbol{p}_t$ and $\boldsymbol{p}_s$ are expected to satisfy the following relation:

$$\tilde{\boldsymbol{p}}_s = \frac{z_t}{z_s} \underbrace{\boldsymbol{K}\boldsymbol{R}\boldsymbol{K}^{-1}}_{\boldsymbol{H}} \tilde{\boldsymbol{p}}_t \simeq \underbrace{\left(\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3\right)^\top}_{\boldsymbol{H}} \tilde{\boldsymbol{p}}_t, \tag{7}$$

where $\boldsymbol{H}$ denotes the homography matrix [45] with $\boldsymbol{h}_i^\top = (h_{i1}, h_{i2}, h_{i3})$ representing the i-th row vector, and the symbol

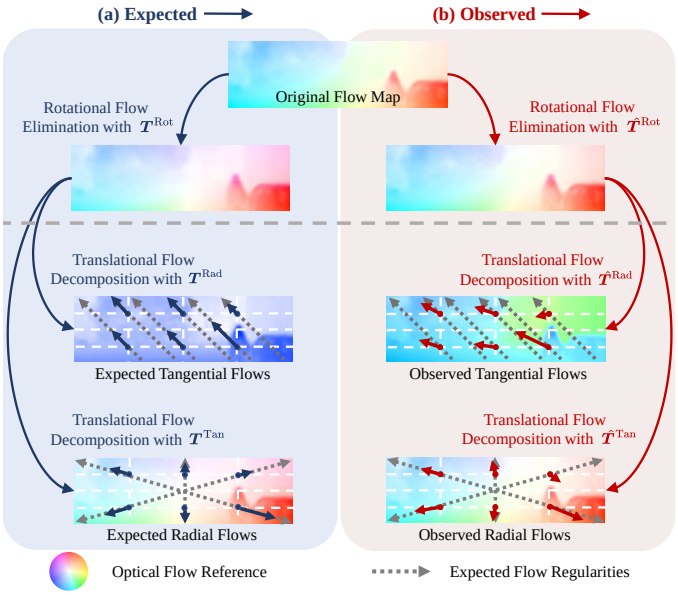

Fig. 2. The discriminative treatment of motion components and the resulting flow decomposition processes.

$\simeq$ indicates that the two vectors are equal up to a scale factor. Rearranging (7) yields

$$\boldsymbol{p}_s = \left(\frac{\boldsymbol{h}_1^\top \tilde{\boldsymbol{p}}_t}{\boldsymbol{h}_3^\top \tilde{\boldsymbol{p}}_t}, \frac{\boldsymbol{h}_2^\top \tilde{\boldsymbol{p}}_t}{\boldsymbol{h}_3^\top \tilde{\boldsymbol{p}}_t} \right)^\top \tag{8}$$

and

$$\boldsymbol{f}_{t\to s}^{\text{Rig}} = \frac{1}{\boldsymbol{h}_3^\top \tilde{\boldsymbol{p}}_t} \begin{pmatrix} \tilde{\boldsymbol{p}}_t^\top \left(-\boldsymbol{h}_3 \boldsymbol{i}_1^\top\right) \tilde{\boldsymbol{p}}_t + \boldsymbol{h}_1^\top \tilde{\boldsymbol{p}}_t \\ \tilde{\boldsymbol{p}}_t^\top \left(-\boldsymbol{h}_3 \boldsymbol{i}_2^\top\right) \tilde{\boldsymbol{p}}_t + \boldsymbol{h}_2^\top \tilde{\boldsymbol{p}}_t \end{pmatrix}, \tag{9}$$

where $\boldsymbol{i}_k \in \mathbb{R}^3$ denotes the k-th column vector of an 3×3 identity matrix. (9) indicates that, when only rotation is involved, the rigid flow becomes a nonlinear function of the pixel coordinates $\boldsymbol{p}_t$. This nonlinearity arises from two sources: (1) the quadratic terms $\tilde{\boldsymbol{p}}_t^\top (-\boldsymbol{h}_3 \boldsymbol{i}_1^\top) \tilde{\boldsymbol{p}}_t = h_{31} u_t^2 + h_{32} u_t v_t + h_{33} u_t$ and $\tilde{\boldsymbol{p}}_t^\top (-\boldsymbol{h}_3 \boldsymbol{i}_2^\top) \tilde{\boldsymbol{p}}_t = h_{31} u_t v_t + h_{32} v_t^2 + h_{33} v_t$, as well as (2) the perspective division term $1/(\boldsymbol{h}_3^\top \tilde{\boldsymbol{p}}_t) = 1/(h_{31} u_t + h_{32} v_t + h_{33})$. As a result, in such cases, the rigid flow is determined solely by the rotation and the intrinsic matrix, and can vary significantly across the entire image, giving rise to a highly irregular rigid flow map[1], as shown in Fig. 1(c).

Therefore, instead of relying on an independent algorithm or auxiliary network to estimate rotation for image warping, we directly utilize $\hat{\boldsymbol{T}}^{\text{Rot}}$ provided by the PoseNet to eliminate the irregular rotational flows from the original optical flow map. Ideally, the resulting rigid flow map retains only the two types of translational flows, induced by $\boldsymbol{T}^{\text{Tan}}$ and $\boldsymbol{T}^{\text{Rad}}$, which are directly relevant to depth estimation, as illustrated in Fig. 2(a).

[1]Although rotational flows also comprises three relatively regular components [47], [48], they cannot be obtained practically by decomposing the original flows. This is because the translational component cannot be eliminated prior to the rotational ones due to the non-commutative nature of rigid-body transformations, as shown in (6).

When the ego-motion corresponds to a pure tangential translation transformation $\boldsymbol{T}^{\text{Tan}}$, the pixel correspondences $\boldsymbol{p}_t$ and $\boldsymbol{p}_s$ are expected to satisfy the following relation:

$$\tilde{\boldsymbol{p}}_s \simeq \boldsymbol{K} \left(z_t \boldsymbol{K}^{-1} \tilde{\boldsymbol{p}}_t + \boldsymbol{t}^{\text{Tan}} \right). \tag{10}$$

Substituting (2) into (10) yields

$$\boldsymbol{p}_s = \left(\frac{z_t u_t + f_u t_x}{z_t}, \frac{z_t v_t + f_v t_y}{z_t} \right)^\top \tag{11}$$

and

$$\boldsymbol{f}_{t\to s}^{\text{Rig}} = \frac{1}{z_t} \begin{pmatrix} f_u t_x \\ f_v t_y \end{pmatrix}, \tag{12}$$

which indicates that the tangential translational flows across the image are oriented in the direction of $(f_u t_x, f_v t_y)^\top$ and vary solely with depth, as illustrated in Fig. 2(a). Similarly, when the ego-motion corresponds to a pure radial translation transformation $\boldsymbol{T}^{\text{Rad}}$, the pixel correspondences $\boldsymbol{p}_t$ and $\boldsymbol{p}_s$ are expected to satisfy the following relation:

$$\tilde{\boldsymbol{p}}_s \simeq \boldsymbol{K} \left(z_t \boldsymbol{K}^{-1} \tilde{\boldsymbol{p}}_t + \boldsymbol{t}^{\text{Rad}} \right). \tag{13}$$

Substituting (2) into (13) yields

$$\boldsymbol{p}_s = \left(\frac{z_t u_t + u_0 t_z}{z_t + t_z}, \frac{z_t v_t + v_0 t_z}{z_t + t_z} \right)^\top \tag{14}$$

and

$$\boldsymbol{f}_{t\to s}^{\text{Rig}} = \frac{-t_z}{z_t + t_z} \begin{pmatrix} u_t - u_0 \\ v_t - v_0 \end{pmatrix}, \tag{15}$$

which indicates that radial translational flows across the image are consistently oriented toward or away from the principal point $\boldsymbol{p}_0$, as illustrated in Fig. 2(a). Consequently, these flows are relevant not only with depth but also with pixel coordinates, with their magnitudes diminishing near $\boldsymbol{p}_0$ due to the reduced distance between $\boldsymbol{p}_t$ and $\boldsymbol{p}_0$. However, when both $\boldsymbol{T}^{\text{Tan}}$ and $\boldsymbol{T}^{\text{Rad}}$ are present, the resulting rigid flows can be expressed as follows:

$$\boldsymbol{f}_{t\to s}^{\text{Rig}} = \frac{1}{z_t + t_z} \begin{pmatrix} f_u t_x - (u_t - u_0) t_z \\ f_v t_y - (v_t - v_0) t_z \end{pmatrix}, \tag{16}$$

which obscures the distinct regularities inherent in both types of flows, thereby hindering the framework from discriminately backpropagating their corresponding supervisory signals to the network outputs. Specifically, the gradients propagated through $\boldsymbol{f}_{t\to s}^{\text{Rig}}$ are given by:

$$\frac{\partial \boldsymbol{f}_{t\to s}^{\text{Rig}}}{\partial z_t} = -\frac{1}{(z_t + t_z)^2} \begin{pmatrix} f_u t_x - (u_t - u_0) t_z \\ f_v t_y - (v_t - v_0) t_z \end{pmatrix}, \tag{17}$$

$$\frac{\partial \boldsymbol{f}_{t\to s}^{\text{Rig}}}{\partial t_x} = \frac{1}{z_t + t_z} \begin{pmatrix} f_u \\ 0 \end{pmatrix}, \tag{18}$$

$$\frac{\partial \boldsymbol{f}_{t\to s}^{\text{Rig}}}{\partial t_y} = \frac{1}{z_t + t_z} \begin{pmatrix} 0 \\ f_v \end{pmatrix}, \tag{19}$$

$$\frac{\partial \boldsymbol{f}_{t\to s}^{\text{Rig}}}{\partial t_z} = -\frac{1}{(z_t + t_z)^2} \begin{pmatrix} (u_t - u_0) z_t + f_u t_x \\ (v_t - v_0) z_t + f_v t_y \end{pmatrix}. \tag{20}$$

(17) indicates that z_t inherently receives smaller gradients near the principal point, which hinders consistent optimization across the image. Moreover, (18) and (19) show that the gradient of tangential translation components are influenced

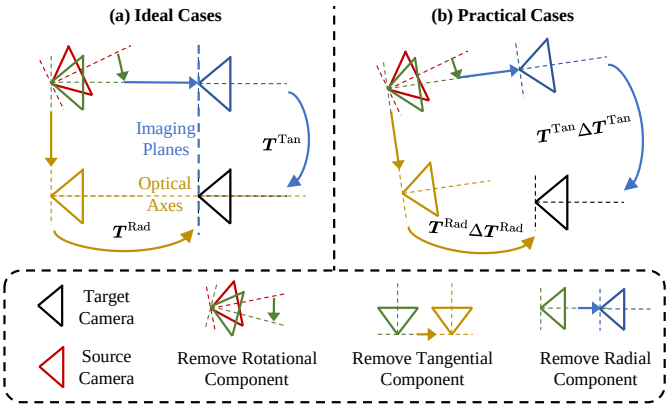

Fig. 3. Ego-motion transformation decomposition: (a) Ideally, ego-motion can be decomposed into pure tangential and radial translation components; (b) In practice, errors in the PoseNet predictions introduce undesired deviations to the decomposed components.

by the radial one, while (20) reveals the opposite influence. As a result, errors in one component can propagate to and impair the gradients for the others, ultimately hindering the convergence of PoseNet training.

To address these issues, after eliminating the irregular rotational flows, we further utilize $\hat{\boldsymbol{T}}^{\text{Tan}}$ and $\hat{\boldsymbol{T}}^{\text{Rad}}$ to decompose the remaining translational flows into the components that are geometrically regular but different with respect to depth, as illustrated in Fig. 2(a). The decomposed tangential and radial flows are then used to incorporate additional geometric constraints (detailed in Sect. V-C) into the joint learning framework, enabling independent refinement of depth and each translational component.

