DCPI-Depth: Explicitly Infusing Dense Correspondence Prior to Unsupervised Monocular Depth Estimation

Mengtan Zhang, Yi Feng, *Student Member, IEEE*, Qijun Chen, *Senior Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—There has been a recent surge of interest in learning to perceive depth from monocular videos in an unsupervised fashion. A key challenge in this field is achieving robust and accurate depth estimation in regions with weak textures or where dynamic objects are present. This study makes three major contributions by delving deeply into dense correspondence priors to provide existing frameworks with explicit geometric constraints. The first novel contribution is a contextual-geometric depth consistency loss, which employs depth maps triangulated from dense correspondences based on estimated ego-motion to guide the learning of depth perception from contextual information, since explicitly triangulated depth maps capture accurate relative distances among pixels. The second novel contribution arises from the observation that there exists an explicit, deducible relationship between optical flow divergence and depth gradient. A differential property correlation loss is therefore designed to refine depth estimation with a specific emphasis on local variations. The third novel contribution is a bidirectional stream co-adjustment strategy that enhances the interaction between rigid and optical flows, encouraging the former towards more accurate correspondence and making the latter more adaptable across various scenarios under the static scene hypotheses. DCPI-Depth, a framework that incorporates all these innovative components and couples two bidirectional and collaborative streams, achieves state-of-the-art performance and generalizability across multiple public datasets, outperforming all existing prior arts. Specifically, it demonstrates accurate depth estimation in texture-less and dynamic regions, and shows more reasonable smoothness. Our source code is publicly available at https://mias.group/DCPI-Depth.**

***Index Terms*—Depth estimation, dynamic object, dense correspondence, geometric constraint.**

Received 16 January 2025; revised 18 April 2025 and 5 June 2025; accepted 11 June 2025. Date of publication 25 June 2025; date of current version 8 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62473288 and Grant 62233013; in part by the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University under Grant HMHAI-202406; in part by the Fundamental Research Funds for the Central Universities; in part by the NIO University Programme (NIO UP); and in part by the Xiaomi Young Talents Program. The associate editor coordinating the review of this article and approving it for publication was Prof. Sebastian Knorr. *(Corresponding author: Rui Fan.)*

Mengtan Zhang is with Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 201210, China (e-mail: 2050026@tongji.edu.cn).

Yi Feng and Qijun Chen are with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: fengyi@ieee.org; qjchen@tongji.edu.cn).

Rui Fan is with the College of Electronics and Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China, and also 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).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2025.3581422, provided by the authors.

Digital Object Identifier 10.1109/TIP.2025.3581422

I. Introduction

MONOCULAR depth estimation, a crucial research field in computer vision and robotics, has applications across various domains, such as autonomous driving [1], augmented reality [2], and embodied artificial intelligence [3]. It provides agents with powerful environmental perception capabilities, enabling robust ego-motion estimation and 3D geometry reconstruction [4]. Early monocular depth estimation approaches [5], [6], developed based on supervised learning, typically require a large amount of well-annotated, per-pixel depth ground truth (generally acquired using high-precision LiDARs [7]) for model training [8]. Nevertheless, collecting and labeling such data is tedious and costly [9]. Thus, the practical use of these supervised approaches remains limited [10].

In recent years, un/self-supervised monocular depth estimation has garnered significant attention [4], [9], [11], [12], [13], [14]. Such approaches obviate the need for extensive depth ground truth by leveraging either stereo image pairs [15], [16] or monocular videos to jointly learn depth and ego-motion estimation [4]. Specifically, given the target and source images, the estimated depth map (at the target view) and camera ego-motion are employed to warp the source image into the target view. The depth estimation network (hereafter referred to as *DepthNet*) and the pose estimation network (hereafter referred to as *PoseNet*) are then jointly trained in an un/self-supervised manner by minimizing the photometric loss, which measures the consistency between the reconstructed and original target images [4], [9], [13], [17].

Despite the progress made, three limitations in existing frameworks continue to impede further advances in monocular depth estimation:

1) DepthNet perceives depth based on the contextual information in RGB images. While it effectively determines whether an object is in front of or behind another, accurately and efficiently learning their relative distance by minimizing the photometric loss is challenging. This is because photometric loss cannot directly reflect the magnitude of depth error, sometimes resulting in

unsuitable gradients for optimizing depth during back-propagation.

2) Local depth variation is commonly constrained by edge-aware smoothness loss, which encourages local smoothness in depth based on image gradients [15]. However, since changes in image intensity do not directly correlate with local depth variation, this enforced smoothing can introduce errors in depth estimation.
3) Comparable photometric losses can result from different rigid flows, which may map a pixel to the wrong candidates with similar pixel intensities. This implies that the supervisory signals provided by the photometric loss to DepthNet and PoseNet are indirect, possibly leading to unsatisfactory robustness of monocular depth estimation, particularly in regions with weak textures.

Therefore, in this article, we introduce a novel unsupervised monocular depth estimation framework, known as **D**ense **C**orrespondence **P**rior-**I**nfused **Depth (DCPI-Depth)**, to overcome the limitations above by exploiting the depth cues in dense correspondence priors. Our DCPI-Depth consists of two bidirectional and collaborative streams: a traditional photometric consistency-guided (PCG) stream and our proposed correspondence prior-guided (CPG) stream. The PCG stream, following the prevalently used methods [4], [9], [13], [14], employs estimated depth and ego-motion information to warp the source frame into the target view, and then computes a photometric loss to provide the supervisory signal. On the other hand, the CPG stream leverages a pre-trained FlowNet [18] to provide dense correspondence priors. These priors are first utilized along with the estimated ego-motion to construct a geometric-based depth map via triangulation [19]. Such an explicitly derived depth map captures accurate relative distances among pixels. By enforcing consistency between these two sources of depth maps through a newly developed contextual-geometric depth consistency (CGDC) loss, we significantly optimize the convergence of DepthNet during training. Moreover, optical flow divergence, a differential property of dense correspondence priors, is found to have an explicit relationship with depth gradient. Building upon this relationship, we develop a novel differential property correlation (DPC) loss to improve depth quality from the aspect of local variation. Finally, a bidirectional stream co-adjustment (BSCA) strategy is adopted to make the two streams complement each other, where the rigid flow in the PCG stream mainly alleviates the misguidance of the CPG steam to depth on dynamic objects, while the optical flow in the CPG stream refines the rigid flow produced in the PCG stream with dense correspondences.

In summary, the main contributions of this article include:

- DCPI-Depth, a novel unsupervised monocular depth estimation framework with a CPG stream developed to infuse the dense correspondence prior into the traditional PCG stream;
- A CGDC loss to optimize the convergence of DepthNet using a geometric-based depth map constructed by triangulating dense correspondence priors with estimated ego-motion;
- A DPC loss to further refine the quality of the estimated depth from the aspect of local variation based on the explicit relationship between optical flow divergence and depth gradient;
- A BSCA strategy to enable the two streams to complement each other, with a specific emphasis on improving depth accuracy in dynamic regions without using masking techniques.

The remainder of this article is organized as follows: Sect. II presents an overview of existing monocular depth estimation methods. In Sect. III, we detail the proposed DCPI-Depth framework. In Sect. IV, we present the experimental results across several public datasets. In Sect. V, we discuss two limitations of the DCPI-Depth framework. Finally, we conclude this article and discuss possible future work in Sect. VI.

II. Related Work

A. Supervised Monocular Depth Estimation

Supervised monocular depth estimation approaches [6], [20], [21], [22] require depth ground truth for model training. As the first attempt, the study [5] proposed a coarse-to-fine architecture and a scale-invariant loss function to perceive depth from a single image. In subsequent studies [23] and [24], monocular depth estimation was reformulated as a per-pixel classification task, where depth ranges instead of exact depth values are predicted. Furthermore, to combine the benefits of both regression and classification tasks, the study [25] redefined this problem as a per-pixel classification-regression task, which utilizes bins to categorize depth values. Building on this approach, the study [26] further enhanced depth estimation by progressively optimizing the search range for high-quality depth within these bins. With the rapid advancement of generative models, the study [27] introduced a diffusion model-based depth estimation approach, which produces highly refined depth predictions through iterative denoising. More recently, Depth Anything [28] has demonstrated impressive performance, primarily due to its powerful backbone (a vision Transformer-based vision foundation model) that is capable of extracting general-purpose, informative deep features. It first reproduces a MiDaS-based [29] teacher model with pre-trained weights from DINOv2 [30], and then utilizes the teacher's predictions as pseudo-labels to train a student model on large-scale unlabeled data. Building on single-image depth estimation, monocular video depth estimation incorporates both temporal and geometric consistencies [2]. Representative works, such as [2] and [31], pioneered the use of correspondences and camera poses to enforce inter-frame depth consistency in 3D space, which inspires us to further explore the way of infusing dense correspondence priors to guide the unsupervised learning for monocular depth estimation.

B. Un/Self-Supervised Monocular Depth Estimation

To liberate monocular depth estimation from dependence on extensive ground-truth data, un/self-supervised approaches [4], [9], [13], [14], [15], [32] have emerged as viable alternatives. These methods typically utilize the estimated depth map to establish a differentiable warping between two images and

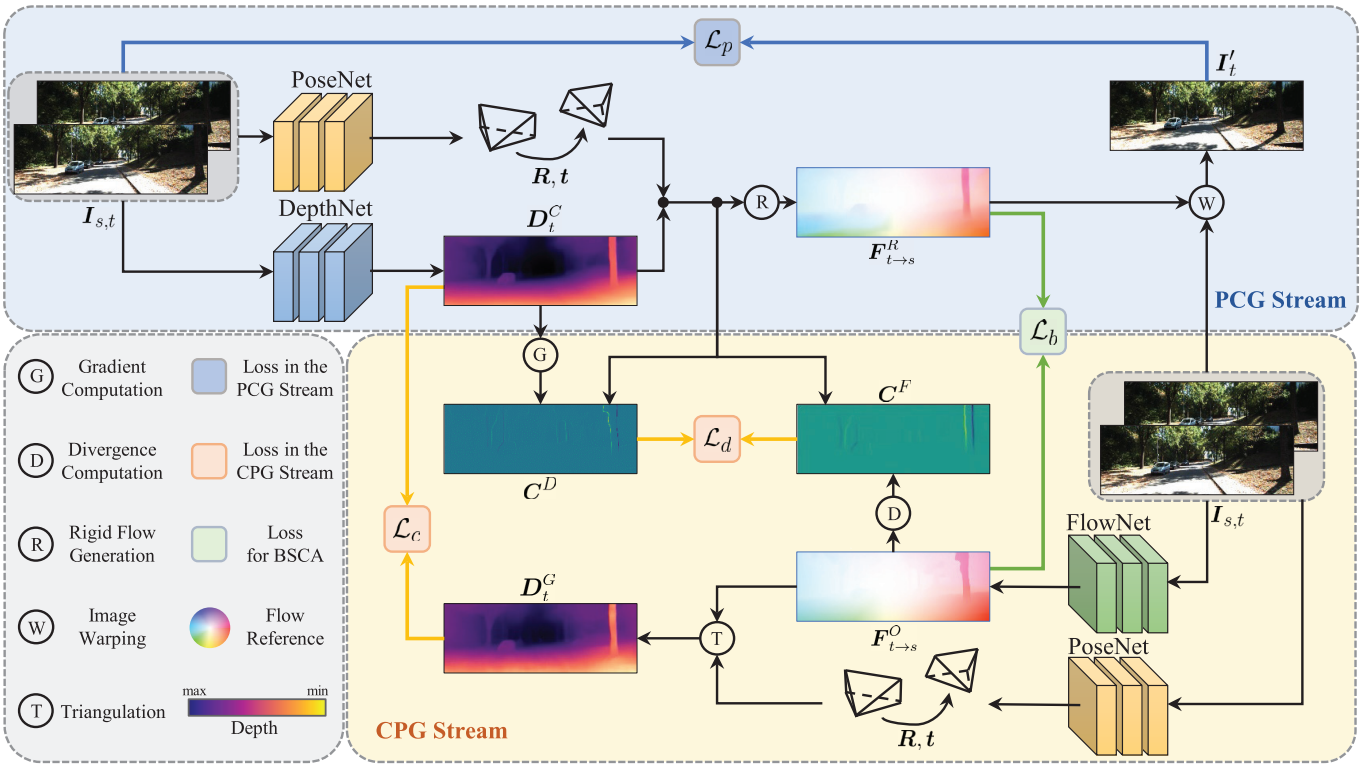

Fig. 1. The overall architecture of our proposed DCPI-Depth framework, which consists of two collaborative and bidirectional streams: PCG and CPG. The input image pairs, PoseNet, and the estimated ego-motion are depicted separately in each stream.

