Integrating Disparity Confidence Estimation Into Relative Depth Prior-Guided Unsupervised Stereo Matching

Chuang-Wei Liu, *Student Member, IEEE*, Mingjian Sun, *Member, IEEE*, Cairong Zhao, *Member, IEEE*, Hanli Wang, *Senior Member, IEEE*, Alexander Dvorkovich, *Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Unsupervised stereo matching has garnered significant attention for its independence from costly disparity annotations. Typical unsupervised methods rely on the multi-view consistency assumption for training networks, which suffer considerably from stereo matching ambiguities, such as repetitive patterns and texture-less regions. A feasible solution lies in transferring 3D geometric knowledge from a relative depth map to the stereo matching networks. However, existing knowledge transfer methods learn depth ranking information from randomly built sparse correspondences, which makes inefficient utilization of 3D geometric knowledge and introduces noise from mistaken disparity estimates. This work proposes a novel unsupervised learning framework to address these challenges, which comprises a plug-and-play disparity confidence estimation algorithm and two depth prior-guided loss functions. Specifically, the local coherence consistency between neighboring disparities and their corresponding relative depths is first checked to obtain disparity confidence. Afterwards, quasi-dense correspondences are built using only confident disparity estimates to facilitate efficient depth ranking learning. Finally, a dual disparity smoothness loss is proposed to boost stereo matching performance at disparity discontinuities. Experimental results demonstrate that our method achieves state-of-the-art stereo matching accuracy on the KITTI Stereo benchmarks among all unsupervised stereo matching methods.**

***Index Terms*—Stereo matching, unsupervised learning, knowledge transfer, disparity confidence.**

Received 24 April 2025; revised 30 June 2025 and 15 July 2025; accepted 31 July 2025. Date of publication 8 August 2025; date of current version 22 January 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62473288, Grant 62473286, Grant 62233013, and Grant 62371343; 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 Program (NIO UP); and in part by the Xiaomi Young Talents Program. This article was recommended by Associate Editor W. Wang. *(Corresponding author: Rui Fan.)*

Chuang-Wei Liu is with the College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: cwliu@tongji.edu.cn).

Mingjian Sun is with the Department of Control Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China, and also with Suzhou Research Institute, Harbin Institute of Technology, Suzhou, Jiangsu 215000, China (e-mail: sunmingjian@hit.edu.cn).

Cairong Zhao is with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: zhaocairong@tongji.edu.cn).

Hanli Wang is with the College of Electronics and 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).

Alexander Dvorkovich is with the Multimedia Technology and Telecom Department, Telecommunications Center, Moscow Institute of Physics and Technology, Dolgoprudny, 141701 Moscow, Russia (e-mail: dvork.alex@gmail.com).

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, Shanghai Key Laboratory of Intelligent Autonomous Systems, the State Key Laboratory of Autonomous Intelligent Unmanned Systems, and the Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), 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/TCSVT.2025.3597128, provided by the authors.

Digital Object Identifier 10.1109/TCSVT.2025.3597128

I. Introduction

STEREO matching [3], [4] refers to a process similar to human binocular vision that provides metric depth perception using a rectified stereo vision system [5]. Among all metric depth perception methods, stereo matching has emerged as a prevalent option due to its practicality and cost-effectiveness [6]. As a result, stereo matching has found wide applications related to geometry 3D reconstruction, including autonomous driving [7], [8], augmented reality [9], image compression [10], remote sensing [11], and defect detection [12], [13].

With recent advances in deep learning, researchers have increasingly resorted to convolutional neural networks (CNNs) to address the stereo matching task [14], [15], [16]. These CNN-based approaches are capable of learning and aggregating abstract features directly from input stereo images and output disparity maps in an end-to-end fashion [17], [18], [19], [20]. Witnessing the impressive performance of vision foundation models (VFMs) in deep feature representation and monocular depth estimation, recent supervised studies [21], [22] propose to integrate geometric constraints with robust priors from monocular depth VFMs, thereby achieving improved stereo matching accuracy and generalizability. However, a persistent challenge in this research domain lies in the absence of large-scale real-world datasets with disparity ground truth, which significantly restrains the supervised learning of stereo matching [23]. Therefore, there is a strong motivation to develop powerful unsupervised stereo matching approaches

that are independent of disparity ground truth for learning the best network parameters.

Existing unsupervised stereo matching methods typically rely on the multi-view consistency assumption for network training, such as the commonly adopted photometric loss and geometry consistency loss that enforce the local block or pixel-wise similarity between correspondences [5]. This assumption provides effective constraints for learning 3D geometric knowledge in general scenes, while exhibiting poor robustness in areas with stereo ambiguities, *e.g.*, repetitive patterns and texture-less regions. The matching clues for disparity estimation at these ambiguous areas are insufficient to guide CNNs toward the correct optimization direction, thereby hindering effective parameter convergence [24]. Inspired by the aforementioned VFM-based yet supervised stereo matching methods, this work aims to address stereo matching ambiguities by integrating robust geometric priors derived from monocular depth VFMs into existing unsupervised learning frameworks.

To address the limitations of the multi-view consistency assumption, recent studies [2], [24], [25] aim at transferring 3D geometric knowledge from depth priors or multi-model data into metric depth perception CNNs. For instance, SC-DepthV3 [2] first separates foreground and background areas based on a relative depth map and then establishes correspondence for each foreground pixel with another randomly selected background pixel. Afterwards, a confident depth ranking (CDR) loss is introduced to enforce divergent estimates between the built correspondences. Nevertheless, this knowledge transfer method still suffers from the following limitations:

- The randomly generated correspondences introduce substantial noise to the loss functions, especially those built upon inaccurate disparity estimates.
- The immensely sparse correspondences lead to inefficient exploitation of the 3D geometric knowledge.
- The simplistic disparity divergence strategy fails to provide precise guidance for disparity estimation.

Motivated by these issues, this work aims at enhancing the efficiency and accuracy of the knowledge transfer process by selectively building quasi-dense correspondences with only reliable disparity estimates. We first propose an unsupervised disparity confidence estimation algorithm, named **Disparity-Depth Consistency Voting (DDCV)**. In DDCV, each disparity undergoes a voting process within its neighborhood system, with the average voting score representing its confidence. Each vote is either 0 or 1, denoting a negative or positive vote, respectively, and is derived by checking the local coherence consistency between the disparity pairs and their corresponding relative depth pairs. Afterwards, we introduce a **local depth ranking (LDR)** loss function to learn the depth rankings by building quasi-dense correspondences with reliable disparities as suggested by DDCV. In general, the quasi-dense correspondences enhance the knowledge transfer efficiency, and the reliable disparities suggested by DDCV effectively suppress noise introduced by incorrect estimates. Moreover, LDR loss penalizes only those disparity correspondences that exhibit inconsistency with their corresponding relative depth rankings. This strategy enables our LDR loss to deliver more precise guidance for network parameter convergence. Finally, we propose a novel **dual disparity smoothness (DDS)** loss function to reinforce the consistency between the estimated disparities and relative depth priors in terms of both local smoothness and depth discontinuities. A comparison between the CDR loss in SC-DepthV3 and our proposed LDR loss is presented in Fig. 1. It can be observed that our proposed LDR loss leads to improved disparity smoothness and more distinct boundaries at disparity discontinuities. We conclude the contributions of this study as follows:

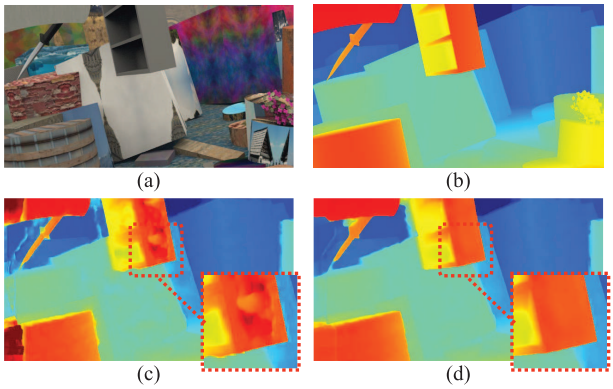

Fig. 1. (a) Left image, (b) disparity ground truth, and estimated disparity maps generated by ViTAStereo [1] trained using (c) the CDR loss in SC-DepthV3 [2] and (d) our proposed LDR loss.

- We discover the significance of disparity confidence in effectively transferring 3D geometric knowledge from relative depth priors to stereo matching networks.
- We propose DDCV, a real-time and plug-and-play algorithm that estimates disparity confidence from any disparity map and its corresponding relative depth map.
- We introduce two relative depth prior-guided loss functions for unsupervised stereo matching, leading to state-of-the-art (SoTA) stereo matching accuracy on the KITTI Stereo benchmark among all existing unsupervised methods.