The above flow decomposition processes are performed by removing the rotational component from the full transformation $\boldsymbol{T}$, followed by the simultaneous and independent removal of the radial and tangential components. The two ideal decomposition processes, as illustrated in Fig. 3(a), can be formulated by rearranging (6) as follows:

$$\begin{cases} \boldsymbol{T}^{\text{Rad}} = \boldsymbol{T}\left(\boldsymbol{T}^{\text{Rot}}\right)^{-1}\left(\boldsymbol{T}^{\text{Tan}}\right)^{-1} \\ \boldsymbol{T}^{\text{Tan}} = \boldsymbol{T}\left(\boldsymbol{T}^{\text{Rot}}\right)^{-1}\left(\boldsymbol{T}^{\text{Rad}}\right)^{-1}. \end{cases} \tag{21}$$

However, in practice, errors in the ego-motion estimated by PoseNet are inevitable, which can introduce deviations $\Delta\boldsymbol{T}^{\text{Rad}}$ and $\Delta\boldsymbol{T}^{\text{Tan}}$ into the above two decomposition processes, respectively, resulting in the following expression:

$$\begin{cases} \boldsymbol{T}^{\text{Rad}}\Delta\boldsymbol{T}^{\text{Rad}} = \boldsymbol{T}\left(\hat{\boldsymbol{T}}^{\text{Rot}}\right)^{-1}\left(\hat{\boldsymbol{T}}^{\text{Tan}}\right)^{-1} \\ \boldsymbol{T}^{\text{Tan}}\Delta\boldsymbol{T}^{\text{Tan}} = \boldsymbol{T}\left(\hat{\boldsymbol{T}}^{\text{Rot}}\right)^{-1}\left(\hat{\boldsymbol{T}}^{\text{Rad}}\right)^{-1}. \end{cases} \tag{22}$$

Such deviations, shown in Fig. 3(b), can be derived from (22) and expressed as follows:

$$\begin{cases} \Delta\boldsymbol{T}^{\text{Rad}} = \begin{pmatrix} \Delta\boldsymbol{R}^{-1} & (\boldsymbol{I}-\Delta\boldsymbol{R}^{-1})\boldsymbol{t}^{\text{Tan}} - \Delta\boldsymbol{R}^{-1}\Delta\boldsymbol{t}^{\text{Tan}} \\ \boldsymbol{0}^{\top} & 1 \end{pmatrix} \\ \Delta\boldsymbol{T}^{\text{Tan}} = \begin{pmatrix} \Delta\boldsymbol{R}^{-1} & (\boldsymbol{I}-\Delta\boldsymbol{R}^{-1})\boldsymbol{t}^{\text{Rad}} - \Delta\boldsymbol{R}^{-1}\Delta\boldsymbol{t}^{\text{Rad}} \\ \boldsymbol{0}^{\top} & 1 \end{pmatrix} \end{cases}. \tag{23}$$

where $\Delta\boldsymbol{t}^{\text{Tan}} = \hat{\boldsymbol{t}}^{\text{Tan}} - \boldsymbol{t}^{\text{Tan}}$ and $\Delta\boldsymbol{t}^{\text{Rad}} = \hat{\boldsymbol{t}}^{\text{Rad}} - \boldsymbol{t}^{\text{Rad}}$ represent the translational deviations in $\hat{\boldsymbol{T}}^{\text{Tan}}$ and $\hat{\boldsymbol{T}}^{\text{Rad}}$, respectively, while $\Delta\boldsymbol{R} = \hat{\boldsymbol{R}}\boldsymbol{R}^{-1}$ denotes the rotational deviation. As shown in (23), $\Delta\boldsymbol{T}^{\text{Rad}}$ and $\Delta\boldsymbol{T}^{\text{Tan}}$ introduce not only the rotational deviation $\Delta\boldsymbol{R}$ but also the other type of translational component into $\boldsymbol{T}^{\text{Rad}}$ and $\boldsymbol{T}^{\text{Tan}}$, respectively. Consequently, as illustrated in Fig. 2(b), the resulting radial and tangential translational flows deviate from their expected ones, thereby impairing the effectiveness of incorporating additional geometric constraints derived from optical flow decomposition.

To address the aforementioned practical challenges, we leverage two geometric constraints (detailed in Sect. V-C) to jointly minimize the deviations between the decomposed translational flows and the expected ones. These two loss functions collaboratively guide the estimation of $\hat{\boldsymbol{T}}^{\text{Rot}}$, while simultaneously providing independent supervisory signals to $\hat{\boldsymbol{T}}^{\text{Tan}}$ and $\hat{\boldsymbol{T}}^{\text{Rad}}$.

In widely used driving datasets, such as KITTI [49] and nuScenes [50], vehicles predominantly move forward, with only gradual rotations occurring during turns. As a result, these datasets present limited challenges for accurate rotation estimation. Nevertheless, since $\boldsymbol{T}^{\text{Rad}}$ typically dominates over $\boldsymbol{T}^{\text{Tan}}$, even a minor rotational deviation $\Delta\boldsymbol{R}$ can induce a noticeable undesired deviation $(\boldsymbol{I}-\Delta\boldsymbol{R})\boldsymbol{t}^{\text{Rad}}$ in $\boldsymbol{t}^{\text{Tan}}$. To further demonstrate the robustness of the proposed method under more challenging conditions involving substantial residual rotations $\Delta\boldsymbol{R}$, we create a new dataset (detailed in Sect. VI-B) containing sequences with pronounced rotational motion and frequent camera shake.

V. DiMoDE

As shown in Fig. 3(a), removing the rotational and tangential components from the full transformation is equivalent to aligning the optical axes of the source and target cameras, while removing the rotational and radial components is equivalent to aligning their imaging planes. To realize the proposed discriminative treatment of motion components in practice, we incorporate such alignment processes into the joint learning framework, thereby introducing explicit geometric constraints that refine the outputs of both networks.

A. Framework Overview

An overview of the proposed DiMoDE framework is presented in Fig. 4. The ego-motion transformation $\hat{\boldsymbol{T}}$, estimated by PoseNet, is decomposed into three components: $\hat{\boldsymbol{T}}^{\text{Rot}}$, $\hat{\boldsymbol{T}}^{\text{Tan}}$ and $\hat{\boldsymbol{T}}^{\text{Rad}}$. First, $\hat{\boldsymbol{T}}^{\text{Rot}}$ is applied to the source camera to align its orientation with that of the target camera. Subsequently, $\hat{\boldsymbol{T}}^{\text{Tan}}$ and $\hat{\boldsymbol{T}}^{\text{Rad}}$ are used to align the optical axes and imaging planes, respectively. In accordance with these two alignment processes, the original optical flows provided by FlowNet are transformed to those from the target image to the aligned source camera image. To refine each component of the PoseNet output, the first set of geometric constraints is imposed by minimizing the deviations between the transformed optical flows and the expected pure radial and tangential ones. Once the PoseNet predictions achieve sufficient accuracy, ensuring proper alignment of optical axes and imaging planes,

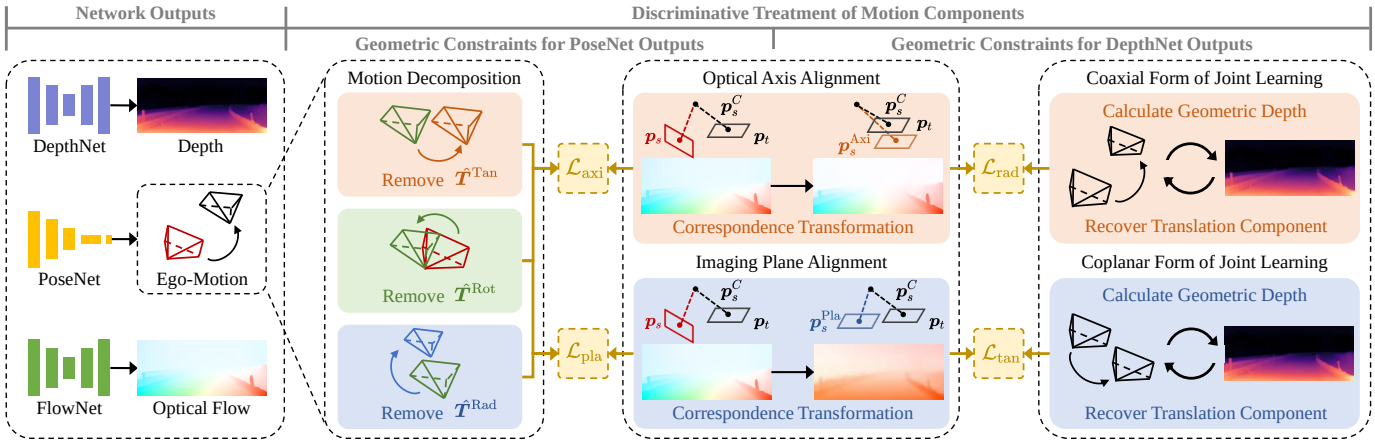

Fig. 4. An overview of the DiMoDE framework. Centered on the core idea of discriminative motion component treatment, depth (from DepthNet) and ego-motion (from PoseNet) are utilized to perform optical axis and imaging plane alignments. FlowNet generates dense correspondences, which are transformed during the alignment processes. The transformed flows are ultimately leveraged to incorporate two sets of geometric constraints into the unsupervised joint learning framework, thereby simultaneously improving both depth and pose estimation performance.

each translation component is incorporated individually to further refine the DepthNet outputs. In such cases, the depth and each translation component can be mutually determined through the decomposed flows, thereby forming constraint cycles that provide robust supervisory signals for DepthNet training under adverse conditions.

B. Optical Axis and Imaging Plane Alignments

To integrate the alignment processes and their corresponding geometric constraints into the joint learning framework, a straightforward approach, adopted in the previous study [11], is to reconstruct the warped image once after removing each motion component and then establish dense correspondences from the warped image pair. However, repeated image warping incurs substantial computational overhead and often introduces holes and artifacts in the reconstructed images [8], which undermine the reliability of correspondence generation and ultimately weaken the effectiveness of the imposed geometric constraints. To address this issue, we adopt a correspondence transformation strategy that directly transforms the original optical flow $\boldsymbol{f}_{t\to s}^{\text{Opt}}$ into the optical flows resulting from optical axis and imaging plane alignment, denoted as $\boldsymbol{f}_{t\to s}^{\text{Axi}}$ and $\boldsymbol{f}_{t\to s}^{\text{Pla}}$, respectively. Specifically, pixel correspondences $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$ are generated once using the target-to-source optical flow map $\boldsymbol{F}_{t\to s}^{\text{Opt}}$ predicted by FlowNet [51]. To backproject $\boldsymbol{p}_s$ into 3D camera coordinates, the source-view depth map $\boldsymbol{D}_s$ is warped into the target view based on the optical flow map, producing the following pixel-aligned depth map [23], [40]:

$$\boldsymbol{D}_{s\to t} = \mathcal{W}\left(\boldsymbol{D}_s, \boldsymbol{F}_{t\to s}^{\text{Opt}}\right). \tag{24}$$

$\boldsymbol{p}_s^C$, the 3D camera coordinates of $\boldsymbol{p}_s$, can be computed using the following expression:

$$\boldsymbol{p}_s^C = \boldsymbol{D}_{s\to t}(\boldsymbol{p}_t)\boldsymbol{K}^{-1}\boldsymbol{p}_s. \tag{25}$$