employ photometric loss to provide supervisory signals [4], [33]. The study [33] represents the first reported attempt to learn monocular depth estimation from stereo image pairs within a self-supervised framework. Subsequently, the study [4] extended this approach by coupling the learning of depth and ego-motion estimation from monocular videos. However, un/self-supervised methods often face challenges with independently moving objects and preserving clear object boundaries, due to multi-view ambiguities [13]. Therefore, in the study [9], a minimum reprojection loss and an auto-masking technique were introduced to exclude such regions during model training, which significantly improves depth estimation performance. Building upon these prior arts, several studies explored more sophisticated network architectures [14], [34], [35], [36], [37], [38] for improved depth estimation performance. Others incorporate additional relevant tasks such as optical flow estimation [39], [40], [41] and semantic segmentation [42], [43], [44] to enhance cross-task consistency or address the dynamic object challenge.

While these efforts have demonstrated promising performance, existing unsupervised frameworks still present significant opportunities for refinement [13]. This limitation primarily arises from the reliance on contextual information to infer the pixel-wise depth map, which is indirectly supervised through the minimization of photometric loss. In the absence of guidance from additional and meaningful prior knowledge, DepthNet struggles to perform robustly in challenging scenarios. Therefore, in study [13], a monocular depth estimation model pre-trained on large-scale datasets is utilized to provide pseudo-depth, a single-image depth prior, and two depth refinement loss functions are also designed to achieve more robust and reliable depth estimation. However, the limited accuracy of the pseudo-depth significantly restricts the refinement capabilities of these loss functions. Therefore, in this article, we resort to dense correspondence priors for depth refinement. Unlike pseudo-depth, these priors offer more direct, reliable, and interpretable geometric guidance through our developed CGDC and DPC losses.

III. Methodology

A. Overall Architecture

As illustrated in Fig. 1, our proposed DCPI-Depth framework comprises two collaborative and bidirectional streams: PCG and CPG. The former, following the prior studies [9], [13], [14], effectively yet indirectly supervises the training of DepthNet and PoseNet through the photometric loss, while the latter infuses dense correspondence priors (provided by a pre-trained FlowNet) into the former to overcome the limitations of current SoTA frameworks [13], [14], [37], which rely solely on the PCG stream. Specifically, within the CPG stream, we introduce two novel loss functions: (1) a CGDC loss that guides the training of DepthNet by enforcing consistency between geometric-based and contextual-based depth maps, enabling DepthNet to capture accurate relative distances among pixels, thereby optimizing its convergence during training; (2) a DPC loss to constrain the local variation of depth based on the explicit relationship between optical flow divergence and depth gradient. Furthermore, we develop a BSCA strategy to collectively improve the aforementioned two streams: the rigid

flow ensures accurate geometric guidance for depth estimation mainly on dynamic objects, while the optical flow refines the rigid flow based on dense correspondence, thus allowing these two streams to effectively complement each other.

B. Contextual-Geometric Depth Consistency Loss

In the conventional PCG stream, given target and source video frames[1] $\boldsymbol{I}_{t,s} \in \mathbb{R}^{H\times W\times 3}$, DepthNet takes $\boldsymbol{I}_t$ as input to infer a depth map $\boldsymbol{D}_t^C \in \mathbb{R}^{H\times W}$ based on contextual information, where H and W represent the height and width of the input image, respectively, while PoseNet estimates the ego-motion, including a rotation matrix $\boldsymbol{R} = [\boldsymbol{r}_1^\top, \boldsymbol{r}_2^\top, \boldsymbol{r}_3^\top]^\top \in SO(3)$ and a translation vector $\boldsymbol{t} = [t_1, t_2, t_3]^\top \in \mathbb{R}^3$ from the source view to the target view. The rigid flow map $\boldsymbol{F}_{t\to s}^R \in \mathbb{R}^{H\times W\times 2}$ from the target view to the source view can then be generated as follows:

$$\begin{bmatrix} \boldsymbol{F}_{t\to s}^R(\boldsymbol{p}_t) \\ 0 \end{bmatrix} + \tilde{\boldsymbol{p}}_t \sim \boldsymbol{K}\begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\end{bmatrix}\begin{bmatrix} \boldsymbol{D}_t^C(\boldsymbol{p}_t)\boldsymbol{K}^{-1}\tilde{\boldsymbol{p}}_t \\ 1 \end{bmatrix}, \tag{1}$$

where the symbol $\sim$ indicates that two vectors are equal up to a scale factor, $\boldsymbol{K}$ represents the camera intrinsic matrix, $\boldsymbol{p}_t = [u, v]^\top$ denotes a 2D pixel, and $\tilde{\boldsymbol{p}}_t$ is its homogeneous coordinates. $\boldsymbol{I}_s$ is then warped into the target view using the rigid flow map $\boldsymbol{F}_{t\to s}^R$, generating $\boldsymbol{I}_t'$. By comparing $\boldsymbol{I}_t'$ with $\boldsymbol{I}_t$, the following photometric loss is computed to provide a supervisory signal for the training of both DepthNet and PoseNet [45]:

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

where SSIM denotes the pixel-wise structural similarity index operation [46], and the weight α is set to 0.85, following the study [9].

DepthNet in the conventional PCG stream is trained to infer depth value per pixel from contextual information by minimizing (2) based on given RGB image pairs. Previous studies have neglected to incorporate geometric guidance into DepthNet training, leading to a significant limitation. While DepthNet can ascertain whether an object is in front of or behind another relative to the camera origin from contextual information, it struggles to effectively learn the extent of their relative distances by solely minimizing the photometric loss. This challenge arises because errors in image intensities do not directly reflect depth errors in terms of magnitude, rendering the gradient from photometric loss during back-propagation not always suitable for depth optimization.

To address these limitations, we resort to dense correspondence priors to generate another depth map based on well-developed and interpretable principles of multi-view geometry, thereby providing an additional constraint on depth estimation from RGB images. We first introduce a pre-trained FlowNet [18] to generate the optical flow map $\boldsymbol{F}_{t\to s}^O \in \mathbb{R}^{H\times W\times 2}$, from which the following dense correspondence priors are derived:

$$\begin{cases} \hat{\boldsymbol{p}}_t^C = \boldsymbol{K}^{-1}\tilde{\boldsymbol{p}}_t \\ \hat{\boldsymbol{p}}_s^C = \boldsymbol{K}^{-1}\left(\tilde{\boldsymbol{p}}_t + \begin{bmatrix}\boldsymbol{F}_{t\to s}^O(\boldsymbol{p}_t) \\ 0\end{bmatrix}\right), \end{cases} \tag{3}$$

[1]In this article, the subscripts "t" and "s" denote "target" and "source", respectively.

where $\hat{\boldsymbol{p}}_t^C$ and $\hat{\boldsymbol{p}}_s^C$ represent a pair of normalized camera coordinates along the optical axis in the target and source views, respectively. We then leverage such dense correspondence priors along with the ego-motion estimated by PoseNet to construct a geometric-based depth map $\boldsymbol{D}_t^G \in \mathbb{R}^{H\times W}$ via triangulation [19] based on the following relationship:

$$\hat{\boldsymbol{p}}_s^C \sim \begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\end{bmatrix}\begin{bmatrix} \boldsymbol{D}_t^G(\boldsymbol{p}_t)\hat{\boldsymbol{p}}_t^C \\ 1 \end{bmatrix}. \tag{4}$$

$\hat{p}_{s,i}^C$, the i-th element in $\hat{\boldsymbol{p}}_s^C$ $(i = \{1, 2\})$, is expressed as follows:

$$\hat{p}_{s,i}^C = \frac{\boldsymbol{D}_t^G(\boldsymbol{p}_t)\boldsymbol{r}_i^\top\hat{\boldsymbol{p}}_t^C + t_i}{\boldsymbol{D}_t^G(\boldsymbol{p}_t)\boldsymbol{r}_3^\top\hat{\boldsymbol{p}}_t^C + t_3}. \tag{5}$$

$\boldsymbol{D}_t^G$ can then be yielded by triangulating the dense correspondences using the following expression:

$$\boldsymbol{D}_t^G(\boldsymbol{p}_t) = \frac{\sum_{i=1}^{2}\left(t_i - \hat{p}_{s,i}^C t_3\right)}{\sum_{i=1}^{2}\left(\hat{p}_{s,i}^C\boldsymbol{r}_3^\top\hat{\boldsymbol{p}}_t^C - \boldsymbol{r}_i^\top\hat{\boldsymbol{p}}_t^C\right)}, \tag{6}$$

which reflects the relative distances among pixels, derived from motion. Therefore, we employ the following CGDC loss:

$$\mathcal{L}_c = \frac{1}{HW}\sum_{\boldsymbol{p}}\frac{\left|\boldsymbol{D}_t^G(\boldsymbol{p}) - \boldsymbol{D}_t^C(\boldsymbol{p})\right|}{\boldsymbol{D}_t^C(\boldsymbol{p})} \tag{7}$$

to provide DepthNet with an additional constraint, enabling it to capture accurate relative distances among pixels, reflected by motion. Since the CGDC loss directly measures the relative differences between two types of depth maps, it provides more informative gradients with respect to depth errors, thereby facilitating the convergence of DepthNet. Considering that higher depth values potentially exhibit greater absolute depth error, we adopt relative error in our CGDC loss to ensure more consistent gradients for back-propagation across all pixels. The effectiveness of our proposed CGDC loss is validated and discussed in Sect. IV-E.