The remainder of this article is structured as follows: related works, including unsupervised stereo matching and disparity confidence estimation methods, are presented in Sect. II. Sect. III details our proposed DDCV algorithm, LDR, and DDS loss functions. Comprehensive ablation studies and comparative experiments are presented in Sect. IV. Finally, in Sect. V, we summarize the results and provide recommendations for future work. Our supplementary material and code are publicly available at https://mias.group/Un-ViTAStereo

II. Related Work

A. Unsupervised Stereo Matching

Extensive efforts have been devoted to unsupervised stereo matching to eliminate the dependence on costly disparity ground truth. Under the commonly adopted multi-view consistency assumption, these methods typically warp the right image and disparity map into the left view using the estimated left disparity map [26], [27]. Afterwards, a hybrid loss, such as

a combination of photometric loss and geometry consistency loss, is employed to enforce similarity between the reconstructed left views and the original ones [28]. Some works attempt to incorporate occlusion inference into the network training process to enforce the structural understanding [29], [30]. For instance, Co-Teaching [31] employs a co-teaching framework, where two networks with different initializations interactively teach each other to improve occlusion handling. Other studies enhance stereo matching accuracy by jointly learning stereo matching with optical flow estimation and egomotion, without relying on any post-processing. For example, Flow2Stereo [32] trains one single network to estimate both flow and disparity, thereby fully exploiting the inherent geometric constraints from stereoscopic videos. In contrast to existing methods, our work novelly proposes to transfer the 3D geometric knowledge from relative depth priors into a stereo matching network. Notably, our method requires no modifications to the network architecture and optimization of the training data, yet achieves SoTA stereo matching accuracy among all unsupervised methods on public benchmarks.

Extensive efforts have been devoted to unsupervised stereo matching to eliminate the dependence on costly disparity ground truth. Under the commonly adopted multi-view consistency assumption, these methods typically warp the right image and disparity map into the left view using the estimated left disparity map [27]. Afterwards, a hybrid loss, such as a combination of photometric loss and geometry consistency loss, is employed to enforce similarity between the reconstructed left views and the original ones [28]. Some works attempt to incorporate occlusion inference into the network training process to enforce the structural understanding [29], [30]. For instance, Co-Teaching [31] employs a co-teaching framework, where two networks with different initializations interactively teach each other to improve occlusion handling. Other studies enhance stereo matching accuracy by jointly learning stereo matching with optical flow estimation and egomotion, without relying on any post-processing. For example, Flow2Stereo [32] trains one single network to estimate both flow and disparity, thereby fully exploiting the inherent geometric constraints from stereoscopic videos. In contrast to existing methods, our work novelly proposes to transfer the 3D geometric knowledge from relative depth priors into a stereo matching network. Notably, our method requires no modifications to the network architecture and optimization of the training data, yet achieves SoTA stereo matching accuracy among all unsupervised methods on public benchmarks.

B. Knowledge Transfer

Aiming to learn the depth rankings from a relative depth map, our proposed loss functions are thus also related to knowledge transfer methods. For instance, SUW-learn [25] learns 3D geometric knowledge from a pre-calculated semantic segmentation map under the assumption that pixels with different semantic labels exhibit significantly different depth rankings. Specifically, correspondences are established between each pixel and another randomly selected pixel. Afterwards, a loss function is introduced to encourage similar disparity estimates for correspondences with the same semantic label, and divergent estimates for those with different semantic labels. SC-DepthV3 [2] builds upon this strategy by learning more fine-grained depth rankings from a pre-calculated relative depth map, while making minor improvements in correspondence density and the disparity divergence strategy. By contrast, our proposed LDR loss significantly improves knowledge transfer accuracy and efficiency by building quasi-dense correspondences with only confident disparity estimates.

C. Disparity Confidence Estimation

Disparity confidence estimation methods are either explicit programming-based or data-driven [33]. The former ones mainly use cost volumes as the source of confidence information, such as the difference or ratio between two cost minima, as in the peak ratio naive [34] and maximum margin [35]. More recent methods introduce various geometric assumptions to estimate confidence directly from the disparity map. For instance, the disparity agreement [36], [37] assigns higher confidence to disparities that exhibit greater local coherence, and the uniqueness constraint [38] requires a reliable disparity to establish exclusive correspondence in the reference view. Early data-driven methods estimate disparity confidence from small patches of the disparity map, such as CCNN [39], which introduces CNN into this research domain for the first time. The local-global confidence network [40] leverages a U-Net architecture [41] and estimates disparity confidence from both the image and disparity map. SEDNet [42] proposes a differentiable soft-histogramming mechanism to model the error between intermediate multi-scale disparity maps for confidence estimation.

However, most data-driven methods are supervised and rely heavily on disparity ground truth. In contrast, unsupervised learning has been barely investigated in this research domain. Existing unsupervised methods generally use a combination of explicit-based methods to generate pseudo-supervision signals for random forest [43] classifiers or CNNs [38], [44]. The study presented in [45] proposes to generate positive and negative training samples by analyzing the consistency between disparity maps of the same static scene. Despite these efforts, a substantial performance gap remains between unsupervised and supervised CNNs. In contrast, our proposed DDCV is an explicit programming-based and, to the best of our knowledge, the first exploration of relative depth prior-guided disparity confidence estimation algorithm. Without requiring any learnable parameters, DDCV achieves comparable accuracy with recent SoTA supervised disparity confidence estimation methods. Therefore, we regard DDCV as a strong baseline for future powerful relative depth-guided disparity confidence estimation methods.

III. Methodology

Recent advances in VFMs [48], [49] have further empowered knowledge transfer-based unsupervised stereo matching methods by providing robust relative depth priors. Additionally, our previous work, ViT Adapter for Stereo (ViTAS) [1], successfully adapts the VFM features to enhance the performance of a stereo matching network. Therefore, in

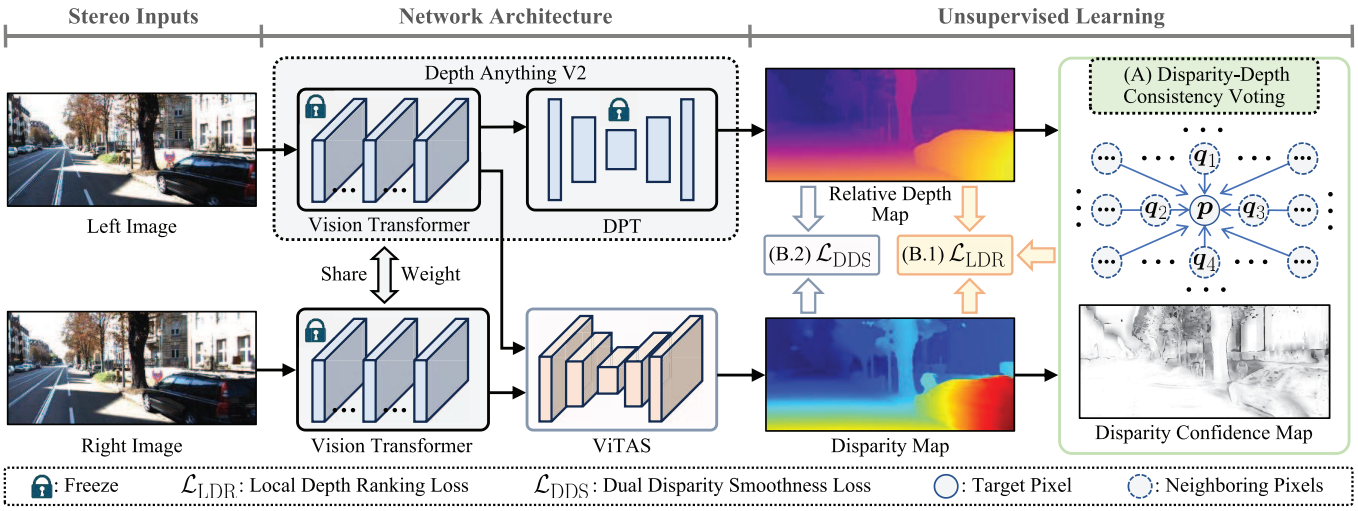

Fig. 2. An illustration of the model architecture and training framework in this work. DPT represents the dense prediction Transformer [46], serving as the decoder of a pre-trained Depth Anything V2 model. Our ViTAStereo shares the feature encoder with the Depth Anything V2 model [47]. The parameters of the feature encoder are frozen.

this work, we incorporate ViTAS and the pre-trained vision Transformer encoder of the Depth Anything V2 model [47] to generate ViTAStereo [1], which is able to generate the disparity map $\boldsymbol{D}$ and its corresponding relative depth map $\widetilde{\boldsymbol{D}}$ at a low computational cost, as visualized in Fig. 2. The yielded disparity and relative depth maps are subsequently used to initialize our unsupervised learning framework. Therefore, this work can also be regarded as the unsupervised learning version of ViTAStereo. Notably, although relative depth maps are required in the network training process, the disparity inference process takes as input only the stereo image pairs, making our unsupervised learning framework compatible with any other existing stereo matching networks. The proposed disparity confidence estimation algorithm, DDCV, is detailed in Sect. III-A. Afterwards, all loss functions adopted in this work are introduced in Sect. III-B.