Transforming $\boldsymbol{p}_s^C$ through the alignment processes and projecting it back into pixel coordinates yields the following expressions:

$$\begin{cases} \boldsymbol{p}_s^{\text{Pla}} \simeq \boldsymbol{K}\left(\hat{\boldsymbol{R}}^{-1}\boldsymbol{p}_s^C - \hat{\boldsymbol{t}}^{\text{Rad}}\right) \\ \boldsymbol{p}_s^{\text{Axi}} \simeq \boldsymbol{K}\left(\hat{\boldsymbol{R}}^{-1}\boldsymbol{p}_s^C - \hat{\boldsymbol{t}}^{\text{Tan}}\right) \end{cases}. \tag{26}$$

Finally, the optical flows $\boldsymbol{f}_{t\to s}^{\text{Pla}}$ and $\boldsymbol{f}_{t\to s}^{\text{Axi}}$ are derived as follows:

$$\begin{cases} \boldsymbol{f}_{t\to s}^{\text{Pla}} = \boldsymbol{p}_s^{\text{Pla}} - \boldsymbol{p}_t = \dfrac{\boldsymbol{K}\left(\hat{\boldsymbol{R}}^{-1}\boldsymbol{p}_s^C - \hat{\boldsymbol{t}}^{\text{Rad}}\right)}{\boldsymbol{i}_3^\top\left(\hat{\boldsymbol{R}}^{-1}\boldsymbol{p}_s^C - \hat{\boldsymbol{t}}^{\text{Rad}}\right)} - \boldsymbol{p}_t \\ \boldsymbol{f}_{t\to s}^{\text{Axi}} = \boldsymbol{p}_s^{\text{Axi}} - \boldsymbol{p}_t = \dfrac{\boldsymbol{K}\left(\hat{\boldsymbol{R}}^{-1}\boldsymbol{p}_s^C - \hat{\boldsymbol{t}}^{\text{Tan}}\right)}{\boldsymbol{i}_3^\top\left(\hat{\boldsymbol{R}}^{-1}\boldsymbol{p}_s^C - \hat{\boldsymbol{t}}^{\text{Tan}}\right)} - \boldsymbol{p}_t \end{cases}. \tag{27}$$

C. Geometric Constraints from Motion Components

As discussed in Sect. IV, $\boldsymbol{f}_{t\to s}^{\text{Pla}}$ and $\boldsymbol{f}_{t\to s}^{\text{Axi}}$ are expected to approximate the tangential and radial translational flows. The former should exhibit a uniform direction across the image, while the latter should point toward or away from the principal point at each pixel. Based on (12) and (15), these properties enable the formulation of two geometric constraints: (1) all $\boldsymbol{f}_{t\to s}^{\text{Pla}}$ vectors form a consistent angle with a global reference vector $\boldsymbol{e}_1$, such that $\frac{\boldsymbol{e}_1^\top \boldsymbol{f}_{t\to s}^{\text{Pla}}}{\left\|\boldsymbol{f}_{t\to s}^{\text{Pla}}\right\|_2}$ remains constant across the image, where $\boldsymbol{e}_k \in \mathbb{R}^2$ denotes the k-th column vector of the 2×2 identity matrix; (2) each $\boldsymbol{f}_{t\to s}^{\text{Axi}}$ vector is expected to be parallel to the direction from its pixel $\boldsymbol{p}_t$ to the principal point $\boldsymbol{p}_0$, such that $\frac{\boldsymbol{f}_{t\to s}^{\text{Axi}}}{\left\|\boldsymbol{f}_{t\to s}^{\text{Axi}}\right\|_2} = \frac{\boldsymbol{p}_t - \boldsymbol{p}_0}{\left\|\boldsymbol{p}_t - \boldsymbol{p}_0\right\|_2}$. In practice, these two geometric constraints are imposed by incorporating the following loss functions into the joint learning framework:

$$\mathcal{L}_{\text{pla}} = \text{Var}\left(\arccos\frac{\boldsymbol{e}_1^\top \boldsymbol{f}_{t\to s}^{\text{Pla}}}{\left\|\boldsymbol{f}_{t\to s}^{\text{Pla}}\right\|_2}\right), \tag{28}$$

and

$$\mathcal{L}_{\text{axi}} = \mathbb{E}\left[\arccos \frac{(\boldsymbol{f}^{\text{Axi}}_{t\to s})^\top (\boldsymbol{p}_t - \boldsymbol{p}_0)}{\|\boldsymbol{f}^{\text{Axi}}_{t\to s}\|_2 \, \|\boldsymbol{p}_t - \boldsymbol{p}_0\|_2}\right]. \tag{29}$$

where the mean $\mathbb{E}[\cdot]$ and variance $\text{Var}(\cdot)$ are computed over all valid pixels in static and non-occluded regions, as determined by the masking technique proposed in the study [39]. Continuously imposing these constraints through (28) and (29) enables more accurate imaging plane and optical axis alignments, respectively. The radial translation (used exclusively for the imaging plane alignment) and the tangential translation (used solely for the optical axes alignment) are therefore optimized independently, while the rotation estimation is jointly constrained by both. The effectiveness of the imposed constraints and the complementary role of (28) and (29) are validated through an ablation study detailed in Sect. VI-F.

Once the PoseNet predictions reach sufficient accuracy, the alignment processes allow the original joint depth and ego-motion learning from monocular video to be reformulated into two geometrically simplified forms: the coaxial form, analogous to learning from axial-motion image pairs, and the coplanar form, analogous to learning from stereo image pairs along the horizontal and vertical directions. In such forms, the depth at each pixel and the ego-motion satisfy the following relations, derived from (12) and (15):

$$\rho_x := \frac{z_t}{t_x} = \frac{f_x}{\boldsymbol{e}_1^\top \boldsymbol{f}^{\text{Pla}}_{t\to s}}, \quad \rho_y := \frac{z_t}{t_y} = \frac{f_y}{\boldsymbol{e}_2^\top \boldsymbol{f}^{\text{Pla}}_{t\to s}}, \tag{30}$$

$$\rho_z := \frac{z_t}{t_z} = -\frac{(\boldsymbol{p}_t - \boldsymbol{p}_0)^\top (\boldsymbol{f}^{\text{Axi}}_{t\to s} + \boldsymbol{p}_t - \boldsymbol{p}_0)}{(\boldsymbol{p}_t - \boldsymbol{p}_0)^\top \boldsymbol{f}^{\text{Axi}}_{t\to s}}, \tag{31}$$

where ρ_x, ρ_y, and ρ_z are defined as the per-pixel ratios between depth and the corresponding translation components. These ratios are determined by the transformed optical flows, with ρ_z additionally depending on the pixel coordinates. According to (30) and (31), at each pixel $\boldsymbol{p}_t$, multiplying the respective translation components by their corresponding ratios yields three geometric depth values: $\rho_x\hat{t}_x$, $\rho_y\hat{t}_y$, and $\rho_z\hat{t}_z$. These values are expected to be consistent with the depth $\hat{z}_t$ predicted by DepthNet, thereby imposing three per-pixel geometric constraints to supervise DepthNet training. Moreover, the simplified geometric relationships enable closed-form inverse computation: given the estimated depth and three ratio maps, each translation component can be recovered as $\mathbb{E}[\hat{z}_t/\rho_x]$, $\mathbb{E}[\hat{z}_t/\rho_y]$, and $\mathbb{E}[\hat{z}_t/\rho_z]$. This eliminates the need for iterative optimization. The recovered translation components are also expected to be consistent with those estimated by PoseNet. Based on these constraints, we introduce the following two loss functions to the joint learning framework:

$$\begin{aligned}\mathcal{L}_{\text{tan}} =& \mathbb{E}\left[\frac{|\rho_x\hat{t}_x - \hat{z}_t|}{\hat{z}_t}\right] + \left|\frac{\mathbb{E}\left[\hat{z}_t/\rho_x\right] - \hat{t}_x}{\hat{t}_x}\right| \\ &+ \mathbb{E}\left[\frac{|\rho_y\hat{t}_y - \hat{z}_t|}{\hat{z}_t}\right] + \left|\frac{\mathbb{E}\left[\hat{z}_t/\rho_y\right] - \hat{t}_y}{\hat{t}_y}\right|.\end{aligned} \tag{32}$$

and

$$\mathcal{L}_{\text{rad}} = \mathbb{E}\left[\frac{|\rho_z\hat{t}_z - \hat{z}_t|}{\hat{z}_t}\right] + \left|\frac{\mathbb{E}\left[\hat{z}_t/\rho_z\right] - \hat{t}_z}{\hat{t}_z}\right|, \tag{33}$$

By eliminating the pixel-coordinate dependence of $\boldsymbol{f}^{\text{Axi}}_{t\to s}$ via (31), the resulting ρ_z becomes a function of depth along. Consequently, the first term in both (32) and (33) can provide consistent gradients with respect to per-pixel depth errors across the entire map. Furthermore, unlike existing methods that impose only pixel-level constraints into DepthNet training [3], [39], [41], the proposed method establishes two constraint cycles via (32) and (33), within which the second terms in both (32) and (33), along with the fourth term in (32), measure the overall depth errors aggregated over all valid pixels. Consequently, in challenging scenarios, such as night-time or adverse weather conditions, where per-pixel supervisory signals may become unreliable, this constraint cycle can help prevent significant degradation in DepthNet performance. To validate its effectiveness, we conduct extensive comparative experiments and ablation studies on the nuScenes dataset [50] (see Sects. VI-E and VI-F), which contains numerous video sequences captured under the above-mentioned challenging conditions.

In addition to the above newly proposed loss functions, we retain the minimum photometric reprojection loss $\mathcal{L}_{\text{pho}}$ adopted in the study [2], as well as the local structure refinement loss $\mathcal{L}_{\text{loc}}$ employed in the study [3] to establish the baseline training framework. The overall loss function is formulated as:

$$\mathcal{L} = \lambda_1(\mathcal{L}_{\text{axi}} + \mathcal{L}_{\text{pla}}) + \lambda_2(\mathcal{L}_{\text{rad}} + \mathcal{L}_{\text{tan}}) + \lambda_3\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{pho}}, \tag{34}$$

where λ_1, λ_2, and λ_3 are weighting coefficients that balance the contributions of the corresponding loss terms. The specific weight settings and the overall training strategy are detailed in Sect. VI-A, and the effect of varying weighting coefficients is further analyzed in Sect. VI-F.

VI. Experiments

Sect. VI-A provides implementation details on the proposed DiMoDE framework and the adopted training settings. Sect. VI-B introduces the six public datasets used for performance comparison, along with a newly created real-world dataset designed to evaluate model performance in complex scenarios. Evaluation metrics are detailed in Sect. VI-C. Sect. VI-D and Sect. VI-E present comprehensive comparisons between our method and previous SoTA approaches in monocular visual odometry and depth estimation, respectively. Finally, Sect. VI-F provides detailed ablation studies to evaluate the efficacy of each innovative design based on discriminative motion component treatment, and details hyperparameter selection.

A. Implementation Details

As DiMoDE is designed as a general framework rather than being tied to a specific network architecture, we evaluate its compatibility with a variety of existing depth and pose estimation networks in our experiments. For DepthNet, we adopt Lite-Mono [4], a lightweight CNN-Transformer hybrid network, as well as D-HRNet [36], a representative CNN-based network, to evaluate DiMoDE's adaptability to different depth estimation network architectures. Given the higher computational cost of D-HRNet, the majority of our

experiments are conducted using Lite-Mono. For PoseNet, we adopt a lightweight design based on a ResNet-18 [29] encoder, following previous studies [2], [23]. To further validate the effectiveness of DiMoDE, we also employ the PoseNet proposed in the study [26], which incorporates spatial clues into the pose estimation process. However, since the latter PoseNet [26] is significantly more computationally intensive and does not support real-time inference, we primarily adopt the basic ResNet-based PoseNet [2], [23] in our experiments. Finally, RAFT [51] is employed as the FlowNet in DiMoDE to generate dense correspondences.