C. Differential Property Correlation Loss

Existing approaches [14], [37] that include only the PCG stream often struggle to distinguish and effectively handle regions with different levels of continuity. Specifically, these methods encounter difficulties in ensuring smooth depth changes in continuous regions and preserving clear boundaries near or at discontinuities. This problem arises primarily due to the lack of proper constraints that encourage DepthNet to consider local depth variations, especially since the depth of each pixel is estimated independently. Several studies [8], [9], [14], [37], [45], [47] introduced an edge-aware smoothness loss based on image gradients, initially presented in the study [15], to encourage local smoothness in depth estimation. However, such a loss function is somewhat problematic and incomplete. While this loss function is effective in regions where depth and image intensity have consistent change trends, it cannot constrain the extent of smoothing. Moreover, in continuous regions with rich texture or at discontinuities with subtle

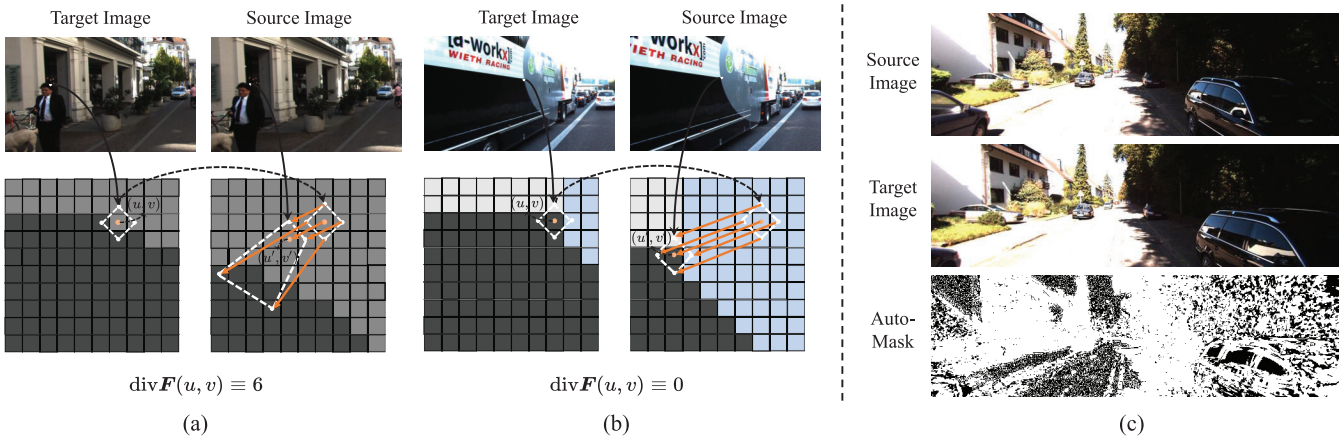

Fig. 2. Illustrations of optical flow divergence and auto-masking result: (a) optical flow divergence for pixels with similar intensities yet being spatially discontinuous; (b) optical flow divergence for pixels with significantly different intensities yet being spatially continuous; (c) auto-masking result for the given source and target images. A given pixel and its four neighbors in the target image are utilized for visualization in (a) and (b), where it can be observed that their correspondences in the source image are widely separated in (a) but similarly distributed in (b). The auto-masking algorithm tends to overly mask static regions, particularly in low-texture areas or when overexposed, and cannot effectively mask dynamic objects.

texture changes, this loss function may prove ineffective or even cause misleading guidance.

This study finds that, compared to pixel intensity, the dense correspondence priors between two images, as provided by FlowNet, have a more direct and deducible relationship with depth changes. As illustrated in Fig. 2(a), adjacent pixels that have similar intensities but are spatially discontinuous (because they are located on different objects) are likely to exhibit significantly different apparent motions, resulting in high optical flow divergence due to the separation of these pixels. In contrast, as illustrated in Fig. 2(b), pixels that have different intensities but are located in continuous regions typically have similar apparent motions, resulting in low optical flow divergence. $\mathrm{div}\boldsymbol{F}$, the divergence map of the given optical flow map $\boldsymbol{F}$, can be numerically calculated through the following expression:

$$\begin{aligned}\mathrm{div}\boldsymbol{F}(u,v) \equiv &-\boldsymbol{n}_u^\top \boldsymbol{F}(u-1,v) + \boldsymbol{n}_u^\top \boldsymbol{F}(u+1,v)\\ &+ \boldsymbol{n}_v^\top \boldsymbol{F}(u,v+1) - \boldsymbol{n}_v^\top \boldsymbol{F}(u,v-1),\end{aligned} \tag{8}$$

where $\boldsymbol{n}_u = [1,0]^\top$ and $\boldsymbol{n}_v = [0,1]^\top$ are two unit vectors in the horizontal and vertical directions, and the symbol $\equiv$ represents discretization. Therefore, we are motivated to establish an explicit constraint on local depth variation by leveraging the two correlated differential properties: optical flow divergence and depth gradient. This allows for more accurate depth estimation by ensuring that changes in depth are consistently aligned with variations in optical flow.

However, it has been proven in the study [48] that rotational flow is independent of depth. Therefore, only the translational component of ego-motion contributes to constraining depth estimation from the aspect of local variation using optical flow divergence, and incorporating rotational flow, particularly when it is substantial, can disrupt this constraint.

To address this issue, we eliminate the rotational apparent motions in the optical flow using the following expression:

$$\boldsymbol{F}^{\mathrm{Tra}}_{t\to s} = \boldsymbol{F}^{O}_{t\to s} - \boldsymbol{F}^{\mathrm{Rot}}_{t\to s}, \tag{9}$$

where $\boldsymbol{F}^{\mathrm{Tra}}_{t\to s}$ denotes the translational optical flow, and $\boldsymbol{F}^{\mathrm{Rot}}_{t\to s}$, the rotational rigid flow, is generated using the estimated $\boldsymbol{R}$ and a translation vector of zeros $\mathbf{0}$ as follows:

$$\begin{bmatrix}\boldsymbol{F}^{\mathrm{Rot}}_{t\to s}(\boldsymbol{p}_t)\\ 0\end{bmatrix} + \tilde{\boldsymbol{p}}_t \sim \boldsymbol{K}\begin{bmatrix}\boldsymbol{R} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{D}^C_t(\boldsymbol{p}_t)\boldsymbol{K}^{-1}\tilde{\boldsymbol{p}}_t\\ 1\end{bmatrix}. \tag{10}$$

The relation between translational apparent motion and depth can then be written as follows:

$$\begin{aligned}\begin{bmatrix}\boldsymbol{F}^{\mathrm{Tra}}_{t\to s}(\boldsymbol{p}_t)\\ 0\end{bmatrix} &= \frac{\boldsymbol{K}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{t}\end{bmatrix}\begin{bmatrix}\boldsymbol{D}^C_t(\boldsymbol{p}_t)\boldsymbol{K}^{-1}\tilde{\boldsymbol{p}}_t\\ 1\end{bmatrix}}{\boldsymbol{D}^C_s(\boldsymbol{p}_s)} - \tilde{\boldsymbol{p}}_t\\ &= \left(\frac{\boldsymbol{D}^C_t(\boldsymbol{p}_t)}{\boldsymbol{D}^C_s(\boldsymbol{p}_s)} - 1\right)\tilde{\boldsymbol{p}}_t - \frac{\boldsymbol{K}\boldsymbol{t}}{\boldsymbol{D}^C_s(\boldsymbol{p}_s)}\\ &= \frac{t_3}{\boldsymbol{D}^C_s(\boldsymbol{p}_s)}\left(\tilde{\boldsymbol{p}}_t - \tilde{\boldsymbol{p}}_o - \boldsymbol{K}\begin{bmatrix}t_1/t_3\\ t_2/t_3\\ 0\end{bmatrix}\right),\end{aligned} \tag{11}$$

where $\boldsymbol{D}^C_s(\boldsymbol{p}_s) = \boldsymbol{D}^C_t(\boldsymbol{p}_t) - t_3$ under the condition of $\boldsymbol{R} = \boldsymbol{I}$, and $\tilde{\boldsymbol{p}}_o$ denotes the homogeneous coordinates of the image principal point. We calculate the divergence of the optical flow at $\boldsymbol{p}_t$ as follows:

$$\begin{aligned}\nabla\cdot\boldsymbol{F}^{\mathrm{Tra}}_{t\to s}(\boldsymbol{p}_t) &= \nabla\cdot\left(\frac{t_3}{\boldsymbol{D}^C_t(\boldsymbol{p}_t) - t_3}\boldsymbol{q}_t\right)\\ &= \boldsymbol{q}_t\cdot\nabla\frac{t_3}{\boldsymbol{D}^C_t(\boldsymbol{p}_t) - t_3} + \frac{t_3}{\boldsymbol{D}^C_t(\boldsymbol{p}_t) - t_3}\nabla\cdot\boldsymbol{p}_t,\end{aligned} \tag{12}$$

where $\nabla = \left[\frac{\partial}{\partial u}, \frac{\partial}{\partial v}\right]^\top$, and $\boldsymbol{q}_t$ is expressed as follows:

$$\begin{bmatrix}\boldsymbol{q}_t\\ 0\end{bmatrix} = \tilde{\boldsymbol{p}}_t - \tilde{\boldsymbol{p}}_o - \boldsymbol{K}\begin{bmatrix}t_1/t_3\\ t_2/t_3\\ 0\end{bmatrix}. \tag{13}$$

Rewriting (12) into the following expression:

$$\underbrace{\frac{\boldsymbol{D}^C_t(\boldsymbol{p}_t) - t_3}{t_3}\nabla\cdot\boldsymbol{F}^{\mathrm{Tra}}_{t\to s}(\boldsymbol{p}_t) - \nabla\cdot\boldsymbol{p}_t}_{\boldsymbol{C}^F(\boldsymbol{p}_t)}$$

$$= \underbrace{-\frac{\boldsymbol{q}_t}{\boldsymbol{D}_t^C(\boldsymbol{p}_t) - t_3}}_{\boldsymbol{C}^D(\boldsymbol{p}_t)} \cdot \nabla \boldsymbol{D}_t^C(\boldsymbol{p}_t). \tag{14}$$

where $\boldsymbol{C}^F \in \mathbb{R}^{H\times W}$ and $\boldsymbol{C}^D \in \mathbb{R}^{H\times W}$ denote the differential properties derived from optical flow divergence and depth gradient, respectively. (14) establishes a mathematical relationship between these two differential properties, which are expected to be numerically equivalent. Therefore, we formulate the following DPC loss:

$$\mathcal{L}_d = \frac{1}{HW}\sum_{\boldsymbol{p}} \frac{\left|\boldsymbol{C}^D(\boldsymbol{p}) - \boldsymbol{C}^F(\boldsymbol{p})\right|}{\left|\boldsymbol{C}^D(\boldsymbol{p})\right|}. \tag{15}$$

Specifically, it encourages DepthNet to produce depth maps whose local variations are consistent with those implied by the optical flow maps. Such a consistency constraint enables more reasonable smoothness, especially in those regions where depth and image intensity do not have consistent change trends. Similar to (7), we consider the relative error in (15). The effectiveness of our proposed DPC loss is validated through an ablation study detailed in Sect. IV-E.