A. Disparity-Depth Consistency Voting

Building upon the concept of local coherence constraint [50], Markov Random Field (MRF)-based stereo matching methods [6], [51] have remained both prevalent and competitive in this research domain, regardless of the era—before or after the emergence of CNNs. These methods suggest that a crucial factor in accurate disparity estimation lies in discovering disparities that exhibit local coherence with their neighboring disparities. Notably, the relative depths derived from the Depth Anything V2 model provide sufficient local coherence priors between neighboring pixels. Drawing inspiration from the success of these MRF-based methods, we introduce DDCV, a disparity confidence estimation algorithm that models the proportion of neighboring disparity pairs that exhibit local coherence consistency with their corresponding relative depth pairs. In DDCV, the confidence of each disparity is calculated through a dense voting process within its neighborhood system using two proposed local coherence-based voting functions: the ranking consistency voting function $\mathbb{F}_{\text{RC}}$ and the variation consistency voting function $\mathbb{F}_{\text{VC}}$. Each vote

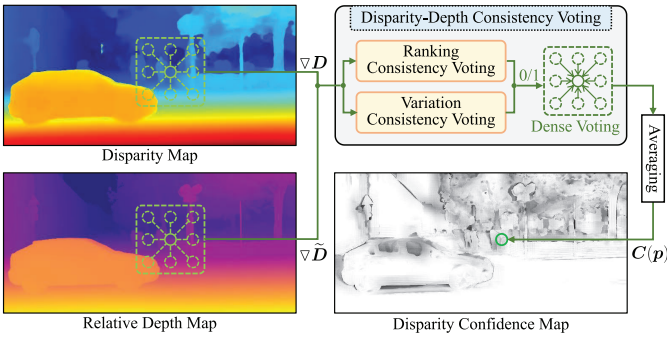

Fig. 3. An illustration of our proposed DDCV.

is either 1 or 0, representing a positive or negative vote, respectively. Finally, the confidence of a disparity is derived by averaging all votes from its neighborhood system.

Given an imperfectly estimated disparity map $\boldsymbol{D}$ and its corresponding relative depth map $\widetilde{\boldsymbol{D}}$. A neighborhood system $\mathcal{N}_p = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m\}$ comprises a collection of 2D pixels centered at a target pixel $\boldsymbol{p}$. $\mathbb{F}_{\text{RC}}$ imposes a local coherence constraint on the depth ranking between each disparity pair and its corresponding relative depth pair within $\mathcal{N}_p$ as follows:

$$\mathbb{F}_{\text{RC}}(\boldsymbol{p}, \boldsymbol{q}) = \Theta(\triangle \widetilde{\boldsymbol{D}}_{pq} \times \triangle \boldsymbol{D}_{pq}), \tag{1}$$

where $\times$ operation denotes the product of two scalar values, $\triangle \widetilde{\boldsymbol{D}}_{pq} = \widetilde{\boldsymbol{D}}(\boldsymbol{p}) - \widetilde{\boldsymbol{D}}(\boldsymbol{q})$ and $\triangle \boldsymbol{D}_{pq} = \boldsymbol{D}(\boldsymbol{p}) - \boldsymbol{D}(\boldsymbol{q})$ represent the disparity variation and depth variation, respectively, and

$$\Theta(x) = \begin{cases} 1, & x \geqslant 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

is a step function to decide the depth ranking consistency vote. In general, a positive depth ranking consistency vote requires the depth ranking between a pair of neighboring pixels to remain consistent between $\boldsymbol{D}$ and $\widetilde{\boldsymbol{D}}$.

The variation consistency voting function $\mathbb{F}_{\text{VC}}$ requires mild disparity variations in regions with stable relative depth, and

significant disparity variations at relative depth discontinuities, as follows:

$$\mathbb{F}_{\mathrm{VC}}(\boldsymbol{p},\boldsymbol{q}) = \Theta(\Theta(|\triangle\widetilde{\boldsymbol{D}}_{pq}| - \gamma \times \sigma) \times (1 - |\triangle\boldsymbol{D}_{pq}|)) \times \Theta(\Theta(\gamma - |\triangle\widetilde{\boldsymbol{D}}_{pq}|) \times (|\triangle\boldsymbol{D}_{pq}| - \sigma)), \quad (3)$$

where

$$\gamma = \frac{\sum_{\boldsymbol{p}\in\widetilde{\boldsymbol{D}}}\sum_{\boldsymbol{q}\in\mathcal{N}_p}|\triangle\widetilde{\boldsymbol{D}}_{pq}|}{\sum_{\boldsymbol{p}\in\boldsymbol{D}}\sum_{\boldsymbol{q}\in\mathcal{N}_p}|\triangle\boldsymbol{D}_{pq}|}, \quad (4)$$

denotes a rough global scale factor, $\sigma \in (1,+\infty)$ is a hyper-parameter to loosen the variation consistency constraint, thereby improving the robustness of $\mathbb{F}_{\mathrm{VC}}$ in areas with a local scale factor that differs significantly from γ. In $\mathbb{F}_{\mathrm{VC}}$, $\Theta(\cdot)$ is also used to identify regions with stable or discontinuous relative depths. Afterwards, the final vote from $\boldsymbol{q}$ to $\boldsymbol{p}$ is calculated as follows:

$$v_{pq} = \mathbb{F}_{\mathrm{RC}}(\boldsymbol{p},\boldsymbol{q}) \times \mathbb{F}_{\mathrm{VC}}(\boldsymbol{p},\boldsymbol{q}), \quad (5)$$

where any local coherence inconsistency results in a negative vote. Finally, the confidence of disparity $\boldsymbol{D}(\boldsymbol{p})$ is computed by averaging all votes from its neighborhood system:

$$\mathcal{C}(\boldsymbol{p}) = \frac{\sum_{\boldsymbol{q}\in\mathcal{N}_p} v_{pq}}{m}. \quad (6)$$

The confidence of all disparities collectively forms the final output confidence map $\mathcal{C}$ of DDCV. More details are illustrated in Fig. 3. Moreover, the relative depths generated by the Depth Anything V2 model present larger values at objects closer to the camera. Therefore, the inverse proportional relationship between the disparity and metric depth is omitted in this work.

B. Unsupervised Loss Functions

The hybrid loss function used in this work comprises four items as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{P}} + \lambda_1 \times \mathcal{L}_{\mathrm{LRC}} + \lambda_2 \times \mathcal{L}_{\mathrm{LDR}} + \lambda_3 \times \mathcal{L}_{\mathrm{DDS}}, \quad (7)$$

where λ_1, λ_2 and λ_3 are three hyper-parameters to weight the contributions of the loss terms, and $\mathcal{L}_{\mathrm{P}}$ and $\mathcal{L}_{\mathrm{LRC}}$, denoting the photometric loss and left-right consistency loss, respectively, are two widely adopted loss functions in unsupervised stereo matching methods [28], [52]. The photometric loss $\mathcal{L}_{\mathrm{P}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{P}} = \frac{1}{N}\sum_{\boldsymbol{p}\in\boldsymbol{I}^L}\Big(0.85 \times \frac{1 - \mathrm{SSIM}(\boldsymbol{I}^L(\boldsymbol{p}), \hat{\boldsymbol{I}}^R(\boldsymbol{p}))}{2} + 0.15 \times \|\boldsymbol{I}^L(\boldsymbol{p}) - \hat{\boldsymbol{I}}^R(\boldsymbol{p})\|_1\Big), \quad (8)$$

where $\boldsymbol{I}^L$ represents the left stereo image, $\hat{\boldsymbol{I}}^R$ is generated by warping the original right stereo image $\boldsymbol{I}^R$ into the left view using the estimated disparity map $\boldsymbol{D}$, N denotes the number of pixels within $\boldsymbol{I}^L$, and SSIM, denoting the structural similarity index measure [53], is used to measure image similarity. The left-right consistency loss $\mathcal{L}_{LRC}$ is defined as follows:

$$\mathcal{L}_{\mathrm{LRC}} = \frac{1}{N}\sum_{\boldsymbol{p}\in\boldsymbol{I}^L}\frac{|\boldsymbol{D}(\boldsymbol{p}) - \hat{\boldsymbol{D}}^R(\boldsymbol{p})|}{\boldsymbol{D}(\boldsymbol{p}) + \hat{\boldsymbol{D}}^R(\boldsymbol{p})}, \quad (9)$$

where $\hat{\boldsymbol{D}}^R$ is derived by warping the right disparity map $\boldsymbol{D}^R$ into the left view using $\boldsymbol{D}$. Our proposed LDR loss $\mathcal{L}_{\mathrm{LDR}}$ and DDS loss $\mathcal{L}_{\mathrm{DDS}}$ are detailed in the subsequent subsections.

1) Local Depth Ranking Loss: We develop our proposed LDR loss $\mathcal{L}_{\mathrm{LDR}}$ based on the disparity continuity hypothesis that disparities change gradually across continuous regions [51]. For each disparity, its neighboring disparities converge toward it from similar values, and the size relationships between them are determined by their corresponding relative depth rankings. Given that most disparities do not correspond to a local depth extremum, their neighboring disparities are distributed on both the greater and smaller sides, allowing these disparities to be precisely confined within a closed interval. As a result, erroneous disparity estimates, particularly those arising from stereo matching ambiguities, can be effectively rectified using these theoretical disparity ranges.

With the disparity confidence map initialized by DDCV, pixels with high confidence disparities are first selected as reference points, serving as local depth anchors for their neighboring pixels. Afterwards, correspondences are established for each disparity with its neighboring reference points. Finally, our proposed $\mathcal{L}_{\mathrm{LDR}}$ learns the 3D geometric knowledge from the relative depth priors by penalizing the disparity correspondences that go against their corresponding relative depth rankings.