All networks are trained on an NVIDIA RTX 4090 GPU. To fully utilize GPU memory, the batch size is set to 8 for images with a resolution of 640×192 pixels. For higher resolutions, such as 640×384 and 512×288, the batch size is reduced to 4 to accommodate memory constraints. Following the study [2], we utilize a snippet of three consecutive video frames as a training sample. For data augmentation, random color jitter and horizontal flips are applied to the images during model training. To minimize loss functions, we use the AdamW [52] optimizer with an initial learning rate of 10^{-4} and a weight decay of 10^{-2}.

The encoders of DepthNet and PoseNet are initialized with pretrained weights from the ImageNet database [53], following the studies [2], [4]. FlowNet is pretrained on the synthetic FlyingChairs dataset [54], following the study [51]. In the early training stage, network predictions tend to be less reliable. Directly enforcing optical axis and imaging plane alignments under such conditions may introduce significant deviations and lead to unstable training. To mitigate this problem, we adopt a progressive training strategy. Specifically, during the first 5 epochs out of the total 30 (Stage 1), we jointly train the three networks in an unsupervised manner using only $\mathcal{L}_{\text{pho}}$ and $\mathcal{L}_{\text{loc}}$, where the weights $(\lambda_1, \lambda_2, \lambda_3)$ in (34) are set to $(0, 0, 0.01)$. In the subsequent 10 epochs (Stage 2), we freeze FlowNet and incorporate the alignment processes to impose the first set of geometric constraints, namely $\mathcal{L}_{\text{pla}}$ and $\mathcal{L}_{\text{axi}}$, primarily to refine pose estimation. The weights $(\lambda_1, \lambda_2, \lambda_3)$ are updated to $(0.05, 0, 0.01)$. Finally, during the last 15 epochs (Stage 3), we introduce the second set of geometric constraints, namely $\mathcal{L}_{\text{tan}}$ and $\mathcal{L}_{\text{rad}}$, to further improve depth estimation performance, with the weights $(\lambda_1, \lambda_2, \lambda_3)$ set to $(0.05, 0.1, 0.01)$.

B. Datasets

The performance of our method on visual odometry is evaluated using both the KITTI Odometry dataset [1] and our newly created real-world MIAS-Odom dataset. The KITTI Odometry dataset contains 22 video sequences in a driving scenario, with ground-truth trajectories available for Seqs. 00-10. Following prior works [2], [23], we use Seqs. 00–08 for unsupervised framework training. Seqs. 09-10 serve as the standard test set, while Seqs. 11–21 are additionally included in this study to further evaluate the model's generalization and convergence performance, which are often overlooked in prior research. Since ground-truth trajectories of Seqs. 11–21 are not publicly available, we generate pseudo ground-truth trajectories using ORB-SLAM3 (Stereo) [16] for performance evaluation. To further validate the practical effectiveness of our method, we create the MIAS-Odom dataset, collected using a handheld setup, as illustrated in Fig. 5. This dataset consists of six (three indoor and three outdoor) video sequences with a total of 11,608 images captured under challenging conditions that are rarely covered by public datasets. These conditions include large rotational motions and frequent camera shakes, which typically lead to substantial rotational errors, as well as overexposure, poor illumination, and motion blur, all of which degrade the reliability of per-pixel photometric supervision. For ground-truth trajectory acquisition, we construct a system consisting of multiple hardware-synchronized sensors, including binocular cameras, a Livox MID-360 solid-state LiDAR, and an onboard IMU. Camera trajectories are obtained using a tightly coupled LiDAR-inertial odometry framework [55] that fuses LiDAR features with IMU measurements to achieve accurate and robust pose estimation.

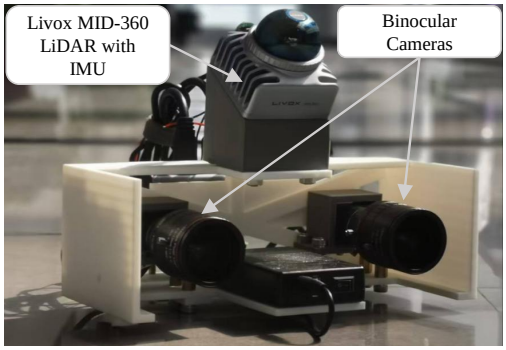

Fig. 5. An illustration of our designed handheld setup equipped with calibrated and synchronized sensors for accurate real-world data collection.

Depth estimation performance is evaluated on five public datasets: KITTI [49], DDAD [56], nuScenes [57], Make3D [57], and DIML [58]. For the KITTI dataset, we adopt the Eigen split [5], which comprises 39,180 monocular triplets for training, 4,424 images for validation, and 697 images for testing. The DDAD dataset is split into a training set of 12,650 images and a test set of 3,950 images. For the nuScenes dataset, we use 79,760 image triplets for training, and evaluate the model's performance on 6,019 front-camera images. As the Make3D and DIML datasets do not provide either stereo image pairs or monocular sequences for unsupervised model training, they are used exclusively to quantify the model's generalizability.

C. Evaluation Metrics

Following the previous study [66], we adopt the average translational error e_t (%), average rotational error e_r ($^\circ$/100m), and the absolute trajectory error (ATE, in meters) to quantify pose estimation performance. Additionally, adhering to the study [2], we quantify depth estimation performance using the mean absolute relative error (Abs Rel), mean squared relative error (Sq Rel), root mean squared error (RMSE), root mean squared log error (RMSE log), and the accuracy under specific thresholds ($\delta_i < 1.25^i$, where $i = 1, 2, 3$).

TABLE I
QUANTITATIVE VISUAL ODOMETRY RESULTS ON THE TEST SETS OF THE KITTI ODOMETRY DATASET [1]. THE BEST RESULTS WITHIN EACH METHOD CATEGORY ARE SHOWN IN BOLD TYPE. 'S' AND 'M' INDICATE TRAINING WITH STEREO IMAGE PAIRS AND MONOCULAR VIDEO SEQUENCES, RESPECTIVELY. RESULTS FOR [15], [16] ARE REPORTED IN MONOCULAR SETTINGS WITH LOOP CLOSURE ENABLED.

Methods	Year	Data	Seq. 09			Seq. 10		
			e_t	e_r	ATE	e_t	e_r	ATE
Geometric-Based Methods								
DSO [17]	2017	–	–	–	52.22	–	–	11.09
ORB-SLAM2 [15]	2017	–	2.88	**0.25**	8.39	3.30	**0.30**	6.63
ORB-SLAM3 [16]	2021	–	**2.35**	0.28	**6.06**	**2.24**	0.31	**4.97**
Hybrid Methods								
DVSO [59]	2018	S	0.83	**0.21**	–	0.74	**0.21**	–
D3VO [27]	2020	S	**0.78**	–	–	**0.62**	–	–
TrianFlow [60]	2020	M	6.93	0.44	–	4.66	0.62	–
pRGBD [28]	2020	M	4.20	1.00	11.97	4.40	1.60	6.35
DF-VO [12]	2021	M	2.40	0.24	8.36	1.82	0.38	3.13
SC-Depth [23]	2021	M	5.08	1.05	13.40	4.32	2.34	7.99
Sun *et al.* [18]	2022	M	–	–	45.95	–	–	8.11
Wang *et al.* [61]	2024	M	1.86	0.29	**6.06**	**1.55**	**0.31**	**2.64**
Learning-Based Methods								
Zhan *et al.* [62]	2018	S	11.89	3.60	52.12	12.82	3.41	24.70
UndeepVO [22]	2018	S	7.01	3.60	–	10.63	4.60	–
SfMLearner [8]	2017	M	19.15	6.82	77.79	40.40	17.69	67.34
Monodepth2 [2]	2019	M	17.17	3.85	76.22	11.68	5.31	20.35
LTMVO [10]	2020	M	3.49	1.00	11.30	5.81	1.80	11.80
SC-Depth [23]	2021	M	7.31	3.05	23.56	7.79	4.90	12.00
MLF-VO [25]	2022	M	3.90	1.41	9.86	4.88	1.38	7.36
MotionHint [63]	2022	M	8.18	1.50	17.82	8.10	2.74	11.63
SCIPaD [26]	2024	M	7.43	2.46	26.15	9.82	3.87	15.51
Lite-SVO [64]	2024	M	7.20	1.50	33.82	8.49	2.40	15.45
Manydepth2 [65]	2025	M	7.01	1.76	–	7.29	2.65	–
DiMoDE (Ours)	–	M	**2.86**	**0.74**	**9.83**	**4.20**	**1.22**	**5.81**

Methods	Year	Data	Seqs. 11-20			Seq. 21		
			e_t	e_r	ATE	e_t	e_r	ATE
Geometric-Based/Hybrid Methods								
ORB-SLAM3 [16]	2021	–	**4.05**	**0.38**	**13.46**	94.16	2.38	829.27
TrianFlow [60]	2020	M	18.77	5.57	76.96	**17.56**	**1.99**	**256.09**
DF-VO [12]	2021	M	11.39	3.91	62.11	43.67	13.84	1491.25
Learning-Based Methods								
Monodepth2 [2]	2019	M	16.16	5.64	53.26	16.88	2.57	526.75
SC-Depth [23]	2021	M	15.52	5.43	55.31	44.79	4.50	800.90
SCIPaD [26]	2024	M	7.98	3.65	35.78	16.27	2.40	**510.05**
Manydepth2 [65]	2025	M	12.91	4.78	39.82	23.79	4.01	571.64
DiMoDE (Ours)	–	M	7.60	**2.80**	**26.64**	16.69	**2.32**	526.86

D. Monocular Visual Odometry Results

This subsection presents a comprehensive evaluation of monocular visual odometry performance. We compare DiMoDE with a wide range of SoTA approaches, including learning-based methods [2], [8], [10], [22], [23], [25], [26], [62]–[64], geometry-based methods [15]–[17], and hybrid methods [12], [18], [23], [27], [28], [59]–[61].

Quantitative results on the KITTI Odometry dataset [1] are presented in Table I. DiMoDE achieves SoTA performance among learning-based methods on the standard test set (Seqs. 09 and 10), exhibiting the lowest trajectory drift (see Fig. 6). However, evaluation on Seqs. 09-10 alone is insufficient to fully demonstrate the model's generalizability

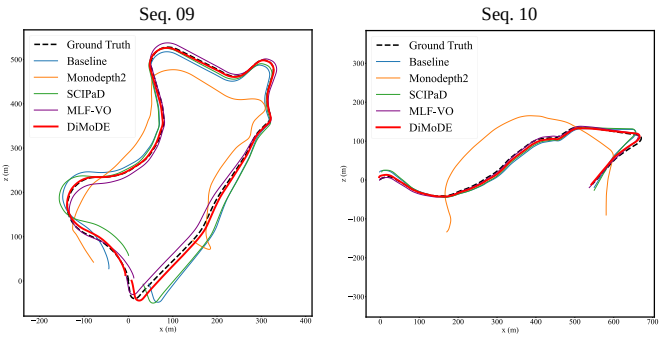

Fig. 6. Comparisons of trajectories on the test sets of the KITTI Odometry dataset [1]. All predicted trajectories are aligned with the ground truth using a 7-DoF similarity transformation.