D. Bidirectional Stream Co-Adjustment Strategy

In the conventional PCG stream, rigid flow is generated using the outputs from DepthNet and PoseNet, which are indirectly supervised by minimizing the photometric loss. DepthNet, when trained in this manner, often struggles in texture-less regions, where a pixel in the target video frame might correspond to multiple pixels with similar intensities in the source frame. Therefore, we are motivated to improve depth estimation in these regions by leveraging the dense correspondences provided by a well-trained FlowNet.

In our proposed CPG stream, despite the effectiveness of infusing dense correspondence priors provided by a pre-trained FlowNet into monocular depth estimation through our developed CGDC and DPC losses, these priors often capture independent apparent motions unrelated to depth when dynamic objects are involved. Additionally, occlusion regions that lack valid correspondences between consecutive frames also lead to unreliable priors. Therefore, directly leveraging the dense correspondence priors from a pre-trained FlowNet can mislead the DepthNet, due to inaccuracies in geometry-based depth estimation and the misalignment between optical flow divergence and depth gradients on dynamic objects and occluded regions. A straightforward solution to this issue is to exclude dynamic and occluded regions using the auto-masking technique developed in the study [9] when computing CGDC and DPC losses. Nonetheless, as illustrated in Fig. 2(c), this technique is not sufficiently robust, as only the dynamic objects that are relatively stationary with respect to ego-motion can be effectively masked [9], and static regions, especially those with low texture, tend to be overly masked [49]. Moreover, when the dense correspondence priors are of low quality due to insufficient iterative updates [18] or significant domain gaps, directly infusing them into the PCG stream without further adjustment may even impair depth estimation performance across all regions.

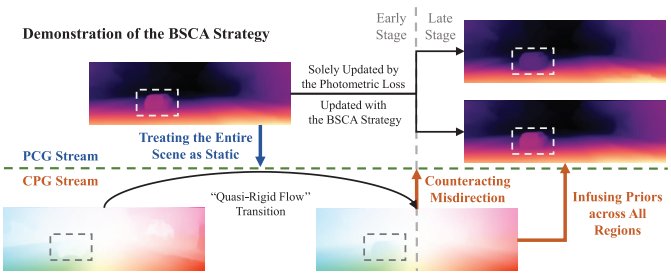

Fig. 3. An illustration of the interaction between PCG and CPG streams through the proposed BSCA strategy to address the challenges posed by dynamic objects.

To address the aforementioned two challenges simultaneously, we propose a simple yet effective BSCA strategy, in which the PCG stream and CPG stream complement each other. We unfreeze the pre-trained FlowNet during training and jointly optimize the optical flow estimated in the CPG stream and the rigid flow generated in the PCG stream by minimizing the following loss function:

$$\mathcal{L}_b = \frac{1}{HW}\sum_{\boldsymbol{p}} \frac{\left\|\boldsymbol{F}_{t\to s}^R(\boldsymbol{p}) - \boldsymbol{F}_{t\to s}^O(\boldsymbol{p})\right\|_1}{\left\|\boldsymbol{F}_{t\to s}^O(\boldsymbol{p})\right\|_1}. \tag{16}$$

In static regions, optical flow and rigid flow should ideally be identical, as demonstrated in prior studies [39], [53]. Similar to the study [39], (16) allows both flows to adjust together without compromising photometric consistency. When DepthNet has not yet been well trained, dense correspondence priors can encourage rigid flow in the PCG stream towards more accurate correspondences. As training progresses, the rigid flow becomes more reliable than the optical flow in occluded regions, where the latter is inherently speculative due to the absence of valid correspondences. (16) then leverages the rigid flow to refine FlowNet's predictions in occluded areas, preventing it from continuously misleading DepthNet. On the other hand, in dynamic regions, optical flow and rigid flow should differ significantly. However, the study [39] introduces an untrained FlowNet to form a joint learning framework, where both FlowNet and DepthNet are trained by minimizing the photometric loss. In this training paradigm, since FlowNet is consistently constrained to capture independent motions, simultaneously minimizing (16) can cause the DepthNet to be further misled in dynamic regions. In contrast, we decouple the FlowNet training from this joint learning framework, generating prior optical flow using a pre-trained FlowNet, which is updated solely by (16). As shown in Fig. 3, by ignoring the photometric consistency constraint on the FlowNet, optical flow in our CPG stream transitions to "quasi-rigid flow" under the guidance of the static scene hypothesis in the PCG stream at the early stage of training. This enables our CGDC and DPC losses in the CPG stream to be effectively applied across the entire image, irrespective of static and dynamic regions. Moreover, as training progresses, photometric loss tends to mislead DepthNet's predictions. At this time, (16) counteracts the misdirection caused by the photometric loss, ensuring that the DepthNet performs accurately in dynamic regions. We

TABLE I

QUANTITATIVE COMPARISON WITH SOTA NETWORKS ON THE KITTI [50], DDAD [35], NUSCENES [51] AND WAYMO OPEN [52] DATASETS. THE BEST RESULTS ARE SHOWN IN BOLD TYPE. THE SYMBOLS ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER PERFORMANCE, RESPECTIVELY. ALL MODELS ARE TRAINED WITH MONOCULAR VIDEO SEQUENCES. THE BASELINE MODELS USED FOR OUR EXPERIMENTS ON EACH DATASET ARE UNDERLINED, RESPECTIVELY

Dataset	Method	Year	Resolution (pixels)	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
KITTI	Monodepth2 [9]	2019	192 × 640	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	DIFFNet [8]	2021	192 × 640	0.102	0.764	4.483	0.180	0.890	0.964	0.983
	Dynamo-Depth [53]	2023	192 × 640	0.112	0.758	4.505	0.183	0.873	0.959	0.984
	SENSE [54]	2023	192 × 640	0.104	0.693	4.294	0.177	0.894	0.965	0.984
	Lite-Mono-8M [14]	2023	192 × 640	0.101	0.729	4.454	0.178	0.897	0.965	0.983
	AQUANet [49]	2024	192 × 640	0.105	**0.621**	**4.227**	0.179	0.889	0.964	0.984
	MonoDiffusion [55]	2024	192 × 640	0.099	0.702	4.385	0.176	0.899	0.966	0.984
	RPrDepth [56]	2024	192 × 640	0.097	0.658	4.279	**0.169**	0.900	0.967	0.985
	DCPI-Depth (Ours)	-	192 × 640	**0.095**	0.662	4.274	0.170	**0.902**	**0.967**	**0.985**
DDAD	Monodepth2 [9]	2019	384 × 640	0.239	12.547	18.392	0.316	0.752	0.899	0.949
	DIFFNet [8]	2021	384 × 640	0.205	12.126	18.461	0.289	0.795	0.916	0.957
	SC-Depth [45]	2021	384 × 640	0.169	3.877	16.290	0.280	0.773	0.905	0.951
	DynamicDepth [57]	2022	384 × 640	0.156	3.305	15.612	0.258	0.785	0.914	0.962
	Lite-Mono [14]	2023	384 × 640	0.161	4.451	16.261	0.271	0.802	0.921	0.962
	Lite-Mono-8M [14]	2023	384 × 640	0.175	6.425	16.687	0.272	0.799	0.920	0.961
	Dynamo-Depth [53]	2023	384 × 640	0.150	3.219	**14.852**	0.246	0.798	0.927	0.969
	SC-DepthV3 [13]	2024	384 × 640	0.142	3.031	15.868	0.248	0.813	0.922	0.963
	DCPI-Depth (Ours)	-	384 × 640	**0.140**	**2.866**	15.786	**0.238**	**0.815**	**0.929**	**0.970**
nuScenes	Monodepth2 [9]	2019	288 × 512	0.425	16.592	10.040	0.402	0.723	0.837	0.887
	DIFFNet [8]	2021	288 × 512	0.228	5.925	8.897	0.290	0.772	0.905	0.950
	MonoViT-tiny [58]	2022	288 × 512	0.412	16.061	10.504	0.385	0.717	0.842	0.898
	Lite-Mono [14]	2023	288 × 512	0.419	15.578	9.807	0.449	0.720	0.831	0.879
	Lite-Mono-8M [14]	2023	288 × 512	0.429	17.058	10.559	0.400	0.709	0.830	0.883
	Dynamo-Depth [53]	2023	288 × 512	0.179	2.118	**7.050**	0.271	0.787	0.896	0.940
	DCPI-Depth (Ours)	-	288 × 512	**0.160**	**1.736**	7.194	**0.248**	**0.793**	**0.921**	**0.966**
Waymo	Monodepth2 [9]	2019	320 × 480	0.173	2.731	7.708	0.227	0.797	0.930	0.968
	DIFFNet [8]	2021	320 × 480	0.149	2.082	7.474	0.200	0.838	0.956	0.981
	Li *et al.* [59]	2021	320 × 480	0.157	1.531	7.090	0.205	–	–	–
	Lite-Mono [14]	2023	320 × 480	0.158	2.305	7.394	0.210	0.816	0.944	0.976
	Lite-Mono-8M [14]	2023	320 × 480	0.154	2.297	7.495	0.209	0.825	0.947	0.975
	Dynamo-Depth [53]	2023	320 × 480	0.116	1.156	6.000	0.166	**0.878**	0.969	0.989
	DCPI-Depth (Ours)	-	320 × 480	**0.116**	**0.963**	**5.642**	**0.162**	0.872	**0.972**	**0.991**

conduct an ablation study in Sect. IV-E to demonstrate the effectiveness of this strategy.

IV. EXPERIMENTS

The performance of our proposed DCPI-Depth is evaluated both qualitatively and quantitatively with extensive experiments in this section. The following subsections provide details on the utilized datasets, practical implementation, evaluation metrics, ablation studies, and comprehensive comparisons with other SoTA methods.

A. Datasets

We conduct our experiments on six public datasets: **KITTI** [50], **DDAD** [35], **nuScenes** [51], **Waymo Open** [52], **Make3D** [60], and **DIML** [61].

For the **KITTI** [50] 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. For the **DDAD** [35] dataset, we follow the prior work [13] to split this dataset into a training set of 12,650 images and a test set of 3,950 images in our experiments. For the **nuScenes** [51] dataset, following the study [53], we use 79,760 image triplets collected by the front camera for model training, and evaluate the model's performance on 6,019 front camera images. For the **Waymo Open** [52] dataset, as in the study [53], we utilize 76,852 front camera image triplets for training, and 2,216 front camera images for evaluation. For the **Make3D** [60] and **DIML** [61] datasets, since neither stereo image pairs nor monocular sequences are provided for unsupervised training, we only use these datasets to quantify the generalizability of models per-trained on the KITTI dataset.