Specifically, for each pixel $\boldsymbol{p}$, a top-k operation first identifies a set of k reference points $\mathcal{R}_p = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_k\}$ with the highest disparity confidence within a dilated sampling window centred at $\boldsymbol{p}$. The usage of a dilated sampling window helps prevent spatially clustered reference points compared with a general window, thereby ensuring a more precise value range for $\boldsymbol{D}(\boldsymbol{p})$. Afterwards, we compute $\mathcal{L}_{\mathrm{LDR}}$ between each disparity correspondence as follows:

$$\phi(\boldsymbol{p}) = \frac{\sum_{\boldsymbol{r}\in\mathcal{R}_p}|1 - \Theta(\boldsymbol{p},\boldsymbol{r})| \times \omega(\triangle\widetilde{\boldsymbol{D}}_{pr}) \times \nu(\triangle\boldsymbol{D}_{pr})}{\sum_{\boldsymbol{r}\in\mathcal{R}_p}|1 - \Theta(\boldsymbol{p},\boldsymbol{r})| \times \omega(\triangle\widetilde{\boldsymbol{D}}_{pr})}, \quad (10)$$

where the $|1 - \Theta(\boldsymbol{p},\boldsymbol{r})|$ operation identifies the disparity correspondences that exhibit inconsistent metric depth rankings with their corresponding relative depth rankings,

$$\omega(\triangle\widetilde{\boldsymbol{D}}_{pr}) = \frac{|\triangle\widetilde{\boldsymbol{D}}_{pr}|}{1 + |\triangle\widetilde{\boldsymbol{D}}_{pr}|}, \quad (11)$$

assigns a greater weight to pixel correspondences with substantially different relative depths, and

$$\nu(\triangle\boldsymbol{D}_{pr}) = \log(1 + |\triangle\boldsymbol{D}_{pr}|), \quad (12)$$

requires the disparity estimates to reside within the disparity value ranges constrained by the relative depth rankings. Finally, our proposed $\mathcal{L}_{\mathrm{LDR}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{LDR}} = \frac{1}{N}\sum_{p\in\boldsymbol{I}^L}\phi(\boldsymbol{p}). \quad (13)$$

By building quasi-dense correspondences only with neighboring reference points associated with reliable disparities, our proposed $\mathcal{L}_{\mathrm{LDR}}$ enables full and accurate utilization of relative depth priors in the unsupervised learning of stereo matching.

2) Dual Disparity Smoothness Loss: The conventional image-based disparity smoothness loss [25] enforces local smoothness in disparity estimations by applying an L_1 penalty

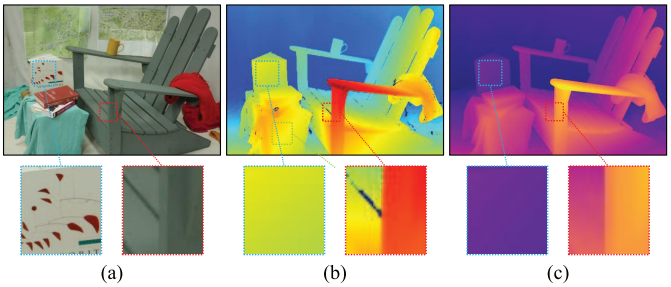

Fig. 4. Illustrations of inconsistency between image intensities and disparities in the (a) left image, (b) ground truth disparity map, and (c) relative depth map. The blue boxes highlight regions with discontinuous pixel intensities but smooth depth variations, while the red boxes indicate regions with smooth image intensities but depth discontinuities.

on the disparity gradients $\partial \boldsymbol{D}$. Given that disparity discontinuities tend to occur at places where image pixel intensity changes greatly, the disparity gradients are weighted with an edge-aware term using the image gradients $\partial \boldsymbol{I}^L$. The general formulation of the image-based disparity smoothness loss is expressed as follows:

$$\mathcal{L}_{\mathrm{DS}}^{I} = \frac{1}{N} \sum_{\boldsymbol{p} \in \boldsymbol{I}^L} |\partial \boldsymbol{D}(\boldsymbol{p})| e^{-\|\partial \boldsymbol{I}^L(\boldsymbol{p})\|}. \tag{14}$$

In general, $\mathcal{L}_{\mathrm{DS}}^{I}$ assumes disparities are locally smooth except in areas with discontinuous image pixel intensities. However, such a simplistic assumption generates incorrect guidance under two common situations: (1) regions with complicated patterns that produce discontinuous image pixel intensities unrelated to depth discontinuities, and (2) depth discontinuities between foreground and background objects that share similar image intensities.

In contrast, monocular depth estimation networks, such as the Depth Anything V2 model, are able to extract and aggregate global texture information, which allows them to overlook the image patterns and accurately identify boundaries between foreground and background objects that have similar image intensities. The generated relative depth map provides a more faithful representation of metric depth smoothness and discontinuities than the original image, as illustrated in Fig. 4. Consequently, we first propose to replace the image gradients with the relative depth gradients, resulting in a relative depth-based disparity smoothness loss as follows:

$$\mathcal{L}_{\mathrm{DS}}^{\widetilde{D}} = \frac{1}{N} \sum_{\boldsymbol{p} \in \boldsymbol{I}^L} |\partial \boldsymbol{D}(\boldsymbol{p})| e^{-\|\partial \widetilde{\boldsymbol{D}}(\boldsymbol{p})\|}. \tag{15}$$

Moreover, leveraging the additional depth discontinuity information provided by the relative depths, we further enhance the relative depth-based disparity smoothness loss with a dual term, which gives a large penalty at areas with discontinuous relative depths but smooth disparities. Finally, our proposed $\mathcal{L}_{\mathrm{DDS}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{DDS}} = \frac{1}{N} \sum_{\boldsymbol{p} \in \boldsymbol{I}^L} |\partial \boldsymbol{D}(\boldsymbol{p})| e^{-\|\partial \widetilde{\boldsymbol{D}}(\boldsymbol{p})\|} + |\partial \widetilde{\boldsymbol{D}}(\boldsymbol{p})| e^{-\|\partial \boldsymbol{D}(\boldsymbol{p})\|}. \tag{16}$$

Compared with the conventional $\mathcal{L}_{\mathrm{DS}}^{I}$, $\mathcal{L}_{\mathrm{DDS}}$ enhances the consistency of both local smoothness and local discontinuity between the estimated disparities and relative depth priors, thereby resulting in better stereo matching performance at depth discontinuities.

IV. Experiments

A. Datasets and Implementation Details

Five public stereo matching datasets are utilized in our experiments for model training and performance evaluation, including two synthetic large-scale datasets:

1) **SceneFlow** [54] consists of a training set (containing 35,454 stereo image pairs) and a test set (often known as the Flying 3D test set, containing 4,370 stereo image pairs) with image resolution of 960×540 pixels.
2) **Virtual KITTI** [55] contains 21,260 stereo image pairs (resolution: $1,242 \times 375$ pixels), generated from five different virtual worlds (created using the Unity game engine and a real-to-virtual cloning method) in urban settings under different imaging and weather conditions,

and three real-world yet small-scale datasets:

1) **KITTI Stereo** contains two subsets: 2012 [56] and 2015 [57], with 192 and 200 training pairs, respectively, and 194 and 200 test pairs, respectively. The image resolution is around $1,240 \times 370$ pixels.
2) **Middlebury** [58] contains 14 pairs of indoor stereo images and is widely recognized for its complicated 3D geometry structures.
3) **ETH3D** [59] contains 27 pairs of stereo grayscale images of both indoor and outdoor scenes (resolution: around 930×490 pixels).

We first conduct ablation studies on the SceneFlow dataset to validate the effectiveness of each proposed component. Subsequently, ViTAStereo is trained on all five datasets and evaluated on the test sets of the KITTI Stereo datasets. Furthermore, we perform comprehensive comparisons between our proposed DDCV and other SoTA disparity confidence estimation methods on the three real-world datasets under various disparity input settings. We select eight disparities with the highest confidences as reference points in $\mathcal{L}_{\mathrm{LDR}}$ (top-k with $k = 8$) as suggested by our ablation study. We also set $\alpha = 2$ in (3), $\lambda_1 = \lambda_3 = 0.1$ and $\lambda_2 = 0.1$ in (7), and use an 11×11 sampling kernel in both DDCV and $\mathcal{L}_{\mathrm{LDR}}$. These parameters are intuitively chosen without extensive hyperparameter tuning. All experiments are conducted on an NVIDIA RTX 4090 GPU. During model training, we randomly crop images to 320×720 pixels and apply conventional data augmentation techniques, including random changes in image color, random rescaling, and random erasing.

B. Evaluation Metrics

Three commonly adopted metrics are computed to quantify stereo matching accuracy: **End-point error (EPE)** measures the average disparity estimation error, **Percentage of error pixels (PEP)** indicates the percentage of incorrect disparities with respect to a tolerance of δ pixels, and **D1** counts the percentage of disparities for which the estimation error exceeds both three pixels and 5% of the ground-truth disparity.

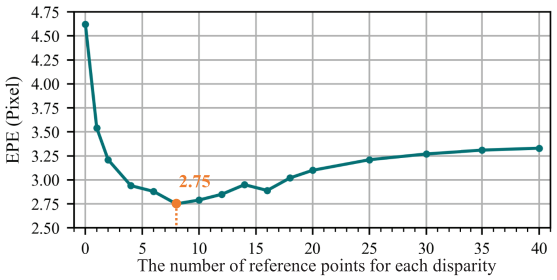

Fig. 5. Ablation study on the top-k operation in $\mathcal{L}_{\text{LDR}}$ w.r.t. different k.

