These Maps Are Made by Propagation: Adapting Deep Stereo Networks to Road Scenarios With Decisive Disparity Diffusion

Chuang-Wei Liu, *Student Member, IEEE*, Yikang Zhang, Qijun Chen, *Senior Member, IEEE*, Ioannis Pitas, *Life Fellow, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Stereo matching has emerged as a cost-effective solution for road surface 3D reconstruction, garnering significant attention towards improving both computational efficiency and accuracy. This article introduces decisive disparity diffusion (D3Stereo), marking the first exploration of dense deep feature matching that adapts pre-trained deep convolutional neural networks (DCNNs) to previously unseen road scenarios. A pyramid of cost volumes is initially created using various levels of learned representations. Subsequently, a novel recursive bilateral filtering algorithm is employed to aggregate these costs. A key innovation of D3Stereo lies in its alternating decisive disparity diffusion strategy, wherein intra-scale diffusion is employed to complete sparse disparity images, while inter-scale inheritance provides valuable prior information for higher resolutions. Extensive experiments conducted on our created UDTIRI-Stereo and Stereo-Road datasets underscore the effectiveness of D3Stereo strategy in adapting pre-trained DCNNs and its superior performance compared to all other explicit programming-based algorithms designed specifically for road surface 3D reconstruction. Additional experiments conducted on the Middlebury dataset with backbone DCNNs pre-trained on the ImageNet database further validate the versatility of D3Stereo strategy in tackling general stereo matching problems. Our source code and supplementary material are publicly available at https://mias.group/D3-Stereo.**

***Index Terms*—Stereo matching, 3D reconstruction, convolutional neural networks, recursive bilateral filtering.**

I. Introduction

ENSURING safe and comfortable driving requires the timely assessment of road conditions and the prompt repair of road defects [1]. With an increasing emphasis on maintaining high-quality road conditions [2], the demand for automated 3D road data acquisition systems has grown more intense than ever [3], [4]. The study presented in [5] employs a laser scanner to collect high-precision 3D road data. Nevertheless, the high equipment costs and the long-term maintenance expenses have limited the widespread adoption of such laser scanner-based systems [6]. Therefore, stereo vision, a process similar to human binocular vision that provides depth perception using dual cameras, has emerged as a practical and cost-effective alternative for accurate 3D road data acquisition [7], [8]. Existing stereo matching approaches are either explicit programming-based or data-driven. The former ones rely on hand-crafted feature extraction and estimate disparities through local block matching or global energy minimization [9]. Nonetheless, hand-crafted feature extraction faces challenges in handling varying lighting conditions and noise. With recent advances in deep learning, researchers have resorted to deep convolutional neural networks (DCNNs) for stereo matching [10], [11]. These data-driven approaches can learn abstract features directly from input stereo images, making them increasingly favored in this research domain. Unfortunately, the limited availability of well-annotated road disparity data restrains the transfer learning of these DCNNs [12]. Therefore, explicitly programming-based stereo matching approaches [7], [13], [14] remain the mainstream in the field of road surface 3D reconstruction.

Building upon the local coherence constraint [15], seed-and-grow stereo matching algorithms [7], [16], [17] have been widely utilized for quasi-dense disparity estimation. Given that road disparities change gradually across continuous regions, our previously published road surface 3D reconstruction algorithm search range propagation (SRP) [7] initializes disparity

Received 8 August 2024; revised 6 November 2024; accepted 25 January 2025. Date of publication 19 February 2025; date of current version 6 March 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 Science and Technology Commission of Shanghai Municipal under Grant 22511104500; in part by the Fundamental Research Funds for the Central Universities, NIO University Program (NIO UP); in part by the Xiaomi Young Talents Program; and in part by the European Commission–European Union through Horizon Europe (Horizon Research and Innovation Actions) under Grant 101093003 (TEMA) HORIZON-CL4-2022-DATA-01-01. The associate editor coordinating the review of this article and approving it for publication was Prof. Wanqing Li. *(Corresponding author: Rui Fan.)*

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

Yikang Zhang is with Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mail: yikangzhang@tongji.edu.cn).

Ioannis Pitas is with the Department of Informatics, University of Thessaloniki, 541 24 Thessaloniki, Greece (e-mail: pitas@csd.auth.gr).

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, the State Key Laboratory of Intelligent Autonomous Systems, and the 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.2025.3540283, provided by the authors.