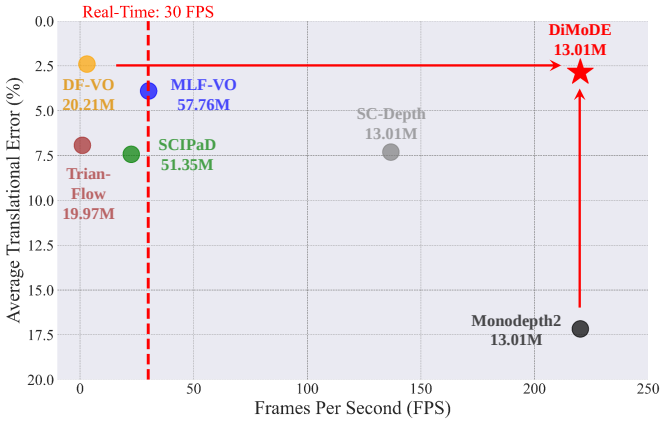

Fig. 7. Comparison of storage overhead and inference speed among representative learning-based and hybrid visual odometry methods.

to unseen environments. To this end, we conduct additional comparative experiments on the remaining 11 sequences (Seqs. 11-21) in the dataset. Across these sequences, DiMoDE consistently achieves the highest average accuracy among all learning-based approaches. The results for Seq. 21 are reported separately due to its extreme difficulty, which causes all existing methods to fail. Furthermore, DiMoDE delivers competitive accuracy compared to geometry-based and hybrid approaches [12], [16], significantly narrowing the performance gap between fully learning-based models and those relying on iterative optimization.

Moreover, we compare representative learning-based and hybrid methods in terms of storage overhead and inference speed, two crucial factors for practical deployment that have received relatively limited attention in previous studies. As shown in Fig. 7, although hybrid methods [12], [60] generally outperform learning-based approaches due to the incorporation of back-end optimization, this advantage comes at the expense of significantly reduced inference speed, making them less suitable for real-world applications. Previous learning-based approaches [10], [25], [26] adopt increasingly sophisticated PoseNet architectures for improved accuracy. Nevertheless, such designs also incur substantial computational and memory

TABLE II

QUANTITATIVE VISUAL ODOMETRY RESULTS ON SEQS. 00, 02, 05, AND 08 OF THE KITTI ODOMETRY DATASET [1]. THE BEST RESULTS WITHIN EACH METHOD CATEGORY ARE SHOWN IN BOLD TYPE. 'M' INDICATES TRAINING ON MONOCULAR VIDEO SEQUENCES. RESULTS FOR [15] ARE REPORTED IN MONOCULAR SETTINGS WITH LOOP CLOSURE ENABLED.

Methods	Year	Data	Seq. 00			Seq. 02		
			e_t	e_r	ATE	e_t	e_r	ATE
Geometric-Based/Hybrid Methods								
ORB-SLAM3 [16]	2021	–	3.13	**0.46**	**8.70**	5.10	0.55	25.11
DF-VO [12]	2021	M	**2.33**	0.63	14.45	**3.24**	**0.49**	**19.69**
Learning-Based Methods								
Monodepth2 [2]	2019	M	11.40	3.26	107.51	6.42	1.51	83.63
SC-Depth [23]	2021	M	11.01	3.39	93.04	6.74	1.96	70.37
SCIPaD [26]	2024	M	6.17	2.28	43.17	4.79	1.66	50.72
Manydepth2 [65]	2025	M	10.12	3.59	72.42	14.30	3.31	94.70
DiMoDE (Ours)	–	M	**2.28**	**0.51**	**10.77**	**2.32**	**0.56**	**17.60**

Methods	Year	Data	Seq. 05			Seq. 08		
			e_t	e_r	ATE	e_t	e_r	ATE
Geometric-Based/Hybrid Methods								
ORB-SLAM3 [16]	2021	–	3.03	0.43	6.57	15.02	**0.28**	50.28
DF-VO [12]	2021	M	**1.09**	**0.25**	**3.61**	**2.18**	0.32	**7.63**
Learning-Based Methods								
Monodepth2 [2]	2019	M	7.21	3.11	42.35	7.08	2.44	56.35
SC-Depth [23]	2021	M	6.70	2.38	40.56	8.11	2.61	56.15
SCIPaD [26]	2024	M	3.77	1.75	13.51	4.58	1.83	22.03
Manydepth2 [65]	2025	M	6.06	2.52	29.04	8.56	3.02	61.90
DiMoDE (Ours)	–	M	**1.63**	**0.62**	**6.47**	**2.10**	**0.65**	**10.01**

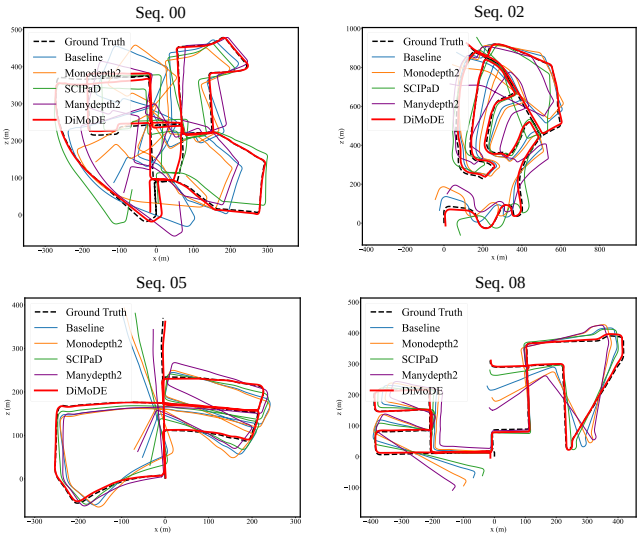

Fig. 8. Comparisons of trajectories on several long-term sequences in the train set of the KITTI Odometry dataset [1]. All predicted trajectories are aligned with the ground truth using a 7-DoF similarity transformation.

costs, making real-time inference infeasible, even on high-end GPUs. In contrast, DiMoDE achieves competitive performance with SoTA hybrid methods while relying solely on a lightweight PoseNet architecture, thereby satisfying real-time inference requirements.

To validate whether the DiMoDE framework provides more effective supervisory signals for pose estimation, we report the results on several long-term sequences (Seqs. 00, 02, 05, and 08) from the training set. As shown in Table II, previous methods, limited by the accuracy of their estimated ego-motion components, tend to accumulate significant errors over these long-term sequences, even when trained on them. In contrast, DiMoDE significantly alleviates this issue, outperforming the previous SoTA method, with ATE reductions of 75.05%, 65.30%, 52.11%, and 54.56%, respectively. Moreover, DiMoDE achieves performance competitive with approaches that rely on iterative optimization. Qualitative comparisons presented in Fig. 8 show that DiMoDE exhibits the least trajectory drift, highlighting its effectiveness in reducing error accumulation over complex and long-term trajectories.

Furthermore, a key strength of unsupervised learning-based

TABLE III

QUANTITATIVE VISUAL ODOMETRY RESULTS ON SEQS. 11-20 OF THE KITTI ODOMETRY DATASET [1]. THE BEST RESULTS ARE SHOWN IN BOLD TYPE. FOR FAIR COMPARISONS, THE DEPTHNET AND FLOWNET IN DF-VO [12], AS WELL AS THE POSENET IN ALL LEARNING-BASED METHODS, ARE FINETUNED FOR THE SAME NUMBER OF ITERATIONS ON THESE SEQUENCES.

Methods	Seq.11	Seq. 12	Seq. 13	Seq. 14	Seq. 15
	ATE	ATE	ATE	ATE	ATE
DF-VO [12]	4.45	50.54	17.24	3.04	6.57
SC-Depth [23]	24.55	14.08	47.30	9.26	41.72
SCIPaD [26]	12.94	22.50	39.96	**1.91**	20.46
Manydepth2 [65]	10.04	20.30	73.38	5.56	11.96
DiMoDE (Ours)	**4.26**	**12.68**	**8.23**	2.17	**9.74**

Methods	Seq.16	Seq. 17	Seq. 18	Seq. 19	Seq. 20
	ATE	ATE	ATE	ATE	ATE
DF-VO [12]	**4.88**	10.31	13.60	58.89	16.40
SC-Depth [23]	16.96	7.40	53.67	202.96	14.86
SCIPaD [26]	14.99	4.97	36.65	150.58	11.92
Manydepth2 [65]	28.51	6.92	15.11	124.18	15.83
DiMoDE (Ours)	6.17	**1.55**	**9.97**	**25.23**	**8.78**

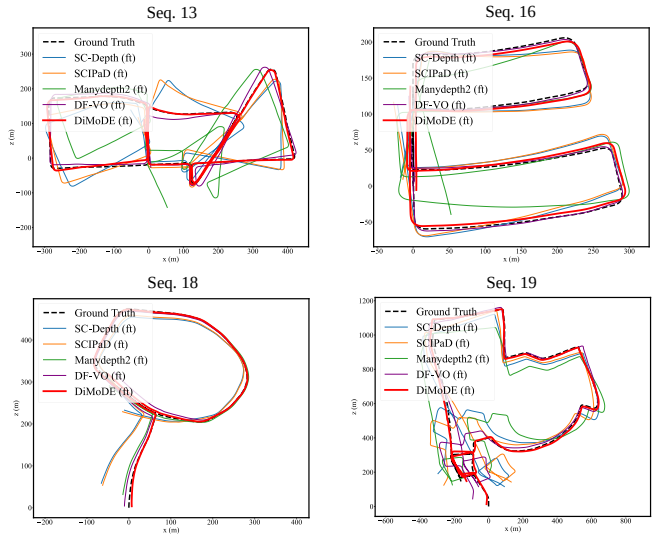

Fig. 9. Comparisons of trajectories produced by models finetuned on the additional sequences of the KITTI Odometry dataset ('ft' denotes finetuned models). All predicted trajectories are aligned with the ground truth using a 7-DoF similarity transformation.

TABLE IV
QUANTITATIVE VISUAL ODOMETRY RESULTS ON OUR NEWLY CREATED MIAS-ODOM DATASET. THE BEST RESULTS ARE SHOWN IN BOLD TYPE. FOR FAIR COMPARISONS, THE DEPTHNET AND FLOWNET IN DF-VO [12], AS WELL AS THE POSENET IN ALL LEARNING-BASED METHODS, ARE FINETUNED FOR THE SAME NUMBER OF ITERATIONS ON THESE SEQUENCES.