TABLE I
ABLATION STUDIES ON THE SCENEFLOW TEST DATASET. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Depth Ranking		Disparity Smoothness			EPE (pixel)	D1 (%)
$\mathcal{L}_{\text{CDR}}$	$\mathcal{L}_{\text{LDR}}$	$\mathcal{L}_{\text{DS}}^{I}$	$\mathcal{L}_{\text{DS}}^{\widetilde{D}}$	$\mathcal{L}_{\text{DDS}}$		
					4.91	16.6
✔					3.46	13.7
	✔				3.09	12.4
✔	✔				3.08	12.2
	✔	✔			2.88	11.5
	✔	✔	✔		2.80	11.0
	✔		✔		2.81	11.0
	✔			✔	**2.75**	**10.7**

Additionally, we generate the density-EPE receiver operating characteristic (ROC) curves and then use the **area under the curve (AUC)** to evaluate the disparity confidence estimation accuracy following previous work [42]. Specifically, the ROC curves in our case measure EPE by successively incorporating disparities in descending order of their confidence. We also introduce the optimal AUC to count the area under the optimal ROC curve that incorporates disparities in ascending order of their EPE values. Moreover, **time/MP** that measures the runtime in megapixels is used to evaluate the disparity confidence estimation efficiency. For all metrics used in our experiments, lower values indicate better performance.

C. Ablation Study

We first investigate the optimal configuration for the top-k operation in $\mathcal{L}_{\text{LDR}}$, referring to the number of reference points for each disparity. As shown in Fig. 5, increasing k from zero initially leads to a marked decrease in EPE. This improvement is attributed to the enhanced availability of 3D geometric information provided by a greater number of reference points, which facilitates more accurate localization of the target disparity. However, an excessively large k leads to a sustained increase in EPE. We attribute this degradation to the incorporation of low-confidence disparities into the reference points, which introduces noise into the computation of $\mathcal{L}_{\text{LDR}}$. Therefore, we set $k = 8$, at which the EPE witnesses the first increase, in all subsequent experiments.

Another ablation study is conducted to demonstrate the superiority of our proposed $\mathcal{L}_{\text{LDR}}$ and $\mathcal{L}_{\text{DDS}}$ over other loss functions related to depth ranking and disparity smoothness, respectively, including the CDR loss $\mathcal{L}_{\text{CDR}}$ proposed in SC-DepthV3 [2], the image-based disparity smoothness loss $\mathcal{L}_{\text{DS}}^{I}$, and the depth-based disparity smoothness loss $\mathcal{L}_{\text{DS}}^{\widetilde{D}}$. The baseline setup includes the photometric loss and left-right consistency loss. As presented in Table I, individually deploying either the CDR loss or LDR loss leads to improved stereo matching accuracy. Specifically, the LDR loss reduces the EPE by approximately 10.7% compared to the CDR loss. Additionally, the incorporation of both $\mathcal{L}_{\text{CDR}}$ and $\mathcal{L}_{\text{LDR}}$ yields similar results to those of using $\mathcal{L}_{\text{LDR}}$ separately, which signifies the ability of $\mathcal{L}_{\text{LDR}}$ to fully substitute for $\mathcal{L}_{\text{CDR}}$. In terms of the disparity smoothness, our proposed $\mathcal{L}_{\text{DDS}}$ reduces the EPE by 4.5% and 2.1% compared with $\mathcal{L}_{\text{DS}}^{I}$ and $\mathcal{L}_{\text{DS}}^{\widetilde{D}}$, respectively. In general, $\mathcal{L}_{\text{LDR}}$ significantly enhances the stereo matching accuracy of ViTAStereo, while $\mathcal{L}_{\text{DDS}}$ provides additional performance gains.

We further present qualitative results of ViTAStereo trained using $\mathcal{L}_{\text{LDR}}$ and $\mathcal{L}_{\text{DDS}}$ in Fig. 6. It can be observed that $\mathcal{L}_{\text{DDS}}$ improves the disparity estimation accuracy in both disparity discontinuities and locally smooth regions, aligning well with our intention in Sect. III-B.2. By contrast, $\mathcal{L}_{\text{LDR}}$ exhibits effectiveness in addressing stereo matching ambiguities, including texture-less regions and even the occluded region on the left side of the left image.

D. Disparity Confidence Estimation

We conduct comparative experiments between DDCV and other disparity confidence estimation methods, including CCNN [39], OTB [38], and SEDNet [42], across various datasets and disparity inputs. All three methods are CNN-based, where CCNN and SEDNet are trained using disparity ground truth, while OTB is trained in an unsupervised manner. The quantitative experimental results are presented in Tables II and III. It can be observed that DDCV outperforms CCNN, LGC, and OTB across all metrics and achieves comparable AUC scores to those of SEDNet. Specifically, DDCV decreases the AUC score by up to 36.2%, 28.8%, and 38.4% compared with CCNN, LGC, and OTB, respectively. In terms of efficiency, although SEDNet reduces time/MP by 29.6% compared to DDCV, our method still exhibits real-time performance at the millisecond level. The qualitative experimental results presented in Fig. 7 show the superiority of our proposed DDCV in identifying erroneous disparities at disparity discontinuities, occlusions, and small-scale objects. In general, our proposed DDCV achieves SoTA disparity confidence estimation accuracy without relying on CNNs and disparity ground truth. We believe that DDCV is able to serve as a strong baseline for future relative depth prior-guided disparity confidence estimation methods.

It is noteworthy that when using disparity maps generated by our unsupervised ViTAStereo (Un-ViTAStereo) as inputs, DDCV achieves lower AUC scores than SEDNet on the KITTI Stereo and Middlebury datasets. The qualitative results on the Middlebury dataset exhibit the superiority of DDCV in handling disparity estimations at occluded areas. The underlying reason is that SEDNet estimates disparity

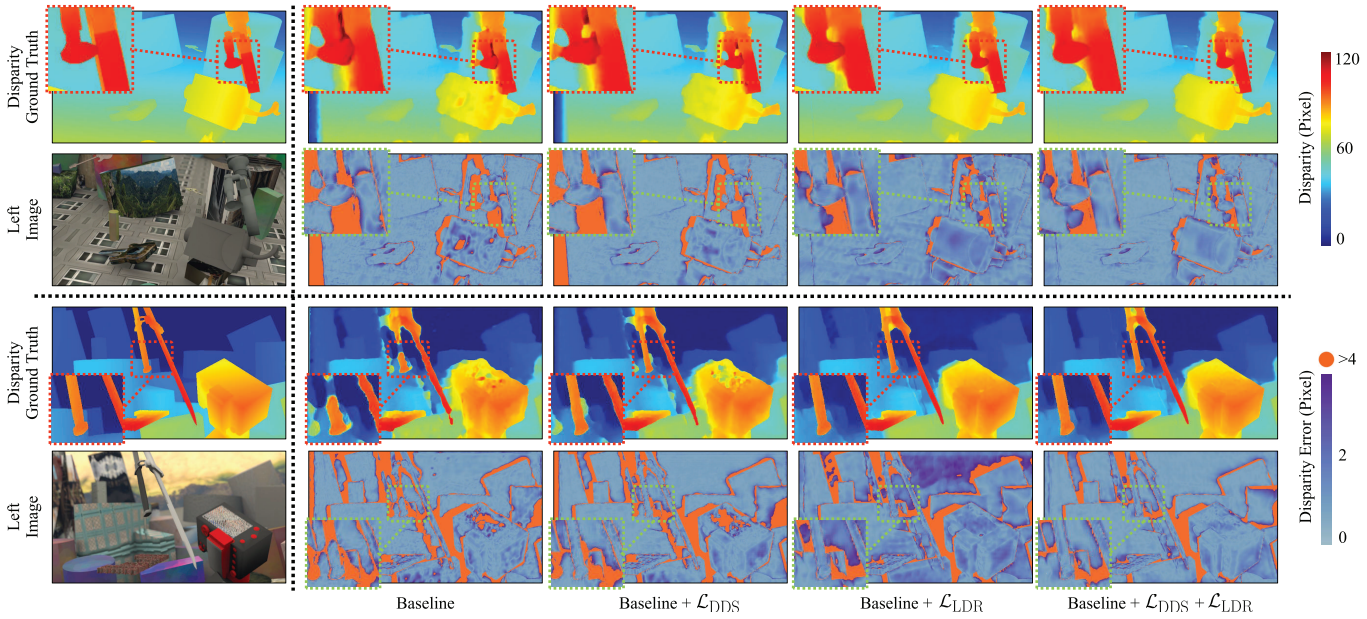

Fig. 6. Examples of disparity estimation results with or without our proposed $\mathcal{L}_{\mathrm{DDS}}$ and $\mathcal{L}_{\mathrm{LDR}}$, where detailed regions are shown with dashed boxes.