Digital Object Identifier 10.1109/TIP.2025.3540283

seeds using a winner-take-all (WTA) strategy at the bottom row of the image and estimates disparities iteratively with the search range propagated from three neighboring seeds. Another significant contribution of [7] lies in the perspective transformation (PT), designed to convert the target view of the road image into a reference view. This transformation helps decrease computations by reducing the disparity search range and improving stereo matching accuracy by increasing the similarity of the compared blocks. While the combination of SRP and PT yields a remarkable 3D geometry reconstruction accuracy of approximately 3 mm, it is noteworthy that the disparity estimation accuracy remains constrained by the reliability of the initial seeds generated using the simple WTA strategy. The unidirectional disparity propagation process further leads to disparity estimation errors on discontinuities, such as road defects. Additionally, both seed-and-grow stereo matching and perspective transformation require a set of sparse yet reliable initial correspondences, and the density and reliability of these correspondences directly affect the efficiency and accuracy of the seed-growing process.

Drawing inspiration from recent advances in plug-and-play sparse correspondence matching [18], [19] approaches, we propose a feasible solution to address these limitations. For example, the deep feature matching (DFM) method [18] utilizes a backbone DCNN pre-trained on the ImageNet database [20] to extract feature pyramids for both views, and subsequently refines the coarse correspondences initialized at the deepest feature layer to former layers following a linear hierarchical manner. These methods have demonstrated the effectiveness of using deep features provided by pre-trained backbones to solve the correspondence matching task. Therefore, our primary motivation is to develop a dense deep feature matching strategy by improving the seed-and-grow stereo matching with the hierarchical refinement strategy in DFM. Leveraging accurate sparse correspondences as disparity seeds, such a dense deep feature matching strategy exhibits compatibility with perspective transformation, thus leading to improvements in both stereo matching accuracy and efficiency compared with the combination of SRP and PT. However, directly incorporating a hierarchical refinement strategy into seed-and-grow stereo matching still has the following limitations:

- The dense matching process in the stereo matching task requires additional matching noise elimination operations in challenging areas with weak/repetitive textures.
- Additional efforts in eliminating inaccurate sparse correspondences are required to mitigate error accumulation and propagation in the seed-growing process.
- The linear hierarchical refinement strategy in DFM is designed to enhance the spatial details of the coarse initial correspondences, while having limited effectiveness in enhancing their density.

To address these limitations, we propose a plug-and-play stereo matching strategy for road surface 3D reconstruction, referred to as **D**ecisive **D**isparity **D**iffusion **Stereo** (**D3Stereo**), serving as the first exploration of dense deep feature matching. D3Stereo is compatible with any hand-crafted feature extraction approaches, stereo matching networks pre-trained on other public datasets, and even backbone DCNNs pre-trained for image classification. We first propose the recursive bilateral filtering (RBF) algorithm, a more efficient alternative to traditional bilateral filtering (BF) [13] for matching cost aggregation. By recursively applying a small filtering kernel, our BRF achieves a significantly expanded receptive field while maintaining the same computational cost as BF, thereby gathering more context information for cost aggregation. The proposed method leverages the powerful semantic feature extraction ability of a pre-trained DCNN backbone in a hierarchical manner. It consists of two algorithms that diffuse decisive disparities at both intra and inter scales, respectively. With a cost volume pyramid built with different layers of feature maps, we first find coarse decisive disparities at the deepest layer. Then, the coarse decisive disparities are adversarially diffused to their neighboring pixels in the same layer to yield a dense disparity map, within which reliable decisive disparities are inherited into the former layer by checking the matching cost local minima consistency between consecutive layers. Our adversarial disparity diffusion process and novel disparity inheritance strategy help eliminate the inaccurate correspondences initialized at the last layer. Afterwards, the derived refinement results activate the decisive disparity intra and inter scale diffusion in the former layer. This process is repeated until a dense disparity map is obtained at the finest resolution layer. In general, the combined usage of diffusing decisive disparities at both intra and inter scales fully exploits the semantic information at different scales of feature maps, thus obtaining improved disparity seeds in terms of both accuracy and distribution uniformity compared with the hierarchical refinement strategy in DFM [18].

Additionally, we create a synthetic road dataset called the UDTIRI-Stereo dataset using the CARLA simulator [21] for disparity estimation evaluation. Although collecting datasets using simulators has emerged as a prevalent alternative for real-world datasets [22], [23], these simulators model the road surface as a ground plane, thereby significantly reducing the complexity of the stereo matching task. In order to narrow the domain gap between the idealized road surface in CARLA and the real-world road surface, we originally augment the road surface mesh model in CARLA with 1) 2D Perlin noise and 2) digital twins of pothole models. By applying linear interpolation between initial random noises, 2D Perlin noise is utilized to generate the natural undulations of the real-world road surface. Moreover, digital twins of pothole models yielded in real-world [6] are randomly transplanted onto the road surface, thus further introducing disparity discontinuities into the UDTIRI-Stereo dataset.