Methods	Indoor Seq.00 ATE	Indoor Seq. 01 ATE	Indoor Seq. 02 ATE
DF-VO [12]	3.53	3.61	5.94
SC-Depth [23]	9.02	5.20	10.67
SCIPaD [26]	5.83	5.42	8.82
Manydepth2 [65]	7.67	4.51	11.21
DiMoDE (Ours)	**1.48**	**3.02**	**5.33**

Methods	Outdoor Seq.00 ATE	Outdoor Seq. 01 ATE	Outdoor Seq. 02 ATE
DF-VO [12]	4.43	4.35	11.07
SC-Depth [23]	6.74	27.63	11.23
SCIPaD [26]	7.32	8.02	9.70
Manydepth2 [65]	6.36	32.92	24.31
DiMoDE (Ours)	**2.71**	**2.22**	**6.48**

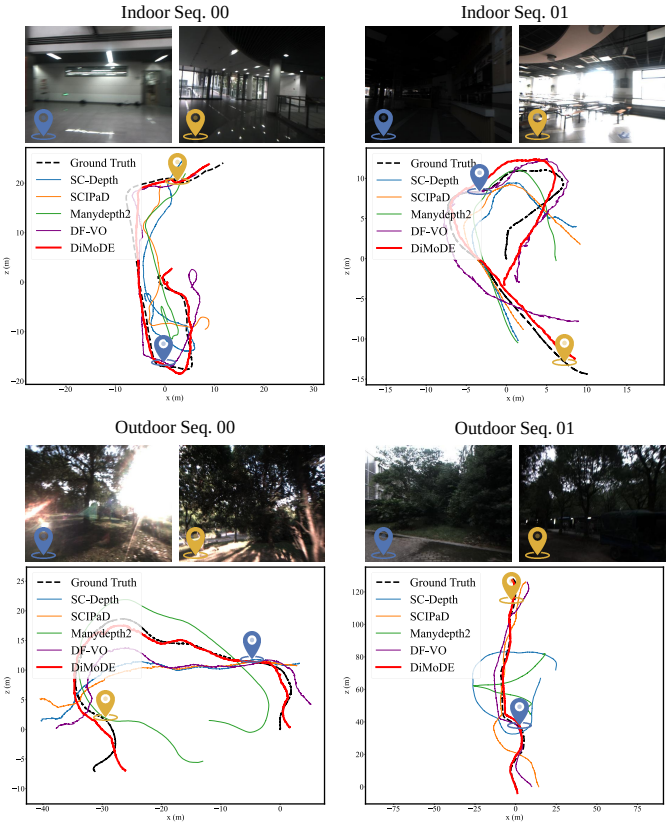

Fig. 10. Comparisons of trajectories produced by models finetuned on the newly created MIAS-Odom dataset. All predicted trajectories are aligned with the ground truth using a 7-DoF similarity transformation.

approaches lies in their inherent ability to adapt the model to previously unseen environments, which is crucial for real-world applications. To validate the effectiveness of the method in this regard, we conduct additional finetuning experiments on Seqs. 11–21 of the KITTI Odometry dataset and on our MIAS-Odom dataset, respectively, by training each model for five additional epochs. As shown in Table III and Fig. 9, while the other methods [23], [26], [65] continue to yield unsatisfactory results after finetuning on Seqs. 11–21 of the KITTI Odometry dataset, DiMoDE enables the estimated trajectories to align closely with the pseudo ground truth. This improvement corroborates our analyses in Sect. IV and demonstrates that decomposing translational components effectively mitigates error propagation and improves convergence, thereby enabling more effective adaptation to previously unseen environments. As shown in Table IV and Fig. 10, DiMoDE demonstrates significantly more robust ego-motion estimation on the MIAS-Odom dataset, effectively handling a variety of real-world challenges, such as overexposure, low illumination, severe motion blur, frequent camera shake, and large-degree turns. In contrast, prior methods consistently fail under these adverse conditions. However, our experimental results suggest that such satisfactory results can only be achieved when the challenging sequences are included in the training set. When a portion or all of these sequences are reserved exclusively for testing, all models, including DiMoDE, fail to generalize effectively. This generalization limitation stands in sharp contrast to the results observed on the KITTI Odometry dataset, and is further analyzed in Sect. VII.

E. Monocular Depth Estimation Results

This subsection presents comprehensive experimental results to evaluate the performance of DiMoDE in monocular depth estimation. We compare our method with representative single-frame approaches [3], [4], [36], [39]–[41], multi-frame approaches [65], [67], [68], and robust depth estimation approaches [42], [43]. All the compared methods provide publicly available implementations, allowing us to conduct additional training and evaluation on the DDAD [56] and nuScenes [50] datasets.

As shown in Table V, our method achieves competitive performance but does not surpass existing SoTA approaches on the KITTI dataset. This is primarily because our method is designed to impose additional constraints for robust joint learning under adverse conditions and to address the inconsistent optimization issue discussed in (17). However, the KITTI dataset is collected under clear daytime conditions, with few close-range objects near the principal point. As such, it lacks the types of challenges that DiMoDE is specifically designed to handle, thereby limiting the extent to which its advantages can be fully demonstrated.

To further validate the effectiveness of the proposed method, we conduct experiments on the DDAD and nuScenes datasets, which contain more challenging scenarios, such as varying illumination, adverse weather, and structurally complex environments. As shown in Tables VI and VII, DiMoDE significantly outperforms previous leading methods [3], [39], [41]. By comparison, several sophisticated networks [4], [36], [40], [65], [68], despite their strong performance on the KITTI dataset, exhibit reduced accuracy or even fail to converge on these more challenging datasets. This performance drop stems from their exclusive reliance on pixel-level supervisory

TABLE V

QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE KITTI [49] DATASET. THE BEST RESULTS FOR SINGLE-FRAME AND MULTI-FRAME APPROACHES ARE SHOWN IN BOLD TYPE, RESPECTIVELY. THE SYMBOLS ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER PERFORMANCE, RESPECTIVELY.

Method	Year	Test frames	Resolution (pixels)	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Multi-Frame Methods										
Manydepth [67]	2021	2(-1, 0)	640 × 192	0.098	0.770	4.459	0.176	0.900	0.965	0.983
DualRefine [68]	2023	2(-1, 0)	640 × 192	0.105	0.787	4.544	0.183	0.891	0.964	0.983
Manydepth2 [65]	2025	2(-1, 0)	640 × 192	**0.091**	**0.649**	**4.232**	**0.170**	**0.909**	**0.968**	**0.984**
Single-Frame Methods										
D-HRNet [36]	2022	1	640 × 192	0.102	0.760	4.479	0.179	0.897	0.965	0.983
Lite-Mono [4]	2023	1	640 × 192	0.101	0.729	4.454	0.178	0.897	0.965	0.983
Dynamo-Depth [39]	2023	1	640 × 192	0.112	0.758	4.505	0.183	0.873	0.959	0.984
SC-DepthV3 [3]	2024	1	640 × 192	0.118	0.756	4.709	0.188	0.864	0.960	**0.985**
MonoDiffusion [40]	2024	1	640 × 192	0.099	0.702	4.385	0.176	0.899	0.966	0.984
DCPI-Depth [41]	2025	1	640 × 192	0.097	0.666	4.388	0.173	0.898	0.966	0.985
DiMoDE (Ours)	-	1	640 × 192	**0.097**	**0.652**	**4.337**	**0.172**	**0.899**	**0.967**	**0.985**

TABLE VI

QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE DDAD [56] DATASET. THE BEST RESULTS FOR SINGLE-FRAME AND MULTI-FRAME APPROACHES ARE SHOWN IN BOLD TYPE, RESPECTIVELY. THE SYMBOLS ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER PERFORMANCE, RESPECTIVELY.

Method	Year	Test frames	Resolution (pixels)	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Multi-Frame Methods										
Manydepth [67]	2021	2(-1, 0)	640 × 384	**0.178**	**3.455**	**15.409**	**0.268**	**0.753**	**0.908**	**0.961**
DualRefine [68]	2023	2(-1, 0)	640 × 384	0.186	4.141	15.741	0.271	0.741	0.907	0.960
Manydepth2 [65]	2025	2(-1, 0)	640 × 384	0.183	4.007	16.014	0.283	0.761	0.903	0.950
Single-Frame Methods										
D-HRNet [36]	2022	1	640 × 384	0.198	8.124	16.316	0.289	0.798	0.914	0.956
Lite-Mono [4]	2023	1	640 × 384	0.175	6.425	16.687	0.272	0.799	0.920	0.961
Dynamo-Depth [39]	2023	1	640 × 384	0.150	3.219	14.852	0.246	0.798	0.927	0.969
SC-DepthV3 [3]	2024	1	640 × 384	0.142	3.031	15.868	0.248	0.813	0.922	0.963
MonoDiffusion [40]	2024	1	640 × 384	0.163	6.362	15.922	0.255	0.815	0.928	0.966
DCPI-Depth [41]	2025	1	640 × 384	0.141	2.711	14.757	0.236	0.813	0.931	0.971
DiMoDE (Ours)	-	1	640 × 384	**0.134**	**2.685**	**14.119**	**0.230**	**0.831**	**0.934**	**0.972**

TABLE VII

QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE NUSCENES [50] DATASET. THE BEST RESULTS FOR SINGLE-FRAME, MULTI-FRAME, AND ROBUST DEPTH ESTIMATION APPROACHES ARE SHOWN IN BOLD TYPE, RESPECTIVELY. THE SYMBOLS ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER PERFORMANCE, RESPECTIVELY. RESULTS FOR [43] ARE REPORTED USING OPTIMAL SCALE AND SHIFT TO ALIGN PREDICTIONS WITH THE GROUND TRUTH, RATHER THAN THE MEDIAN ALIGNMENT USED IN OTHER METHODS, WHICH IS NECESSARY BECAUSE THE MODEL PREDICTS SCALE-SHIFT-INVARIANT DEPTH AND REQUIRES BIAS CORRECTION FOR ALIGNMENT [69].

Method	Year	Test frames	Resolution (pixels)	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Robust Depth Estimation Methods										
Syn2Real-Depth [42]	2025	2(-1, 0)	576 × 320	**0.152**	**1.884**	7.442	**0.244**	**0.813**	**0.930**	**0.968**
DepthAnything-AC [43]	2025	1	512 × 288	0.208	3.230	**7.035**	0.281	0.754	0.903	0.954
Multi-Frame Methods										
Manydepth [67]	2021	2(-1, 0)	512 × 288	0.264	3.418	9.316	0.351	0.621	0.836	0.924
DualRefine [68]	2023	2(-1, 0)	512 × 288	**0.255**	**3.204**	**9.129**	**0.334**	**0.635**	**0.844**	**0.929**
Manydepth2 [65]	2025	2(-1, 0)	512 × 288	0.287	4.964	11.132	0.380	0.612	0.812	0.902
Single-Frame Methods										
D-HRNet [36]	2022	1	512 × 288	0.396	15.087	10.414	0.386	0.715	0.838	0.892
Lite-Mono [4]	2023	1	512 × 288	0.419	15.578	9.807	0.449	0.720	0.831	0.879
Dynamo-Depth [39]	2023	1	512 × 288	0.179	2.118	7.050	0.271	0.787	0.896	0.940
SC-DepthV3 [3]	2024	1	512 × 288	0.167	2.003	7.701	0.251	0.784	0.917	0.965
MonoDiffusion [40]	2024	1	512 × 288	0.223	3.564	8.906	0.299	0.723	0.887	0.947
DCPI-Depth [41]	2025	1	512 × 288	0.157	1.795	7.192	0.255	0.790	0.914	0.959
DiMoDE (Ours)	-	1	512 × 288	**0.139**	**1.472**	**6.254**	**0.226**	**0.836**	**0.939**	**0.973**

signals, which are less robust under challenging conditions and often lead to suboptimal DepthNet predictions. In contrast, DiMoDE establishes a constraint cycle via (32) and (33), which incorporates global geometric constraints into

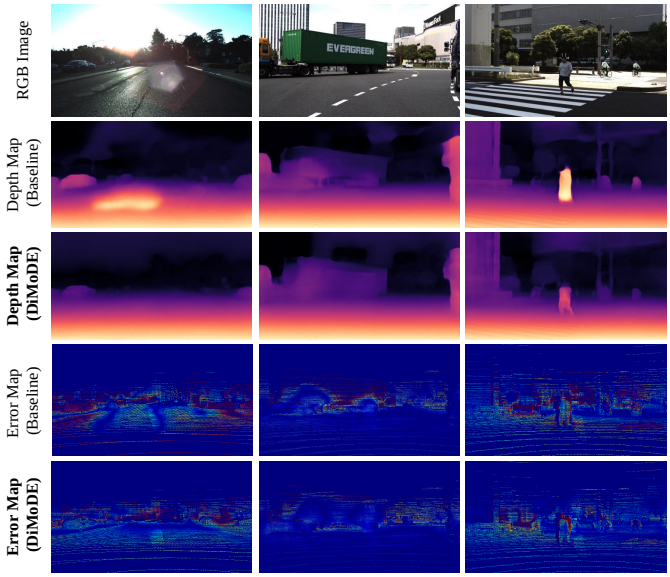

Fig. 11. Qualitative results on the DDAD [56] dataset.