TABLE II

COMPARISONS WITH SOTA DISPARITY CONFIDENCE ESTIMATION METHODS. DISPARITY MAP INPUTS ARE GENERATED BY SEDNET [42], OUR UN-VITASTEREO, IGEVSTEREO [60] AND CRESTEREO [61]

Dataset	Input Disparity Map		AUC					
	Method	EPE (pixel)	CCNN [39]	LGC [40]	OTB [38]	SEDNet [42]	**DDCV (ours)**	Optimal
KITTI	SEDNet [42]	1.16	72.5	67.6	73.1	66.9	**62.1**	26.4
	Un-ViTAStereo [1]	1.02	64.2	58.4	74.0	55.9	**54.4**	25.1
	IGEVStereo [60]	1.01	75.0	70.3	79.5	**66.8**	69.0	30.1
	CreStereo [61]	0.92	56.2	59.2	64.5	**45.9**	51.2	22.4
Middlebury	SEDNet [42]	2.24	111	99.5	115	**52.1**	70.8	27.2
	Un-ViTAStereo [1]	1.96	114	103	113	81.9	**75.4**	30.1
	IGEVStereo [60]	1.01	76.9	65.3	70.1	**44.6**	46.1	16.7
	CreStereo [61]	0.64	37.1	31.9	35.7	27.8	**26.5**	9.51
ETH3D	SEDNet [42]	0.78	37.3	37.1	36.8	37.6	**36.7**	13.8
	Un-ViTAStereo [1]	0.36	27.8	25.7	24.9	**17.9**	21.7	9.13
	IGEVStereo [60]	0.09	8.38	7.50	7.92	**5.89**	7.49	2.72
	CreStereo [61]	0.21	15.8	15.0	14.2	**12.0**	13.9	5.68

TABLE III

COMPARISONS WITH SOTA DISPARITY CONFIDENCE ESTIMATION METHODS IN TERMS OF DISPARITY CONFIDENCE ESTIMATION EFFICIENCY

Method	CCNN [39]	OTB [38]	SEDNet [42]	**DDCV (ours)**
time/MP (ms)	47.2	678	**6.42**	9.13

confidence based on the inconsistency between the estimated multi-scale disparities. However, in occluded regions, unsupervised stereo matching networks tend to produce erroneous yet consistent disparities, rendering the multi-scale inconsistency strategy less effective in detecting these unreliable estimates. In contrast, relative depth maps generated by monocular depth estimation networks are able to provide accurate depth ranking information in occluded regions. As a result, our proposed DDCV exhibits robust disparity confidence estimation performance in such regions. Additionally, the absence of ground-truth disparities in occluded regions within the ETH3D dataset accounts for the exceptional case on this dataset. Therefore, our proposed DDCV stands out as a more suitable and effective alternative for integrating unsupervised stereo matching methods.

E. Unsupervised Stereo Matching

Upon submitting our best results to the online KITTI Stereo 2015 and 2012 benchmark suites, we conduct a comparative analysis with other unsupervised stereo matching networks

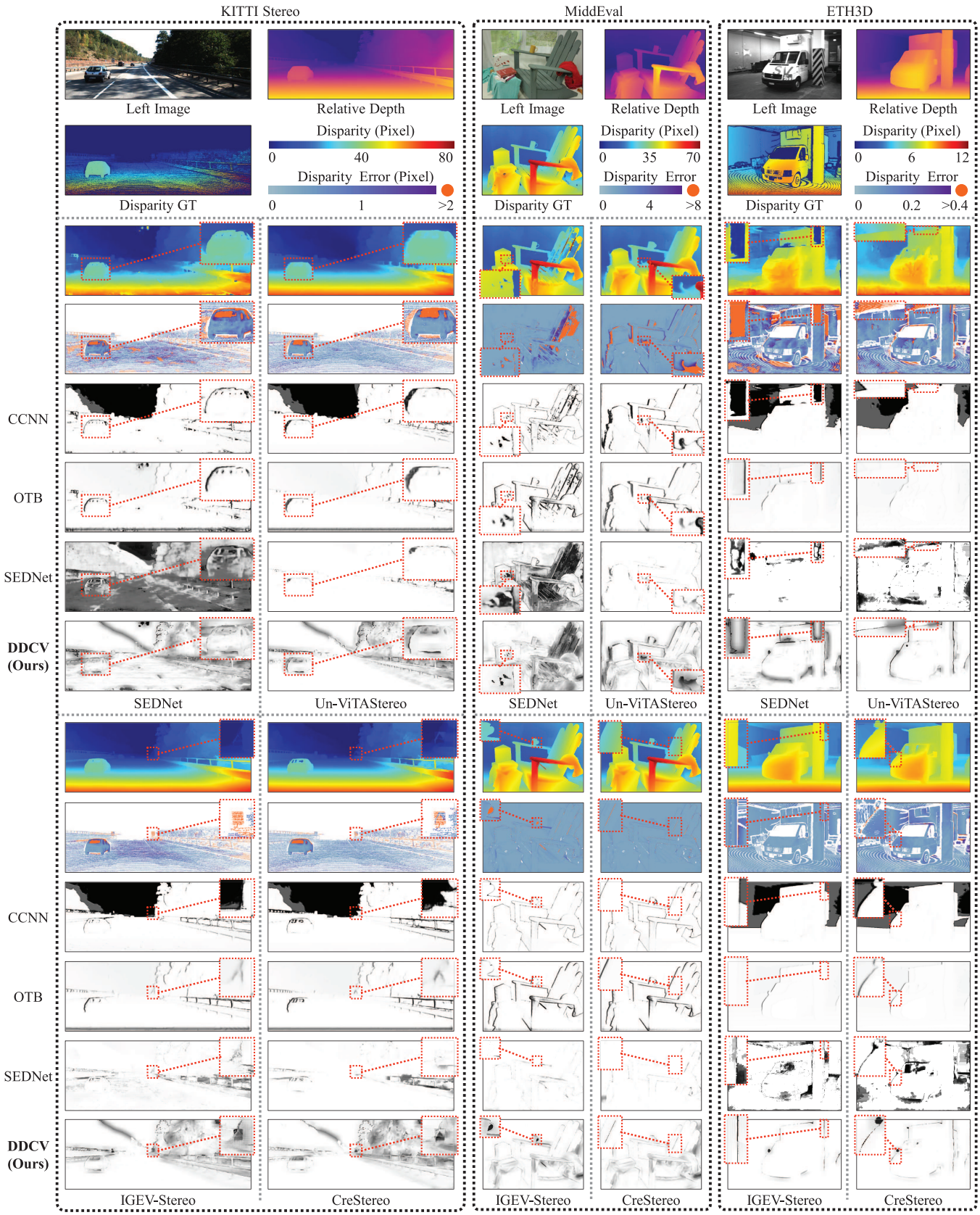

Fig. 7. Qualitative experimental results of CCNN [39], OTB [38], SEDNet [42], and our proposed DDCV on the KITTI Stereo, Middlebury, and ETH3D datasets, where significantly improved regions are shown with dashed boxes.

TABLE IV

COMPARISONS WITH SOTA UNSUPERVISED STEREO MATCHING NETWORKS PUBLISHED ON THE KITTI STEREO BENCHMARK [57]. "NOC" DENOTES METRICS FOR NON-OCCLUDED PIXELS. D1-BG, D1-FG, AND D1-ALL DENOTE D1 FOR BACKGROUND, FOREGROUND, AND ALL PIXELS, RESPECTIVELY

Network	KITTI Stereo 2012				KITTI Stereo 2015						Runtime (s)
	ALL Pixels		Noc Pixels		All Pixels			Noc Pixels			
	EPE (px)	PEP-3 (%)	EPE (px)	PEP-3 (%)	D1-bg (%)	D1-fg (%)	D1-all (%)	D1-bg (%)	D1-fg (%)	D1-all (%)	
SegStereo [24]	2.4	9.64	1.6	7.89	-	-	8.79	-	-	7.70	0.60
Flow2Stereo [32]	1.1	5.11	1.0	4.58	5.01	14.62	6.61	4.77	14.03	6.29	**0.05**
OASM-Net [29]	2.0	8.60	1.3	6.39	6.89	19.42	8.98	5.44	17.30	7.39	0.73
PASMnet_192 [63]	-	-	-	-	5.41	16.36	7.23	5.02	15.16	6.69	0.50
OASM-DDS [30]	1.2	5.69	1.1	4.81	4.46	15.76	6.51	-	-	6.02	-
Tong et al. [64]	-	-	-	-	6.55	15.94	8.11	5.86	14.68	7.32	0.17
SDCO [65]	-	-	-	-	-	-	-	5.28	17.35	7.27	-
Permutation-Stereo [62]	1.8	8.48	1.6	7.39	5.53	15.47	7.18	5.18	14.51	6.72	30.0
SPSMnet [27]	-	-	-	-	5.42	12.84	6.65	4.94	12.01	6.10	0.80
Un-ViTAStereo (ours)	**0.9**	**4.16**	**0.8**	**3.46**	**3.58**	**12.30**	**5.03**	**3.28**	**11.32**	**4.61**	0.22