B. Experimental Setup

Our experiments are conducted on an NVIDIA RTX 4090 GPU with a batch size of 12. Following the study [9], we adopt

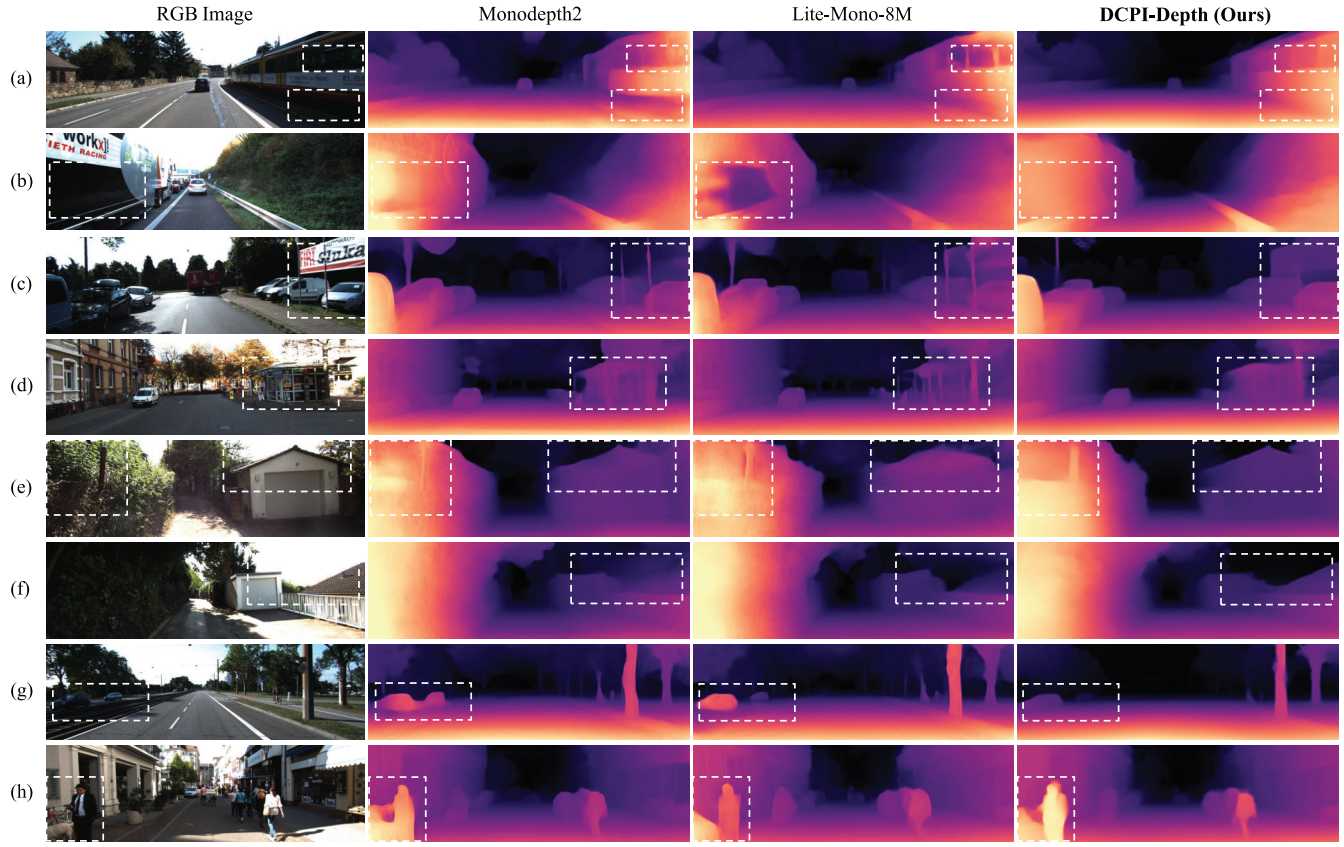

Fig. 4. Qualitative comparisons among Monodepth2, Lite-Mono-8M, and our proposed DCPI-Depth on the KITTI [50] dataset. (a)-(b), (c)-(d), (e)-(f), and (g)-(h) demonstrate the robustness of DCPI-Depth in texture-less regions, in texture-rich regions, at static object boundaries, and on dynamic objects, respectively.

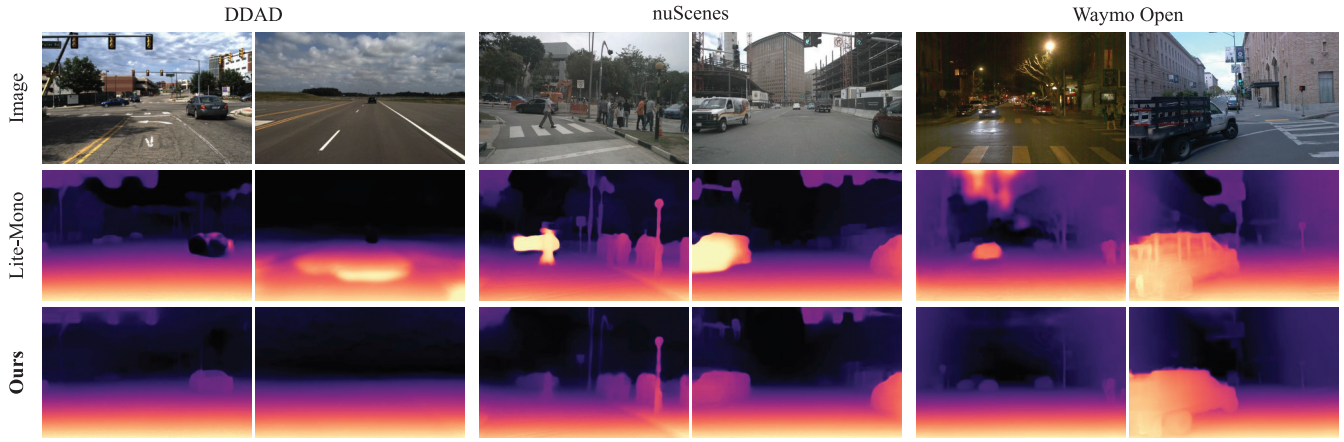

Fig. 5. Qualitative comparisons between Lite-Mono and our proposed DCPI-Depth on the DDAD [35], nuScenes [51], and Waymo Open [52] datasets.

a training approach wherein a snippet of three consecutive video frames is utilized as a training sample. To augment the dataset, random color jitter and horizontal flips are applied to the images during model training. To minimize the loss functions, we employ the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}. The learning rate is adjusted using a cosine annealing scheduler with periodic restarts, decaying from 1×10^{-4} to 5×10^{-6} over 31 epochs. A decay factor of $\gamma = 0.9$ is applied after each cycle to ensure gradual reduction in learning rates. Following the studies [9], [14], the network's encoder is initialized using pre-trained weights from the ImageNet database [64]. RAFT [18] is employed as the FlowNet in our framework to provide dense correspondence prior. It is pre-trained on the KITTI Flow 2015 [65] dataset, which contains only 200 sets of two consecutive frames and has a small overlap with the KITTI Eigen split [5].

TABLE II

QUANTITATIVE COMPARISON WITH SOTA NETWORKS ON THE KITTI [50] DATASET USING THE IMPROVED KITTI GROUND TRUTH FROM [62] FOR MODEL TESTING. THE BEST RESULTS ARE SHOWN IN BOLD TYPE. THE SYMBOLS ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER PERFORMANCE, RESPECTIVELY. "M" DENOTES TRAINING WITH MONOCULAR VIDEO SEQUENCES, AND "MS" DENOTES TRAINING WITH BOTH MONOCULAR VIDEO SEQUENCES AND STEREO IMAGE PAIRS. † INDICATES THE RESULTS ACHIEVED USING THE SAME WEIGHTS IN TABLE I

Method	Year	Resolution (pixels)	Data	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Monodepth2 [9]	2019	192 × 640	MS	0.080	0.466	3.681	0.127	0.926	0.985	0.995
DIFFNet† [8]	2021	192 × 640	M	0.076	0.414	3.492	0.119	0.936	0.988	0.996
RA-Depth [63]	2022	192 × 640	M	0.074	0.363	3.349	0.114	0.940	0.990	0.997
Lite-Mono† [14]	2023	192 × 640	M	0.077	0.413	3.482	0.119	0.933	0.988	0.997
Lite-Mono-8M† [14]	2023	192 × 640	M	0.077	0.423	3.527	0.119	0.934	0.988	0.997
SENSE [54]	2023	192 × 640	M	0.071	0.339	3.175	0.109	0.945	0.990	**0.998**
MonoDiffusion [55]	2024	192 × 640	M	0.073	0.377	3.451	0.115	0.935	0.988	0.997
AQUANet [49]	2024	192 × 640	M	0.070	**0.285**	**2.988**	**0.107**	0.948	**0.992**	**0.998**
DCPI-Depth† (Ours)	-	192 × 640	M	**0.066**	0.326	3.257	**0.107**	**0.949**	0.990	0.997

C. Evaluation Metrics

We employ seven metrics to quantify the model's performance: 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 $i = 1, 2, 3$). Detailed expressions for these metrics can be found in the study [5].

D. Comparison With SoTA Approaches

The quantitative experimental results presented in Table I demonstrate that DCPI-Depth achieves SoTA performance across the KITTI [50], DDAD [35], nuScenes [51], and Waymo Open [52] datasets. Notably, the employed baseline model, Lite-mono [14], that performs unsatisfactorily on these datasets, achieves significant performance improvements, with error reductions ranging from 13.04% to 61.81% on Abs Rel. This substantial enhancement enables DCPI-Depth to surpass previous leading methods [13], [53] on each of the respective datasets, suggesting the effectiveness of our proposed framework.

The qualitative experimental results on the KITTI dataset are shown in Fig. 4. Our DCPI-Depth exhibits superior performance compared to previous SoTA methods. This is particularly evident in texture-less regions, such as (a) and (b). Additionally, our approach ensures smoother and more continuous depth changes and generates clearer boundaries, as exemplified in (c) to (f). These improvements are attributed to the dense correspondence priors infused via our proposed CGDC and DPC losses within the CPG stream, which provide refined and direct geometric cues for depth estimation. Furthermore, DCPI-Depth maintains accurate depth estimation on dynamic objects, such as (g) and (h), unaffected by independent motion. This advantage, observed on dynamic objects, stems from our BSCA strategy, which not only prevents the misleading effects of independent motion captured by the optical flow within the CPG stream but also reinforces accurate estimations in the early training phase, preserving them through to the final results. As shown in Fig. 5, the baseline model performs poorly on the DDAD, nuScenes, and Waymo Open datasets, but demonstrates dramatic performance improvements after being trained under our proposed framework, further demonstrating the robustness of DCPI-Depth across diverse scenarios.

To further investigate the impact of image resolutions and ground truth quality, We evaluate our network's performance using (1) images at a resolution of 320 × 1024 pixels and (2) the improved ground truth [62] from the KITTI dataset. It can be observed in Table II and III that DCPI-Depth consistently achieves SoTA performances compared with existing methods. Although certain metrics exhibit suboptimal results, potentially attributable to the inherent limitation of the network backbone, the proposed training framework, infused with dense correspondence priors, demonstrates a marked improvement in baseline performance and robustness.