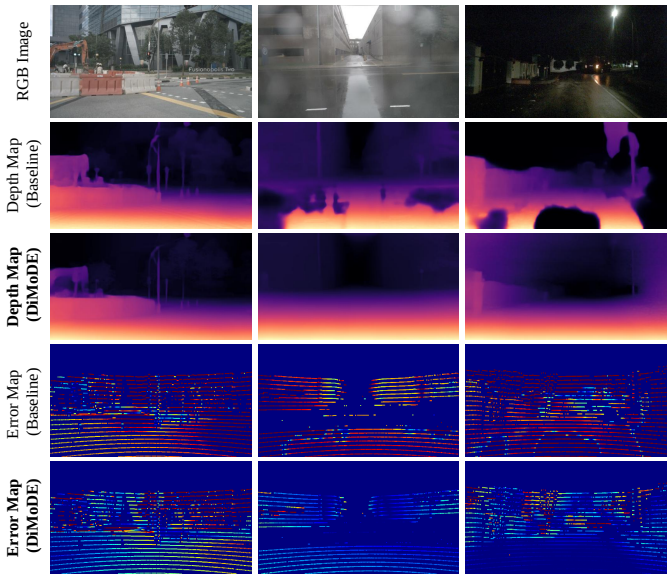

Fig. 12. Qualitative results on the nuScenes [50] dataset.

the training of DepthNet, enabling DiMoDE to effectively overcome the limitations of prior methods. Qualitative results are shown in Figs. 11 and 12, demonstrating that DiMoDE consistently outperforms the baseline model across all aforementioned challenging conditions, which further supports our findings. Furthermore, we compare DiMoDE with two recent SoTA robust depth estimation methods [42] and [43] on the nuScenes dataset. Both approaches rely on specialized domain adaptation strategies, with the method [43] additionally incorporating a depth foundation model [34]. Remarkably, DiMoDE consistently delivers superior performance, underscoring its effectiveness and adaptability as a general-purpose framework for depth and ego-motion joint learning.

Notably, we conduct additional evaluations on the DDAD dataset using models trained on either the KITTI or nuScenes dataset to investigate how variations in environmental conditions within the training data affect the model's generalizability. As shown in Table VIII, methods that perform well on the KITTI dataset fail to generalize effectively to the DDAD dataset. In contrast, models converged on the nuScenes dataset demonstrate significantly better generalizability, achieving performance even comparable to those trained directly on the DDAD dataset. We attribute this phenomenon to the greater environmental complexity and diversity in the nuScenes dataset, which compels models to learn more robust and generalizable representations, rather than overfitting to repetitive, structured scenes in the KITTI dataset. This finding highlights the importance of developing generalizable frameworks that support robust learning across diverse environments, a central objective of DiMoDE.

Finally, we conduct zero-shot evaluations on the Make3D [57] and DIML [58] datasets using models trained on the KITTI dataset to evaluate our model's generalizability to broader, more diverse scenarios. As shown in Table IX, our method achieves slightly better quantitative metrics. Nonetheless, the qualitative comparisons presented in Fig. 13 reveal more pronounced improvements. In particular, the first row of Fig. 13(a), as well as the second and third rows in Fig. 13(b), demonstrate that our method effectively avoids overly distant

TABLE VIII
ZERO-SHOT DEPTH ESTIMATION RESULTS ON THE DDAD [57] DATASET. ALL MODELS ARE TRAINED ON THE KITTI OR NUSCENES DATASET.

Method	Training Data	Abs Rel ↓	Sq Rel ↓	$\delta < 1.25$ ↑
Dynamo-Depth [39]	KITTI	0.191	4.091	0.696
	nuScenes	0.153	3.418	0.785
SC-DepthV3 [3]	KITTI	0.258	5.904	0.554
	nuScenes	0.218	3.868	0.627
MonoDiffusion [40]	KITTI	0.176	3.791	0.738
	nuScenes	0.634	15.197	0.232
Syn2Real-Depth [42]	KITTI	–	–	–
	nuScenes	0.152	3.289	0.782
DiMoDE (Lite-Mono)	KITTI	0.186	4.113	0.710
	nuScenes	**0.142**	**3.159**	**0.797**

TABLE IX
ZERO-SHOT DEPTH ESTIMATION RESULTS ON THE MAKE3D [57] AND DIML [58] DATASETS. ALL MODELS ARE TRAINED ON THE KITTI DATASET.

Dataset	Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓
Make3D	Lite-Mono [4]	0.305	3.060	6.981	0.158
	Dynamo-Depth [39]	0.314	3.259	7.040	0.157
	SC-DepthV3 [3]	0.373	5.584	8.469	0.177
	MonoDiffusion [40]	0.297	2.871	6.877	0.156
	DCPI-Depth [41]	0.291	2.944	6.817	**0.150**
	DiMoDE (Ours)	**0.289**	**2.692**	**6.634**	0.151
DIML	Lite-Mono [4]	0.173	0.271	1.108	0.239
	Dynamo-Depth [39]	0.170	0.262	1.063	0.227
	SC-DepthV3 [3]	0.213	0.407	1.274	0.275
	MonoDiffusion [40]	0.166	0.256	1.084	0.232
	DCPI-Depth [41]	0.163	0.237	1.038	0.226
	DiMoDE (Ours)	**0.160**	**0.237**	**1.033**	**0.224**

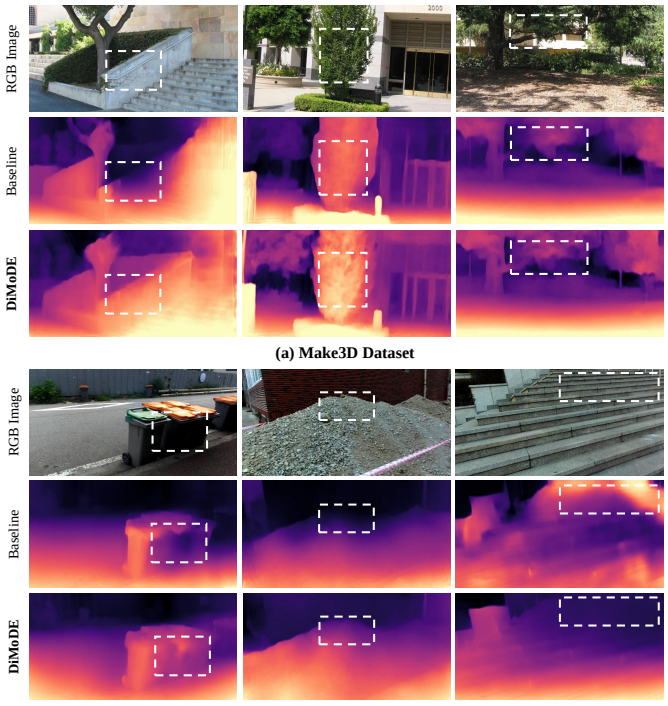

Fig. 13. Qualitative zero-shot monocular depth estimation results on the Make3D [57] and DIML [58] datasets.

predictions near the principal point, a common bias easily learned from the high-frequency scenes in the KITTI dataset. This finding further underscores the benefit of addressing perspective scaling effects, which helps maintain consistent optimization across the entire depth map. Moreover, additional qualitative examples in Fig. 13 show that our DepthNet produces fine-grained predictions in texture-rich regions and demonstrates greater reliability in texture-less areas, compared to the baseline model. These results collectively emphasize that the introduced geometric constraints provide complementary visual cues on top of the photometric supervision, thereby enhancing the robustness and fidelity of depth estimation.

F. Ablation Studies and Hyperparameter Selection

The first ablation study, presented in Table X, validates the effectiveness of the proposed constraints for refining PoseNet outputs. When both $\mathcal{L}_{\text{axi}}$ and $\mathcal{L}_{\text{pla}}$ are removed, a substantial degradation in visual odometry performance is observed, demonstrating the efficacy of imposing optical axis and imaging plane alignments. Furthermore, removing either $\mathcal{L}_{\text{axi}}$ or $\mathcal{L}_{\text{pla}}$ individually results in notable declines in both translation and rotation accuracy, comparable to those observed when both constraints are removed. These findings indicate that the two constraints play complementary roles in constraining rotation estimation, which are consistent with our analyses in Sect. IV. Notably, effective refinement of translational components is only achieved when rotation is well constrained, thereby suppressing irregular rotational flows, as illustrated in Fig. 3. Additionally, the conventional ResNet-based PoseNet is replaced with the architecture adopted in the study [26] for

TABLE X
ABLATION STUDY ON THE EFFECTIVENESS OF THE ADDITIONALLY INCORPORATED GEOMETRIC CONSTRAINTS FOR REFINING POSENET OUTPUTS ON THE KITTI ODOMETRY DATASET ACROSS DIFFERENT POSENET ARCHITECTURES.

Configuration	Seq. 09			Seq. 10		
	e_t	e_r	ATE	e_t	e_r	ATE
Full Implem. (ResNet-18)	**2.86**	**0.74**	**9.83**	**4.20**	**1.22**	**5.81**
Full Implem. (PoseNet from [26])	5.43	2.28	16.29	7.88	2.39	10.20
w/o $\mathcal{L}_{\text{axi}}$ and $\mathcal{L}_{\text{pla}}$ (ResNet-18)	5.84	1.24	26.39	4.89	1.49	6.40
w/o $\mathcal{L}_{\text{axi}}$ (ResNet-18)	5.29	1.12	23.63	4.71	1.31	6.02
w/o $\mathcal{L}_{\text{pla}}$ (ResNet-18)	4.86	1.23	21.96	4.63	1.59	6.24
Baseline (ResNet-18)	7.72	1.71	35.71	8.12	2.96	11.65
Baseline (PoseNet from [26])	7.33	2.61	30.43	10.55	3.89	16.20

TABLE XI
ABLATION STUDY ON THE EFFICACY OF THE GEOMETRIC CONSTRAINTS FROM DECOMPOSED FLOWS ACROSS DIFFERENT DEPTHNET ARCHITECTURES, COMPARED WITH THE CROSS-TASK CONSISTENCY LOSS [38] AND THE CONTEXTUAL-GEOMETRIC DEPTH CONSISTENCY LOSS [41], BOTH OF WHICH RELY ON ORIGINAL OPTICAL FLOWS.

Configuration	KITTI Raw		DDAD		nuScenes	
	Abs Rel	Sq Rel	Abs Rel	Sq Rel	Abs Rel	Sq Rel
Lite-Mono	0.101	0.729	0.162	4.451	0.419	15.578
D-HRNet	0.102	0.760	0.198	8.124	0.396	15.087
Baseline (Lite-Mono)	0.104	0.759	0.150	3.255	0.160	7.250
Baseline (D-HRNet)	0.104	0.810	0.155	3.408	0.160	7.188
w/ [38] (Lite-Mono)	0.102	0.736	0.146	3.288	0.158	2.561
w/ [41] (Lite-Mono)	0.099	0.694	0.143	3.214	0.152	2.130
w/ Ours (Lite-Mono)	**0.097**	**0.652**	**0.134**	2.685	0.139	1.472
w/ Ours (D-HRNet)	0.100	0.729	0.138	**2.588**	**0.138**	**1.404**

both the full implementation and the baseline configuration. The results demonstrate that our framework is compatible with this alternative network and consistently yields significant performance gains.

The second ablation study investigates the effectiveness of the proposed constraints for refining DepthNet outputs, which are derived from decomposed flows following the alignment processes. While prior works have already explored the use of optical flow to constrain depth estimation with all motion types, we not only validate the effectiveness of our method but also compare it with previous approaches. The compared approaches include the cross-task consistency loss introduced in the study [38], which enforces consistency between rigid and optical flows, and the contextual-geometric depth consistency loss proposed in the study [41], where geometric depth is obtained via triangulation. As shown in Table XI, our proposed method significantly outperforms both prior arts across three datasets, supporting our claim that decomposing flows provides more effective geometric constraints. In addition, we replace the default DepthNet with the architecture proposed in [36] to further evaluate the compatibility of DiMoDE. Despite the fundamental differences in architectural design, our method consistently delivers substantial performance gains, demonstrating its robustness and architectural flexibility.