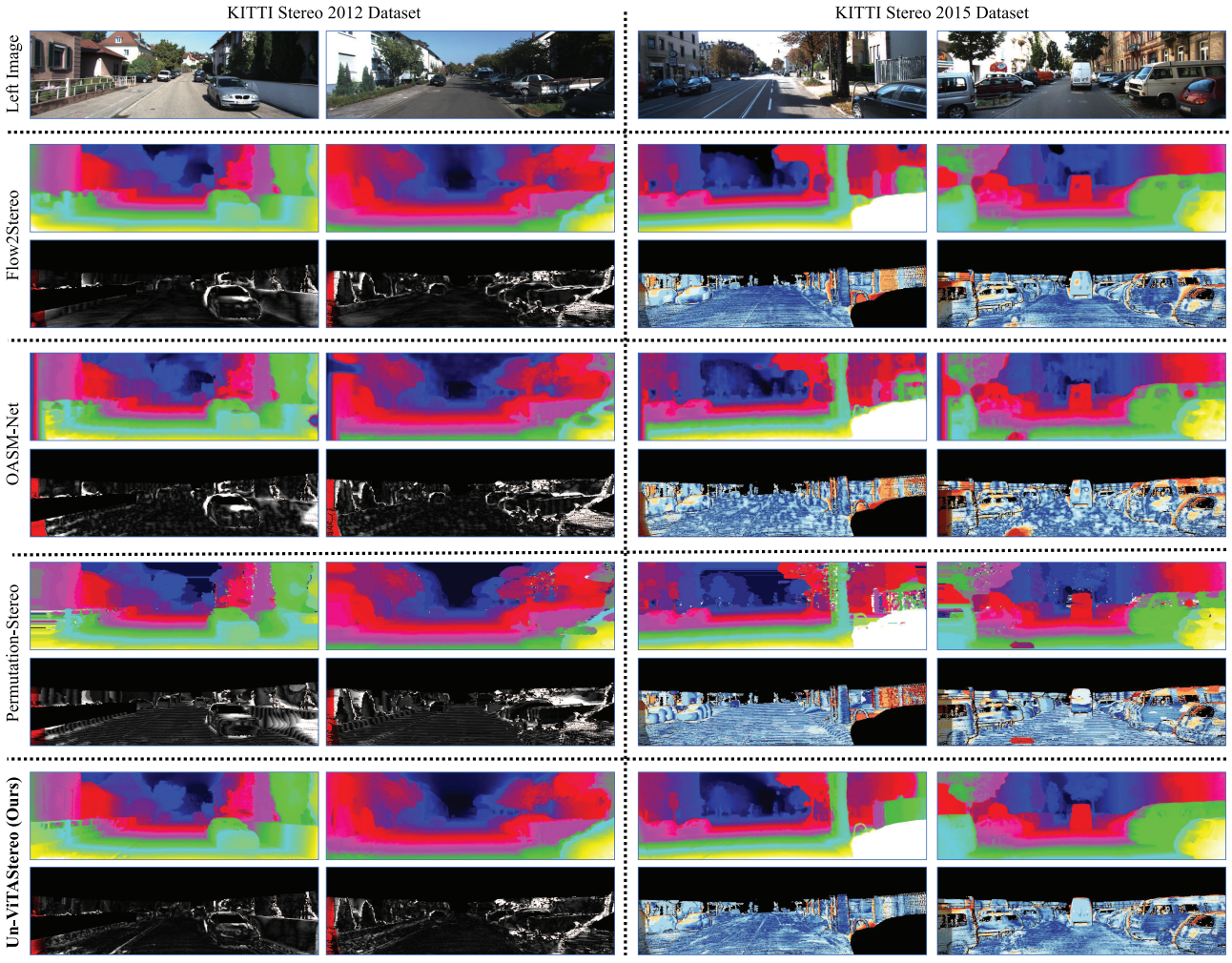

Fig. 8. Qualitative experimental results of Flow2Stereo [32], OASM-Net [29], Permutation-Stereo [62] and this work on the KITTI Stereo datasets [56], [57]. The images in the first row of each method represent the estimated disparity maps and images in the second row denote the visualizations of D1 error. We collect these images from the online benchmark suites.

published on these benchmarks, as presented in Table IV. Surprisingly, our Un-ViTAStereo outperforms all other SoTA stereo matching networks in terms of all evaluation metrics regarding stereo matching accuracy. Specifically, on the KITTI Stereo 2015 benchmark, Un-ViTAStereo outperforms OASM-DDS [30], the second-best network, by up to 22.7% and 23.4% in D1 for all pixels and non-occluded pixels, respectively. On the KITTI Stereo 2012 benchmark, Un-ViTAStereo reduces the PEP-3 and EPE by up to 24.5% and 20%, respectively, compared with Flow2Stereo [32].

The qualitative experimental results, as illustrated in Fig. 8, further demonstrate that our Un-ViTAStereo surpasses other networks in handling stereo matching ambiguities. This advantage is evident in both texture-less regions (illustrated in columns 1 and 2) and repetitive patterns (illustrated in column 3). Additionally, our Un-ViTAStereo achieves higher stereo matching accuracy at disparity discontinuities, such as the vehicle boundaries and background regions (illustrated in column 4). These significant performance gaps demonstrate the effectiveness and efficiency of our proposed $\mathcal{L}_{\text{LDR}}$ and $\mathcal{L}_{\text{DDS}}$ in transferring 3D geometric knowledge from relative depths to a stereo matching network. More visualizations on the KITTI Stereo, Middlebury and ETH3D datasets are provided in our supplementary material at https://mias.group/Un-ViTAStereo to facilitate further evaluation of the robustness of our proposed Un-ViTAStereo.

V. Conclusion and Future Work

This article introduces a relative depth prior-guided unsupervised stereo matching method, comprising a plug-and-play disparity confidence estimation algorithm and two novel 3D geometric knowledge transfer loss functions. By building quasi-dense correspondences with only reliable disparity estimates, our method improves the knowledge transfer efficiency and accuracy. Our study has yielded several key findings: (1) local coherence consistency between depth priors and estimated disparities is highly indicative of disparity estimation accuracy, and (2) building correspondences with only reliable disparity estimates effectively suppresses noise in the knowledge transfer process. Our method achieves both SoTA stereo matching accuracy on the KITTI Stereo benchmarks, as well as SoTA disparity confidence estimation accuracy among unsupervised methods.

While the contributions of this study are significant, it is noted that the design of DDCV remains empirical. For instance, we assume pixel correspondence with disparity variation less than 1 pixel as disparity stable. Additionally, the large number of hyper-parameters makes it challenging for our method to achieve optimal performance. Therefore, in future work, we intend to build upon the above-mentioned findings by developing a disparity confidence estimation network that adaptively learns all parameters in an unsupervised manner. Additionally, given that LiDAR can efficiently generate sparse disparity ground truth, enabling the development of real-world datasets including KITTI Stereo, we aim to further explore the potential of LDR loss in supervised stereo matching by selecting pixels with available disparity annotations as reference points.

References

[1] C.-W. Liu, Q. Chen, and R. Fan, "Playing to vision foundation model's strengths in stereo matching," *IEEE Trans. Intell. Vehicles*, early access, Sep. 25, 2024, doi: 10.1109/TIV.2024.3467287.

[2] 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.

[3] Z. Lu, J. Wang, Z. Li, S. Chen, and F. Wu, "A resource-efficient pipelined architecture for real-time semi-global stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 660–673, Feb. 2022.

[4] K. Zeng, H. Zhang, W. Wang, Y. Wang, and J. Mao, "Deep stereo network with MRF-based cost aggregation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2426–2438, Apr. 2024.

[5] C.-W. Liu, H. Wang, S. Guo, M. J. Bocus, Q. Chen, and R. Fan, "Stereo matching: Fundamentals, state-of-the-art, and existing challenges," in *Autonomous Driving Perception: Fundamentals and Applications*. Singapore: Springer, 2023, pp. 63–100.

[6] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3025–3035, Jun. 2018.

[7] R. Fan et al., "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 1, pp. 225–233, Feb. 2022.

[8] R. Fan, H. Wang, P. Cai, and M. Liu, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 340–356.

[9] H. Liu et al., "A low-cost and scalable framework to build large-scale localization benchmark for augmented reality," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2274–2288, Apr. 2024.

[10] X. Deng, Y. Deng, R. Yang, W. Yang, R. Timofte, and M. Xu, "MASIC: Deep mask stereo image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 6026–6040, Oct. 2023.

[11] S. Li, S. He, S. Jiang, W. Jiang, and L. Zhang, "WHU-stereo: A challenging benchmark for stereo matching of high-resolution satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5603914, doi: 10.1109/TGRS.2023.324520.

[12] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2020.

[13] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Trans. Image Process.*, vol. 30, pp. 8144–8154, 2021.

[14] Y. Lee and H. Kim, "A high-throughput depth estimation processor for accurate semiglobal stereo matching using pipelined inter-pixel aggregation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 411–422, Jan. 2022.

[15] Q. Chen, B. Ge, and J. Quan, "Unambiguous pyramid cost volumes fusion for stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 9223–9236, Oct. 2024.

[16] R. Fan, U. Ozgunalp, Y. Wang, M. Liu, and I. Pitas, "Rethinking road surface 3-D reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5799–5808, Jul. 2022.

[17] H. Zhang, X. Ye, S. Chen, Z. Wang, H. Li, and W. Ouyang, "The farther the better: Balanced stereo matching via depth-based sampling and adaptive feature refinement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4613–4625, Jul. 2022.

[18] P. Li, C. Yao, Y. Jia, and Y. Wu, "Inter-scale similarity guided cost aggregation for stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 1, pp. 134–147, Jan. 2025.

[19] J. Okae, B. Li, J. Du, and Y. Hu, "Robust scale-aware stereo matching network," *IEEE Trans. Artif. Intell.*, vol. 3, no. 2, pp. 244–253, Apr. 2022.