To further validate the effectiveness of DCPI-Depth in leveraging dense correspondence priors, we deploy our framework to SC-DepthV3 [13], which leverages pseudo-depth to achieve robust and highly reliable depth estimation. This study also provides a comprehensive protocol for evaluating depth performance across dynamic objects, static areas, and full images. As shown in Table IV, DCPI-Depth achieves significant improvements, outperforming the baseline method by a considerable margin across all metrics for full image, static regions, and dynamic regions. Furthermore, we provide comparisons of learning curves among SC-DepthV1, SC-DepthV3, and DCPI-Depth. As illustrated in Fig. 6(a), DCPI-Depth significantly improves depth estimation performance compared to SC-DepthV3. These results demonstrate that the dense correspondence priors additionally incorporated provide meaningful geometric cues on top of pseudo-depth. Moreover, it can be observed in Fig. 6(b) that the convergence of the photometric loss is comparable among all three models. This observation suggests that merely minimizing photometric loss, which provides an indirect supervisory signal, presents challenges in further improving depth estimation performance. In contrast, our approach provides a more direct and effective constraint for depth estimation.

TABLE III

QUANTITATIVE COMPARISON WITH SOTA NETWORKS ON THE KITTI [50] DATASET USING HIGHER-RESOLUTION IMAGES FOR MODEL TRAINING. THE BEST RESULTS ARE SHOWN IN BOLD TYPE. THE SYMBOLS ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER PERFORMANCE, RESPECTIVELY. "M" DENOTES TRAINING WITH MONOCULAR VIDEO SEQUENCES

Method	Year	Resolution (pixels)	Data	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth2 [9]	2019	320 × 1024	M	0.115	0.882	4.701	0.190	0.879	0.961	0.982
DIFFNet [8]	2021	320 × 1024	M	0.097	0.722	4.435	0.174	0.907	0.967	0.984
MonoViT-tiny [58]	2022	320 × 1024	M	0.096	0.714	4.292	0.172	0.908	0.968	0.984
DaCCN [37]	2023	320 × 1024	M	0.094	0.624	4.145	0.169	0.909	**0.970**	**0.985**
SENSE [54]	2023	320 × 1024	M	0.099	**0.617**	**4.079**	0.172	0.902	0.968	**0.985**
MonoDiffusion [55]	2024	320 × 1024	M	0.095	0.670	4.219	0.171	0.909	0.968	0.984
DCPI-Depth (Ours)	-	320 × 1024	M	**0.090**	0.655	4.113	**0.167**	**0.914**	0.969	**0.985**

TABLE IV

QUANTITATIVE RESULTS ON THE KITTI [50] DATASET FOR THE FULL IMAGES, STATIC REGIONS, AND DYNAMIC REGIONS. THE BEST RESULTS ARE SHOWN IN BOLD TYPE. THE SYMBOLS ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER PERFORMANCE, RESPECTIVELY. "M" DENOTES TRAINING WITH MONOCULAR VIDEO SEQUENCES

Region	Method	Year	Resolution (pixels)	Data	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Full Image	SC-Depth [45]	2021	256 × 832	M	0.118	0.870	4.997	0.197	0.860	0.956	0.981
	SC-DepthV2 [48]	2022	256 × 832	M	0.118	0.861	4.803	0.193	0.866	0.958	0.981
	SC-DepthV3 [13]	2023	256 × 832	M	0.118	0.756	4.709	0.188	0.864	0.960	0.984
	DCPI-Depth (Ours)	-	256 × 832	M	**0.109**	**0.679**	**4.496**	**0.180**	**0.878**	**0.965**	**0.985**
Static Regions	SC-Depth [45]	2021	256 × 832	M	0.106	0.704	4.702	0.170	0.874	0.966	0.989
	SC-DepthV3 [13]	2023	256 × 832	M	0.108	0.636	4.438	0.163	0.881	0.971	0.991
	DCPI-Depth (Ours)	-	256 × 832	M	**0.101**	**0.584**	**4.235**	**0.156**	**0.892**	**0.974**	**0.991**
Dynamic Regions	SC-Depth [45]	2021	256 × 832	M	0.243	3.890	8.533	0.321	0.689	0.849	0.921
	SC-DepthV3 [13]	2023	256 × 832	M	0.205	2.283	7.356	0.290	0.703	0.884	0.945
	DCPI-Depth (Ours)	-	256 × 832	M	**0.186**	**1.948**	**7.028**	**0.281**	**0.732**	**0.895**	**0.950**

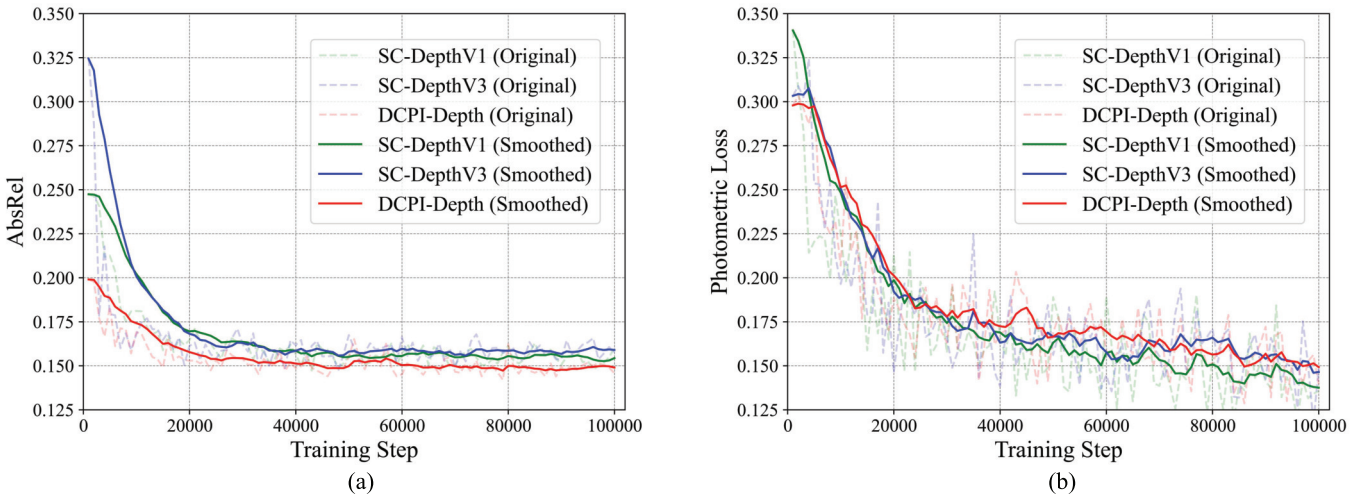

Fig. 6. Learning curve comparisons among SC-DepthV1, SC-DepthV3, and our proposed DCPI-Depth on the KITTI [50] dataset: (a) demonstrates that our DCPI-Depth consistently achieves a lower Abs Rel throughout training compared to other models; (b) illustrates that the convergence of the photometric loss among these models is comparable.

E. Ablation Studies

Table V presents comprehensive ablation studies conducted with SC-DepthV3 to validate the effectiveness of our contributed components. The first ablation study validates the internal design of the CPG stream by comparing the overall performance with and without the incorporation of the CGDC loss and DPC loss, respectively. Key findings from this study include: (1) Incorporating the CGDC loss results in the most significant performance improvements; (2) The DPC loss also plays an important role in improving the performance of our framework. Without the DPC loss, the Abs Rel metric,

TABLE V

ABLATION STUDIES ON THE KITTI [50] AND THE DDAD [35] DATASETS. THE BEST RESULTS ARE SHOWN IN BOLD TYPE. THE SYMBOLS ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES CORRESPOND TO BETTER PERFORMANCE, RESPECTIVELY

DepthNet Backbone	CPG Stream		BSCA Strategy	KITTI (Full Image)				KITTI (Dynamic Regions)			DDAD (Full Image)			
	CGDC loss	DPC loss		Abs Rel ↓	Sq Rel ↓	δ_1 ↑	δ_2 ↑	Abs Rel ↓	Sq Rel ↓	δ_1 ↑	Abs Rel ↓	Sq Rel ↓	δ_1 ↑	δ_2 ↑
ResNet18 [66]	Baseline			0.118	0.756	0.864	0.960	0.205	2.283	0.703	0.149	3.062	0.798	0.920
	✓		✓	0.111	0.698	0.874	0.963	0.193	2.125	0.718	0.143	3.156	0.801	0.915
		✓	✓	0.115	0.716	0.869	0.963	0.201	2.234	0.710	0.148	3.092	0.799	0.921
	✓	✓		0.113	0.707	0.873	0.964	0.211	2.510	0.700	0.145	3.066	0.805	0.919
	✓	✓	✓	**0.109**	**0.679**	**0.878**	**0.965**	**0.186**	**1.948**	**0.732**	**0.143**	**2.963**	**0.812**	**0.925**
DIFFNet [8]	Baseline			0.108	0.696	0.878	0.964	0.197	2.188	0.717	0.145	3.104	0.804	0.922
	✓	✓	✓	**0.103**	**0.653**	**0.886**	**0.966**	**0.191**	**2.098**	**0.719**	**0.137**	**2.812**	**0.816**	**0.926**

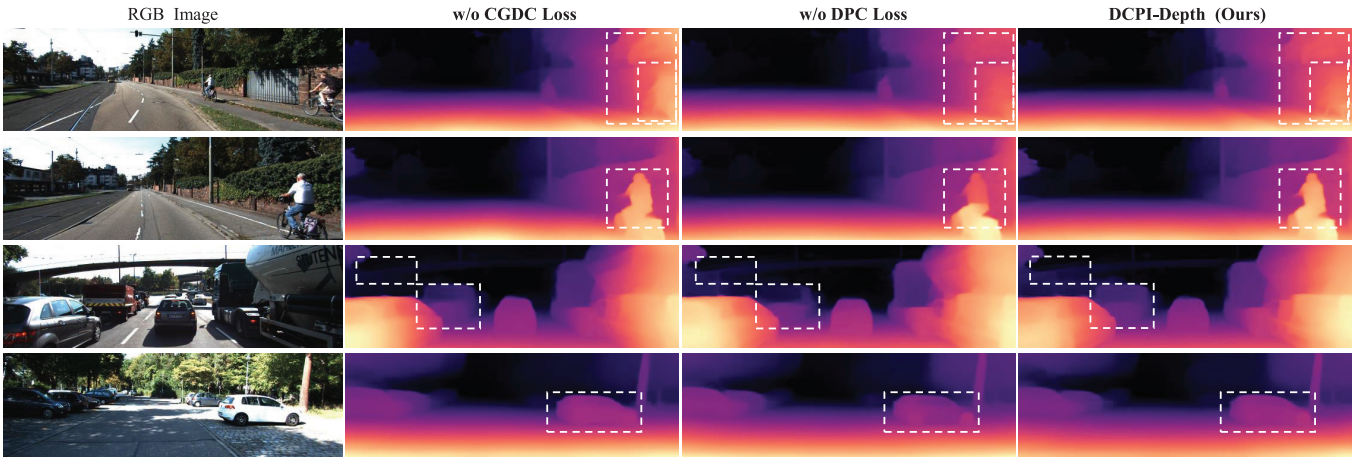

Fig. 7. Qualitative results on the KITTI [50] dataset to demonstrate the effectiveness of the CGDC loss and DPC loss.

which measures the relative depth error in a similar way to the CGDC loss, remains steady. However, other metrics decrease significantly, especially on the DDAD dataset. Fig. 7 provides quantitative experimental results to demonstrate the effectiveness of the CGDC and DPC losses. It is obvious that removing the CGDC loss can lead to significant depth distribution shifts (see the first and third lines), which aligns with our expectations. Furthermore, despite achieving satisfactory quantitative results, removing the DPC loss can result in erroneous depth estimations in continuous regions such as vehicles and cyclists, as illustrated in the second to fourth lines.