Moreover, we analyze the evolution of the validation metric (Abs Rel) on the nuScenes dataset to analyze the impact of our

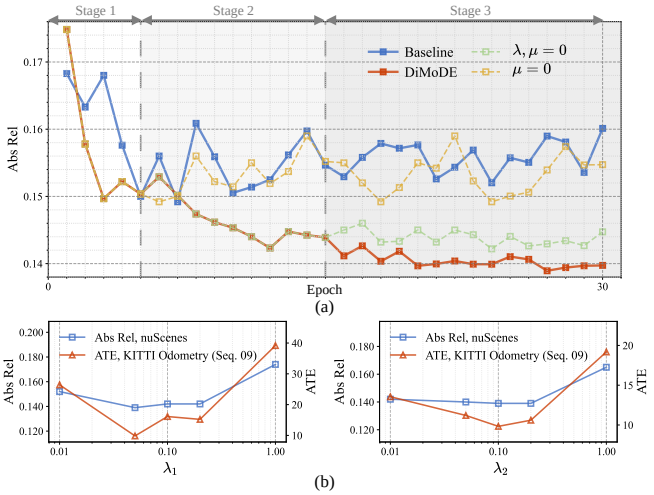

Fig. 14. Quantitative results demonstrating DiMoDE's convergence behavior and robustness to hyperparameter variations: (a) comparison of validation curves for the baseline model, DiMoDE, and its ablation variants; (b) analysis of the impact of varying hyperparameters λ_1 and λ_2 on the performance of both tasks.

method. As shown in Fig. 14(a), the baseline model exhibits noticeable oscillations in its validation curve, indicating that photometric loss alone does not provide sufficiently effective supervision for depth estimation in relatively complex scenes. In contrast, our method achieves significantly smoother convergence, emphasizing the effectiveness of the incorporated geometric constraints. Moreover, we validate the effectiveness of our training strategy (see Sect. VI-A) through two ablation studies: one by removing $\mathcal{L}_{\text{tan}}$ and $\mathcal{L}_{\text{rad}}$ (*i.e.*, setting $\lambda_2 = 0$) during epochs 16-30, and another by additionally removing $\mathcal{L}_{\text{axi}}$ and $\mathcal{L}_{\text{pla}}$ (*i.e.*, setting $\lambda_1, \lambda_2 = 0$) throughout the entire training process. The results confirm that each component contributes to suppressing oscillations and improving overall performance.

Finally, we conduct a hyperparameter selection experiment to determine the optimal weights in (34). As shown in Fig. 14(b), while the default setting ($\lambda_1 = 0.05$, $\lambda_2 = 0.1$) achieves the best performance across both tasks, adjusting the weights within a moderate range (from 0.05 to 0.2) yields comparable results. These results suggest that the performance is relatively robust to these hyperparameters. Nonetheless, performance degrades significantly when the weights are reduced to 0.01 or increased to 1. This degradation is attributed to insufficient gradient strength at lower values, which impedes effective optimization, and to the dominance of the geometric loss at higher values, which disrupts the balance with photometric supervision and leads to unstable training.

VII. Discussion

This section discusses two primary limitations of DiMoDE and analyzes their underlying causes. First, despite notable improvements over existing methods, DiMoDE still struggles to produce reliable predictions under extremely low illumination, as illustrated by the qualitative results on Indoor Seq. 01 in Fig. 10. This limitation primarily arises from the severe degradation of both photometric cues used in the reconstruction loss and the diminished contextual features for reliable dense correspondences generation. Second, although DiMoDE with enhanced constraints demonstrates more robust learning under challenging conditions, it fails to generalize effectively to other unseen real-world sequences. This limitation can be attributed to the significantly greater environmental and motion complexity present in our collected dataset compared to conventional driving scenarios. Consequently, the current network architecture and training strategies fall short in capturing generalizable patterns in such highly complex and diverse environments.

VIII. Conclusion and Future Works

This article presented DiMoDE, an unsupervised joint depth and ego-motion learning framework that discriminately treats motion components. This study first revisited the rigid flows resulting from various motion types, uncovering their distinct characteristics as well as the long-overlooked limitations caused by indiscriminately mixing these flows when generating supervisory signals. To overcome these limitations, DiMoDE introduced explicit geometric constraints from motion components through the alignments of the optical axes and imaging planes between the source and target cameras. These constraints enable targeted optimization for each ego-motion component and enhance depth supervision. Extensive experiments conducted on six public datasets and a newly collected real-world dataset demonstrated that DiMoDE consistently achieves robust learning across diverse real-world environments, delivering impressive performance in both visual odometry and depth estimation tasks.

Future work will focus on addressing existing limitations, including enabling robust training and reliable predictions under extreme illumination, and enhancing generalizability in complex real-world environments through more interpretable motion representations. In addition, we plan to exploit unsupervised learning across diverse scenes for scalable model training, paving the way toward a new paradigm for developing depth foundation models.

References

[1] A. Geiger *et al.*, "Vision meets robotics: the KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] C. Godard *et al.*, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3828–3838.

[3] L. Sun *et al.*, "SC-DepthV3: Robust self-supervised monocular depth estimation for dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 497–508, 2023.

[4] N. Zhang *et al.*, "Lite-Mono: A lightweight CNN and Transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 537–18 546.

[5] D. Eigen *et al.*, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[6] S. Shao *et al.*, "NDDepth: Normal-distance assisted monocular depth estimation and completion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8883–8899, 2024.

[7] L. Yang *et al.*, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 10 371–10 381.

[8] T. Zhou *et al.*, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1851–1858.

[9] N. Yang *et al.*, "Challenges in monocular visual odometry: Photometric calibration, motion bias, and rolling shutter effect," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 2878–2885, 2018.

[10] Y. Zou *et al.*, "Learning monocular visual odometry via self-supervised long-term modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 12359. Springer, 2020, pp. 710–727.

[11] J.-W. Bian *et al.*, "Auto-rectify network for unsupervised indoor depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9802–9813, 2021.

[12] H. Zhan *et al.*, "Visual odometry revisited: What should be learnt?" in *International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4203–4210.

[13] A. Geiger *et al.*, "StereoScan: Dense 3d reconstruction in real-time," in *IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 963–968.

[14] R. Mur-Artal *et al.*, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[15] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[16] C. Campos *et al.*, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[17] J. Engel *et al.*, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[18] L. Sun *et al.*, "Improving monocular visual odometry using learned depth," *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 3173–3186, 2022.

[19] S. Wang *et al.*, "DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2043–2050.

[20] T. Shen *et al.*, "Beyond photometric loss for self-supervised ego-motion estimation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6359–6365.

[21] S. Wang *et al.*, "End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 513–542, 2018.

[22] R. Li *et al.*, "UnDeepVO: Monocular visual odometry through unsupervised deep learning," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7286–7291.

[23] J.-W. Bian *et al.*, "Unsupervised scale-consistent depth learning from video," *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2548–2564, 2021.

[24] S. Zhang *et al.*, "Information-theoretic odometry learning," *International Journal of Computer Vision*, vol. 130, no. 11, pp. 2553–2570, 2022.

[25] Z. Jiang *et al.*, "Self-supervised ego-motion estimation based on multi-layer fusion of rgb and inferred depth," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[26] Y. Feng *et al.*, "SCIPaD: Incorporating spatial clues into unsupervised pose-depth joint learning," *IEEE Transactions on Intelligent Vehicles*, 2024, DOI: 10.1109/TIV.2024.3460868.

[27] N. Yang *et al.*, "D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1281–1292.

[28] L. Tiwari *et al.*, "Pseudo RGB-D for self-improving monocular SLAM and depth prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 437–455.

[29] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[30] Y. Cao *et al.*, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2017.

[31] H. Fu *et al.*, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2002–2011.

[32] S. F. Bhat *et al.*, "AdaBins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4009–4018.

[33] S. Shao *et al.*, "IEBins: Iterative elastic bins for monocular depth estimation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 53 025–53 037, 2023.

[34] L. Yang *et al.*, "Depth anything v2," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 21 875–21 911, 2025.

[35] M. Hu *et al.*, "Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[36] M. He *et al.*, "RA-Depth: Resolution adaptive self-supervised monocular depth estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 565–581.

[37] Z. Zhou *et al.*, "Recurrent multiscale feature modulation for geometry consistent depth learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9551–9566, 2024.

[38] Y. Zou *et al.*, "DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 36–53.

[39] Y. Sun and B. Hariharan, "Dynamo-Depth: Fixing unsupervised depth estimation for dynamical scenes," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[40] S. Shao *et al.*, "MonoDiffusion: Self-Supervised Monocular Depth Estimation Using Diffusion Model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, pp. 3664–3678, 2024.

[41] M. Zhang *et al.*, "DCPI-Depth: Explicitly infusing dense correspondence prior to unsupervised monocular depth estimation," *IEEE Transactions on Image Processing*, vol. 34, pp. 4258–4272, 2025.

[42] W. Yan *et al.*, "Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 21 880–21 890.

[43] B. Sun *et al.*, "Depth anything at any condition," *CoRR*, 2025.

[44] Z. Wang *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[45] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[46] D. Nister *et al.*, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[47] R. S. Bowen *et al.*, "Dimensions of motion: Monocular prediction through flow subspaces," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 454–464.

[48] M. Poggi and F. Tosi, "FlowSeek: Optical flow made easier with depth foundation models and motion bases," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 5667–5679.

[49] A. Geiger *et al.*, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[50] H. Caesar *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631.

[51] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 402–419.

[52] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *CoRR*, 2017.

[53] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009, pp. 248–255.

[54] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2758–2766.

[55] W. Xu *et al.*, "Fast-LIO2: Fast direct lidar-inertial odometry," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.

[56] V. Guizilini *et al.*, "3D packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2485–2494.

[57] A. Saxena *et al.*, "Make3D: Learning 3D scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2008.

[58] J. Cho *et al.*, "DIML/CVL RGB-D dataset: 2m RGB-D images of natural indoor and outdoor scenes," *arXiv preprint arXiv:2110.11590*, 2021.

[59] N. Yang *et al.*, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 817–833.
[60] W. Zhao *et al.*, "Towards better generalization: Joint depth-pose learning without posenet," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9151–9161.
[61] C. Wang *et al.*, "Self-supervised learning of monocular visual odometry and depth with uncertainty-aware scale consistency," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3984–3990.
[62] H. Zhan *et al.*, "Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 340–349.
[63] C. Wang *et al.*, "Motionhint: Self-supervised monocular visual odometry with motion constraints," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1265–1272.
[64] W. Wei *et al.*, "Lite-SVO: Towards a lightweight self-supervised semantic visual odometry exploiting multi-feature sharing architecture," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 10 305–10 311.
[65] K. Zhou *et al.*, "Manydepth2: Motion-aware self-supervised monocular depth estimation in dynamic scenes," *IEEE Robotics and Automation Letters*, 2025.
[66] J. Sturm *et al.*, "A benchmark for the evaluation of RGB-D SLAM systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2012, pp. 573–580.
[67] J. Watson *et al.*, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 1164–1174.
[68] A. Bangunharcana *et al.*, "DualRefine: Self-supervised depth and pose estimation through iterative epipolar sampling and refinement toward equilibrium," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 726–738.
[69] J. Wang *et al.*, "Jasmine: Harnessing diffusion prior for self-supervised depth estimation," *CoRR*, 2025.