[20] C. Yao, Y. Jia, H. Di, P. Li, and Y. Wu, "A decomposition model for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6087–6096.

[21] L. Bartolomei et al., "Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail," in *Proc. Comput. Vis. Pattern Recognit. Conf. (CVPR)*, Aug. 2025, pp. 1013–1027.

[22] J. Cheng et al., "MonSter: Marry monodepth to stereo unleashes power," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2025, pp. 6273–6282.

[23] S. Joung, S. Kim, K. Park, and K. Sohn, "Unsupervised stereo matching using confidential correspondence consistency," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 2190–2203, May 2020.

[24] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "SegStereo: Exploiting semantic information for disparity estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 636–651.

[25] H. Ren, A. Raj, M. El-Khamy, and J. Lee, "SUW-Learn: Joint supervised, unsupervised, weakly supervised deep learning for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 750–751.

[26] S. M. N. Uddin, S. H. Ahmed, and Y. J. Jung, "Unsupervised deep event stereo for depth estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7489–7504, Nov. 2022.
[27] X. Yang, H. Lai, B. Zou, H. Fu, and Q. Long, "Self-supervised learning of PSMNet via generative adversarial networks," in *Proc. Int. Conf. Intell. Comput. (ICIC)*. Cham, Switzerland: Springer, 2024, pp. 469–479.
[28] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," 2017, *arXiv:1709.00930*.
[29] A. Li and Z. Yuan, "Occlusion aware stereo matching via cooperative unsupervised learning," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2019, pp. 197–213.
[30] A. Li, Z. Yuan, Y. Ling, W. Chi, S. Zhang, and C. Zhang, "Unsupervised occlusion-aware stereo matching with directed disparity smoothing," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7457–7468, Jul. 2022.
[31] H. Wang, R. Fan, and M. Liu, "Co-teaching: An ark to unsupervised stereo matching," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3328–3332.
[32] P. Liu, I. King, M. R. Lyu, and J. Xu, "Flow2Stereo: Effective self-supervised learning of optical flow and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6647–6656.
[33] M. Poggi et al., "On the confidence of stereo matching in a deep-learning era: A quantitative evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5293–5313, Sep. 2022.
[34] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2121–2133, Nov. 2012.
[35] M. Poggi, F. Tosi, and S. Mattoccia, "Quantitative evaluation of confidence measures in a machine learning world," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5238–5247.
[36] M. Poggi and S. Mattoccia, "Learning a general-purpose confidence measure based on O(1) features and a smarter aggregation strategy for semi global matching," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 509–518.
[37] M. Poggi, F. Tosi, and S. Mattoccia, "Learning a confidence measure in the disparity domain from O(1) features," *Comput. Vis. Image Understand.*, vol. 193, pp. 2905–2913, Apr. 2020.
[38] M. Poggi, F. Aleotti, F. Tosi, G. Zaccaroni, and S. Mattoccia, "Self-adapting confidence estimation for stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 715–733.
[39] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 46.1–46.13.
[40] F. Tosi, M. Poggi, A. Benincasa, and S. Mattoccia, "Beyond local reasoning for stereo confidence estimation with deep learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 323–338.
[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Munich, Germany: Springer, Oct. 2015, pp. 234–241.
[42] L. Chen, W. Wang, and P. Mordohai, "Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17235–17244.
[43] F. Tosi, M. Poggi, S. Mattoccia, A. Tonioni, and L. D. Stefano, "Learning confidence measures in the wild," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–13.
[44] M.-G. Park and K.-J. Yoon, "Learning and selecting confidence measures for robust stereo matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1397–1411, Jun. 2019.
[45] C. Mostegel, M. Rumpler, F. Fraundorfer, and H. Bischof, "Using self-contradiction to learn confidence measures in stereo vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4067–4076.
[46] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, p. 20. [Online]. Available: https://openreview.net/forum?id=plKu2GByCNW
[47] L. Yang et al., "Depth anything V2," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024, pp. 21875–21911.
[48] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.
[49] 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.
[50] S. Roy, "Stereo without epipolar lines: A maximum-flow formulation," *Int. J. Comput. Vis.*, vol. 34, no. 2, pp. 147–161, 1999.
[51] C.-W. Liu, Y. Zhang, Q. Chen, I. Pitas, and R. Fan, "These maps are made by propagation: Adapting deep stereo networks to road scenarios with decisive disparity diffusion," *IEEE Trans. Image Process.*, vol. 34, pp. 1516–1528, 2025.
[52] H. Wang, R. Fan, P. Cai, and M. Liu, "PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4353–4360, Jul. 2021.
[53] Z. Wang, 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, no. 4, pp. 600–612, Apr. 2004.
[54] N. Mayer et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4040–4048.
[55] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," 2020, *arXiv:2001.10773*.
[56] 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.
[57] 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.
[58] D. Scharstein et al., "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.
[59] T. Schops et al., "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3260–3269.
[60] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21919–21928.
[61] J. Li et al., "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16263–16272.
[62] P.-A. Brousseau and S. Roy, "A permutation model for the self-supervised stereo matching problem," in *Proc. 19th Conf. Robots Vis. (CRV)*, May 2022, pp. 122–131.
[63] L. Wang et al., "Parallax attention for unsupervised stereo correspondence learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2108–2125, Apr. 2022.
[64] K. W. Tong, P. Z. H. Sun, E. Q. Wu, C. Wu, and Z. Jiang, "Adaptive cost volume representation for unsupervised high-resolution stereo matching," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 912–922, Jan. 2023.
[65] X. Cheng et al., "A novel cell structure-based disparity estimation for unsupervised stereo matching," *IET Image Process.*, vol. 16, no. 6, pp. 1678–1693, May 2022.

Chuang-Wei Liu (Student Member, IEEE) received the B.E. degree in automation from Tongji University in 2020. He is currently pursuing the Ph.D. degree with the Machine Intelligence and Autonomous Systems (MIAS) Group, Robotics and Artificial Intelligence Laboratory (RAIL), Tongji University, supervised by Prof. Rui Fan. His research interests include computer stereo vision, especially for unsupervised approaches, and long-term learning.

Mingjian Sun (Member, IEEE) was born in Weifang, Shandong, China, in 1980. He received the B.S. degree in automation, the M.S. degree in fluid machinery and engineering, and the Ph.D. degree in control science and engineering from Harbin Institute of Technology, China, in 2003, 2005, and 2011, respectively. He is currently the Ph.D. Supervisor with Harbin Institute of Technology. His research interests include medical ultrasound and photoacoustic imaging, AI-based medical image analysis, and autonomous control and applications of unmanned systems.

Cairong Zhao (Member, IEEE) received the B.S. degree from Jilin University in 2003, the M.S. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2006, and the Ph.D. degree from Nanjing University of Science and Technology in 2011. He is a Professor with the College of Electronic and Information Engineering, Tongji University. He has published over 50 papers in leading international conferences and journals, including CVPR, ICCV, ICML, ICLR, AAAI, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. His research focuses on visual and intelligent learning, with particular interests in computer vision, pattern recognition, and visual surveillance. He serves as the Chair for the Computer Vision Special Committee of the Shanghai Computer Society. He is an active reviewer of more than ten top-tier AI-related journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, NeurIPS, ICML, and AAAI.

Hanli Wang (Senior Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2007. From 2007 to 2008, he was a Research Fellow with the Department of Computer Science, City University of Hong Kong and concurrently a Visiting Scholar with Stanford University, Palo Alto, CA, USA. From 2008 to 2009, he was a Research Engineer with Precoad Inc., Menlo Park, CA. He was an Alexander von Humboldt Research Fellow with the University of Hagen, Germany, from 2009 to 2010. Since 2010, he has been a Professor with the College of Electronic and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, multimedia signal processing, and machine learning.

Alexander Dvorkovich (Member, IEEE) received the M.Sc. degree in applied physics and mathematics from Moscow Institute of Physics and Technology (MIPT) in 1990, the Ph.D. degree in chemical physics from MIPT in 1993, supervised by Prof. Arkadiy D. Margolin, and the Doctor of Engineering Science degree in radio engineering from Bauman Moscow State Technical University (BMSTU) in 2007. He has been working in the field of telecommunications since 1998, initially as a Senior Researcher with the Radio Research and Development Institute, Moscow, Russia. In 2007, he became the Head of the Scientific Research Laboratory, General Radio Frequency Center, Moscow. Since 2015, he has been the Head of the Multimedia Systems and Technology Laboratory, MIPT. He is currently the Head of the Multimedia Technology and Telecom Department, MIPT; and a Professor with the Radiotechnical Devices and Antenna Systems Department, Moscow Power Engineering Institute (National Research University). In 2016, he was elected as a Corresponding Member of Russian Academy of Sciences (RAS).

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 with The Hong Kong University of Science and Technology from 2018 to 2020 and a Post-Doctoral Scholar-Employee with the University of California 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 with the College of Electronics and Information Engineering and 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 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 received the Shanghai Science and Technology 35 Under 35 honor in 2024 as its youngest recipient. He organized several impactful workshops and special sessions in conjunction with WACV 2021, ICIP 2021/2022/2023, ICCV 2021/2025, and ECCV 2022. He served as an Associate Editor for ICRA'23/25 and IROS'23/24, the Area Chair for ICIP'24, and a Senior Program Committee Member for AAAI'23/24/25/26.