Additionally, we conduct another ablation study where we omit the BSCA strategy while retaining the full configuration of the CPG stream to validate its efficacy. The quantitative results reveal a significant decline in performance in dynamic regions, even falling below that of the baseline network, while the depth estimation accuracy in static regions remains unaffected. The qualitative results are provided in Fig. 8. It can be observed that the moving car exhibits independent apparent motions in the optical flow provided by the frozen FlowNet, leading to erroneous depth guidance via triangulation. As a result, the rigid flow eventually tends to contain the independent apparent motions, and the estimated depth is farther than the actual. In contrast, the model trained with the BSCA strategy effectively eliminates independent apparent motions in both flows for the moving car. This results in more accurate geometric and contextual-based depth estimations for the moving car. The above observations are consistent with our initial motivation for introducing the BSCA strategy, emphasizing its critical role in improving depth estimation performance when dynamic objects are involved.

Thirdly, we conduct an ablation study using a better-performing DepthNet backbone within our full configuration. The results show that our contributed techniques are compatible with this backbone and consistently achieve significant improvements. These findings indicate that our contributions provide distinct advantages that differentiate them from those provided by more advanced networks and demonstrate the potential to deliver improvements across a wider range of models.

Finally, the results achieved by our approach with respect to different qualities of dense correspondence priors are presented in Table VI. These results demonstrate that our method remains highly robust when the dense correspondence priors are of low quality, particularly when FlowNet undergoes insufficient iterative updates or is pre-trained solely on synthetic data. This is because even low-quality dense correspondence priors can help perceive relative distances among pixels and reflect depth variations to some extent, thereby providing

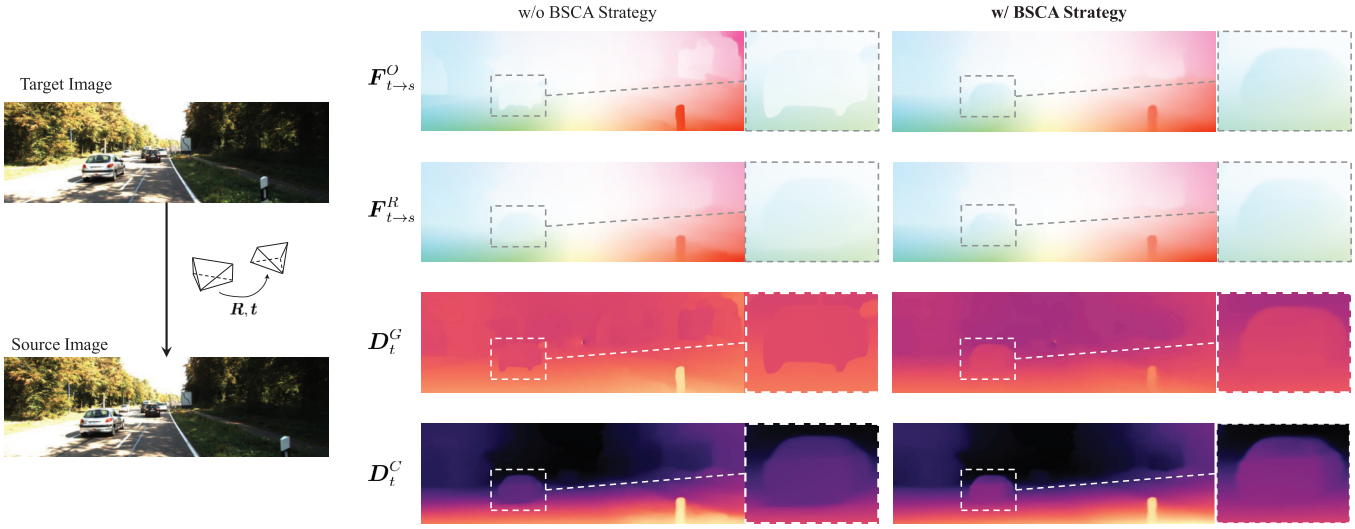

Fig. 8. Qualitative results on the KITTI [50] dataset to demonstrate the effectiveness of the BSCA strategy.

TABLE VI

AN ABLATION STUDY TO DEMONSTRATE THE EFFECTIVENESS OF OUR APPROACH WITH DIFFERENT DENSE CORRESPONDENCE QUALITIES, WHERE FLOWNET IS PRE-TRAINED ON DIFFERENT DATASETS AND APPLIES DIFFERENT NUMBERS OF ITERATIVE UPDATES [18].THE METRIC "F1-EPE" (PIXELS) QUANTIFIES THE ACCURACY OF OPTICAL FLOW ON THE KITTI FLOW 2015 DATASET, WHERE LOWER VALUES INDICATE BETTER PERFORMANCE. DEPTH ESTIMATION IS EVALUATED USING THE KITTI EIGEN SPLIT [5]

Pre-trained Dataset	Iterative Updates	Initial Flow		Final Flow	Depth Estimation	
		Full Image	Static Region	Static Region		
		F1-EPE	F1-EPE	F1-EPE	Abs Rel	δ_1
KITTI Flow 2015	24	**0.63**	**0.66**	4.92	**0.109**	**0.878**
KITTI Flow 2015	1	1.94	1.95	9.73	0.113	0.871
FlyingChairs	24	10.67	10.30	6.53	0.112	0.872
w/o	-	-	-	-	0.118	0.864

TABLE VII

QUANTITATIVE RESULTS ACHIEVED USING ADDITIONAL MONOCULAR DEPTH ESTIMATION MODELS, WITH AND WITHOUT OUR PROPOSED DCPI-DEPTH FRAMEWORK EMPLOYED. WE USE THE KITTI EIGEN SPLIT [5] FOR BOTH MODEL TRAINING AND EVALUATION. ALL MODELS ARE TRAINED USING IMAGES AT A RESOLUTION OF 192 × 640 PIXELS

Method	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
Monodepth2 [9]	0.115	0.903	4.863	0.193	0.877	0.959	0.981
+Ours	**0.110**	**0.749**	**4.559**	**0.183**	**0.879**	**0.962**	**0.984**
Swin-Depth [68]	0.106	0.739	4.510	0.182	0.890	0.964	0.984
+Ours	**0.104**	**0.704**	**4.443**	**0.180**	**0.893**	**0.965**	0.984
DIFFNet [8]	0.102	0.764	4.483	0.180	0.896	0.965	0.983
+Ours	**0.096**	**0.670**	**4.310**	**0.172**	**0.901**	**0.966**	**0.984**
Lite-Mono-3M [14]	0.107	0.765	4.561	0.183	0.886	0.963	0.983
+Ours	**0.101**	**0.707**	**4.447**	**0.176**	**0.893**	**0.964**	**0.984**

valuable guidance during the training of DepthNet. Furthermore, this effectiveness also benefits from the proposed BSCA strategy, which enables the two types of flow to complement each other. Such findings are supported by the final quasi-rigid flow results shown in Table VI. When FlowNet is pre-trained solely on the synthetic FlyingChairs [67] dataset, the BSCA strategy effectively refines correspondences in static regions that are initially affected by either occlusions or domain gaps. Notably, such improvements may not be reflected in the utilized evaluation metrics when the FlowNet is pre-trained on the KITTI Flow 2015 dataset in a supervised manner. This is because unsupervised training provides FlowNet with relatively weaker constraints compared to supervised training, while the KITTI Eigen split [5] used for training differs significantly in data distribution from the dataset used for FlowNet pre-training and evaluation. Fortunately, our approach still helps refine dense correspondence priors across this larger and more diverse data split, as suggested by the observed improvement in depth estimation performance.

F. Generalizability Evaluation

We incorporate the CPG stream and the BSCA strategy into existing open-source methods to further demonstrate the adaptability of our contributions to other SoTA methods. As shown in Table VII, our proposed stream and strategy significantly improve the performance of the original networks across all metrics, which consistently validates the scalability of our methods, as detailed in Sect. IV-E.

Finally, we conduct zero-shot experiments on the Make3D [60] and DIML [61] datasets using the models pre-trained on the KITTI dataset. As shown in Table VIII, our model outperforms all other methods across these two datasets, demonstrating its ability to generalize to new, unseen scenes.

V. DISCUSSION

Despite the effectiveness of our proposed DCPI-Depth framework, it also has two main limitations. First, the

TABLE VIII

QUANTITATIVE RESULTS ON THE MAKE3D [60] AND DIML [61] DATASETS. ALL MODELS ARE TRAINED ON THE KITTI [50] DATASET (IMAGE RESOLUTION: 192 × 640 PIXELS)

Dataset	Method	Abs Rel	Sq Rel	RMSE	RMSE log
Make3D	Monodepth2 [9]	0.321	3.378	7.252	0.163
	HR-Depth [47]	0.305	2.944	6.857	0.157
	R-MSFM6 [69]	0.334	3.285	7.212	0.169
	Lite-Mono [14]	0.305	3.060	6.981	0.158
	MonoDiffusion [55]	0.297	**2.871**	6.877	0.156
	DCPI-Depth (Ours)	**0.291**	2.944	**6.817**	**0.150**
DIML	Monodepth2 [9]	0.185	0.298	1.140	0.249
	HR-Depth [47]	0.183	0.296	1.128	0.248
	R-MSFM6 [69]	0.181	0.301	1.132	0.243
	Lite-Mono [14]	0.173	0.271	1.108	0.239
	MonoDiffusion [55]	0.166	0.256	1.084	0.232
	DCPI-Depth (Ours)	**0.163**	**0.237**	**1.038**	**0.226**

incorporation of the bidirectional PCG and CPG streams inevitably increases the computational and memory overhead during training. Nevertheless, the additional overhead remains within an acceptable range. The introduced FlowNet, which contains 5.3M learnable parameters, and loss functions reduce the training speed from 5.39 it/s to 3.18 it/s. Furthermore, infusing dense correspondence priors into DepthNet through the proposed training framework enhances depth estimation performance while maintaining low computational overhead. Notably, the method preserves a strict monocular setting during the inference phase. Second, while our proposed framework has been shown to be effective when applied to several SoTA monocular depth estimation models, its potential when combined with more advanced network architectures, particularly those based on vision foundation models, remains underexplored and will be left for future work.

VI. CONCLUSION AND FUTURE WORK

This article presented DCPI-Depth, a novel unsupervised monocular depth estimation framework with two bidirectional and collaborative streams: a conventional PCG stream and a newly developed CPG stream. The latter was designed specifically to infuse dense correspondence priors into monocular depth estimation. It consists of a CGDC loss to provide contextual-based depth with geometric guidance obtained from ego-motion and dense correspondence priors, and a DPC loss to constrain the local depth variation using the explicit relationship between the differential properties of depth and optical flow. Moreover, a BSCA strategy was developed to enhance the interaction between the two flow types, encouraging the rigid flow towards more accurate correspondence and making the optical flow more adaptable across various scenarios under the static scene hypotheses. Compared to previous works, our DCPI-Depth framework has demonstrated impressive performance and superior generalizability across six public datasets.

Future work will focus on exploring more advanced optical flow estimation techniques to enhance the reliability and generalizability of dense correspondence priors and to extend the techniques in DCPI-Depth into a joint learning framework where multiple tasks can be explicitly coupled to mutually enhance each other.

REFERENCES

[1] J. Zhang, S. Huang, J. Liu, X. Zhu, and F. Xu, "PYRF-PCR: A robust three-stage 3D point cloud registration for outdoor scene," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 1270–1281, Jan. 2024.

[2] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, and J. Kopf, "Consistent video depth estimation," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 1–71, Aug. 2020.

[3] J. Li, W. Wang, Y. Peng, C. Shen, Y. Zhu, and Z. Xu, "Visual robotic manipulation with depth-aware pretraining," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2024, pp. 843–850.

[4] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Jun. 2014.

[6] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2015.

[7] Z. Huang, Y. Zhang, Q. Chen, and R. Fan, "Online, target-free LiDAR-camera extrinsic calibration via cross-modal mask matching," *IEEE Trans. Intell. Vehicles*, early access, Sep. 11, 2024, doi: 10.1109/TIV.2024.3456299.

[8] H. Zhou, D. Greenwood, and S. Taylor, "Self-supervised monocular depth estimation with internal feature fusion," 2021, *arXiv:2110.09482*.

[9] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3828–3838.

[10] J. Zhang, Y. Liu, G. Ding, B. Tang, and Y. Chen, "Adaptive decomposition and extraction network of individual fingerprint features for specific emitter identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 8515–8528, 2024.

[11] X. Xu, Z. Chen, and F. Yin, "Multi-scale spatial attention-guided monocular depth estimation with semantic enhancement," *IEEE Trans. Image Process.*, vol. 30, pp. 8811–8822, 2021.

[12] Y. Feng, Z. Guo, Q. Chen, and R. Fan, "SCIPaD: Incorporating spatial clues into unsupervised pose-depth joint learning," *IEEE Trans. Intell. Vehicles*, early access, Sep. 16, 2024, doi: 10.1109/TIV.2024.3460868.

[13] L. Sun, J.-W. Bian, H. Zhan, W. Yin, I. Reid, and C. Shen, "SC-DepthV3: Robust self-supervised monocular depth estimation for dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 497–508, Jan. 2024.

[14] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-Mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 18537–18546.

[15] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[16] X. Ye, X. Fan, M. Zhang, R. Xu, and W. Zhong, "Unsupervised monocular depth estimation via recursive stereo distillation," *IEEE Trans. Image Process.*, vol. 30, pp. 4492–4504, 2021.

[17] R. Fan et al., *Autonomous Driving Perception*. Cham, Switzerland: Springer, 2023.

[18] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 402–419.

[19] R. I. Hartley and P. Sturm, "Triangulation," *Comput. Vis. Image Understand.*, vol. 68, no. 2, pp. 146–157, 1997.

[20] J. Han Lee, M.-K. Han, D. Wook Ko, and I. Hong Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv:1907.10326*.

[21] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1043–1051.

[22] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16269–16279.

[23] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2018.

[24] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[25] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jan. 2021, pp. 4009–4018.

[26] S. Shao, Z. Pei, X. Wu, Z. Liu, W. Chen, and Z. Li, "IEBins: Iterative elastic bins for monocular depth estimation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Jan. 2023, pp. 53025–53037.

[27] Y. Duan, X. Guo, and Z. Zhu, "DiffusionDepth: Diffusion denoising approach for monocular depth estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2024, pp. 432–449.

[28] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 10371–10381.

[29] R. Birkl, D. Wofk, and M. Müller, "MiDaS v3.1—A model zoo for robust monocular relative depth estimation," 2023, *arXiv:2307.14460*.

[30] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, vol. 2023, pp. 1–13, Jan. 2023. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt

[31] J. Kopf, X. Rong, and J. Huang, "Robust consistent video depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1611–1621.

[32] J. Liu, L. Kong, J. Yang, and W. Liu, "Towards better data exploitation in self-supervised monocular depth estimation," *IEEE Robot. Autom. Lett.*, vol. 9, no. 1, pp. 763–770, Jan. 2024.

[33] R. Garg, B. G. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 740–756.

[34] Y. Zhang, S. Xu, B. Wu, J. Shi, W. Meng, and X. Zhang, "Unsupervised multi-view constrained convolutional network for accurate depth estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 7019–7031, 2020.

[35] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2485–2494.

[36] Y. Zhang, M. Gong, J. Li, M. Zhang, F. Jiang, and H. Zhao, "Self-supervised monocular depth estimation with multiscale perception," *IEEE Trans. Image Process.*, vol. 31, pp. 3251–3266, 2022.

[37] W. Han, J. Yin, and J. Shen, "Self-supervised monocular depth estimation by direction-aware cumulative convolution network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8613–8623.

[38] Z. He, Y. Zhang, J. Mu, X. Gu, and T. Gu, "LiteGfm: A lightweight self-supervised monocular depth estimation framework for artifacts reduction via guided image filtering," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 8903–8912.

[39] Y. Zou, Z. Luo, and J.-B. Huang, "Df-Net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 36–53.

[40] Z.-C. Yin and J.-P. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1983–1992.

[41] A. Ranjan et al., "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12240–12249.

[42] X. Chen, R. Zhang, J. Jiang, Y. Wang, G. Li, and T. H. Li, "Self-supervised monocular depth estimation: Solving the edge-fattening problem," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5765–5775.

[43] H. Jung, E. Park, and S. Yoo, "Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12642–12652.

[44] J. Zhang, Q. Su, B. Tang, C. Wang, and Y. Li, "DPSNet: Multitask learning using geometry reasoning for scene depth and semantics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 2710–2721, Jun. 2021.

[45] J.-W. Bian et al., "Unsupervised scale-consistent depth learning from video," *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2548–2564, Sep. 2021.

[46] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

[47] X. Lyu et al., "HR-Depth: High resolution self-supervised monocular depth estimation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2294–2301.

[48] J.-W. Bian, H. Zhan, N. Wang, T.-J. Chin, C. Shen, and I. Reid, "Auto-rectify network for unsupervised indoor depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9802–9813, Dec. 2022.

[49] J. L. Gonzalez Bello, J. Moon, and M. Kim, "Self-supervised monocular depth estimation with positional shift depth variance and adaptive disparity quantization," *IEEE Trans. Image Process.*, vol. 33, pp. 2074–2089, 2024.

[50] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[51] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[52] J. Mei et al., "Waymo open dataset: Panoramic video panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2022, pp. 53–72.

[53] Y. Sun and B. Hariharan, "Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Jan. 2023.

[54] G. Li, R. Huang, H. Li, Z. You, and W. Chen, "SENSE: Self-evolving learning for self-supervised monocular depth estimation," *IEEE Trans. Image Process.*, vol. 33, pp. 439–450, 2024.

[55] S. Shao, Z. Pei, W. Chen, D. Sun, P. C. Y. Chen, and Z. Li, "MonoDiffusion: Self-supervised monocular depth estimation using diffusion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3664–3678, Apr. 2025.

[56] W. Han and J. Shen, "High-precision self-supervised monocular depth estimation with rich-resource prior," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2024, pp. 146–162.

[57] Z. Feng, L. Yang, L. Jing, H. Wang, Y. Tian, and B. Li, "Disentangling object motion and occlusion for unsupervised multi-frame monocular depth," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 228–244.

[58] C. Zhao et al., "MonoViT: Self-supervised monocular depth estimation with a vision transformer," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2022, pp. 668–678.

[59] H. Li, A. Gordon, H. Zhao, V. Casser, and A. Angelova, "Unsupervised monocular depth learning in dynamic scenes," in *Proc. Conf. Robot. Learn.*, vol. 155, 16–18, Nov. 2021, pp. 1908–1917.

[60] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2008.

[61] J. Cho, D. Min, Y. Kim, and K. Sohn, "DIML/CVL RGB-D dataset: 2M RGB-D images of natural indoor and outdoor scenes," 2021, *arXiv:2110.11590*.

[62] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant CNNs," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 11–20.

[63] M. He, L. Hui, Y. Bian, J. Ren, J. Xie, and J. Yang, "RA-depth: Resolution adaptive self-supervised monocular depth estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 565–581.

[64] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[65] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3061–3070.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[67] A. Dosovitskiy et al., "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[68] D. Shim and H. J. Kim, "SwinDepth: Unsupervised depth estimation using monocular sequences via Swin transformer and densely cascaded network," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 4983–4990.

[69] J. Yan, H. Zhao, P. Bu, and Y. Jin, "Channel-wise attention-based network for self-supervised monocular depth estimation," in *Proc. Int. Conf. 3D Vis. (3DV)*, Dec. 2021, pp. 464–473.

Mengtan Zhang received the B.S. degree in physics from Tongji University, Shanghai, China, in 2024, where he is currently pursuing the Ph.D. degree with the MIAS Group, Shanghai Research Institute for Intelligent Autonomous Systems, supervised by Prof. Rui Fan. His research interests include computer vision and deep learning.

Yi Feng (Student Member, IEEE) received the B.E. degree in automation from Tongji University, Shanghai, China, in 2022, where he is currently pursuing the Ph.D. degree with the MIAS Group, College of Electronics and Information Engineering, supervised by Prof. Rui Fan. His research interests include computer vision and deep learning.

Qijun Chen (Senior Member, IEEE) received the B.S. degree in automation from the Huazhong University of Science and Technology, Wuhan, China, in 1987, the M.S. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 1999. He is currently a Full Professor with the College of Electronics and Information Engineering, Tongji University. His research interests include robotics control, environmental perception, and understanding of mobile robots and bioinspired control.

Rui Fan (Senior Member, IEEE) received the B.Eng. degree in automation from Harbin Institute of Technology in 2015 and the Ph.D. degree in electrical and electronic engineering from the University of Bristol in 2018. He was a Research Associate at The Hong Kong University of Science and Technology from 2018 to 2020 and a Postdoctoral Scholar-Employee at the University of California at San Diego from 2020 to 2021. He began his faculty career as a Full Research Professor with the College of Electronics and Information Engineering, Tongji University, in 2021. He was promoted to a Full Professor in 2022 and attained tenure in 2024, both in the same college and at Shanghai Research Institute for Intelligent Autonomous Systems. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis. He served as a Senior Program Committee Member for AAAI'23/24/25. He organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, ECCV'22, and ICCV'25. He was honored by being included in the Stanford University List of Top 2% Scientists Worldwide from 2022 to 2024, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, acknowledged as one of Xiaomi Young Talents in 2023, and awarded Shanghai Science and Technology 35 Under 35 honor in 2024 as its youngest recipient. He served as an Area Chair for ICIP'24 and an Associate Editor for ICRA'23/25 and IROS'23/24.