II. Related Works

A. Stereo Matching for Road Surface 3D Reconstruction

Several stereo matching approaches [7], [13], [14], developed specifically for road surface 3D reconstruction, have been proposed since 2014 [24]. The first reported effort in this area of research was our proposed iterative stereo matching algorithm SRP [7]. Despite the remarkable 3D

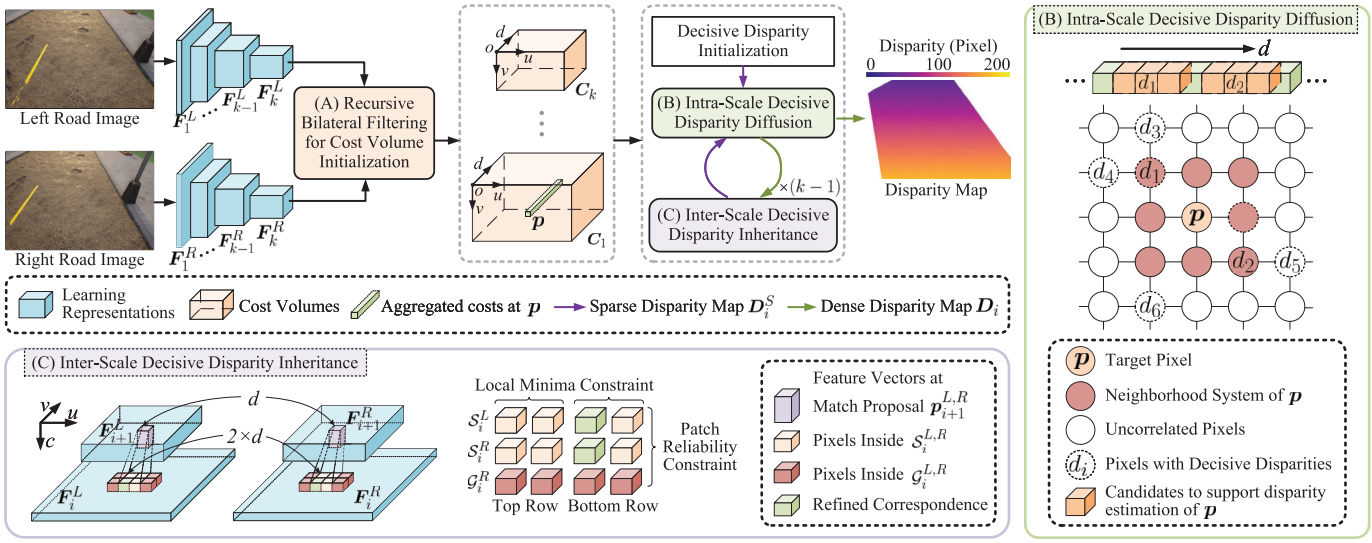


Fig. 1. An illustration of our proposed D3Stereo strategy. Cost volume pyramid is first initialized with RBF. Afterwards, coarse decisive disparities initialized at the deepest layer are hierarchically propagated into former layers with alternating decisive disparity intra-scale diffusion and inter-scale inheritance algorithms.

geometry reconstruction accuracy yielded by SRP, its row-by-row disparity propagation process is challenging to implement in parallel on GPUs. To address this issue, [13] proposed a GPU-friendly algorithm for road disparity estimation based on fast bilateral stereo (FBS) that can be embedded in a drone for real-time road surface 3D reconstruction. Nevertheless, the bilateral filtering process in FBS is computationally intensive, especially when a large filter kernel is employed, thus further increasing the memory burden on the embedded computers. As a result, semi-global matching (SGM) was used in conjunction with PT in [14] for road disparity estimation. Experimental results suggest that SGM outperforms both SRP and FBS when PT is incorporated. In general, D3Stereo continues the search range propagation strategy in SRP, while the seed-growing process is executed within a single instruction multiple data architecture for better leveraging parallel computing resources. Additionally, a recursive bilateral filter is proposed for more efficient cost aggregation compared with FBS.

On the other hand, recent domain generalization-aimed stereo matching networks [25], [26], [27], [28], [29] achieve remarkable generalizability across various scenarios. Common strategies to maintain the performance of stereo matching networks under scene changes include narrowing the cross-domain feature inconsistency [26], [28], [29] and enhancing the ability of DCNNs to learn more image structure information [25]. However, stereo matching in road scenes additionally emphasizes the ability of networks to handle fine-grained disparity variations compared to general domain adaptation tasks, and the performance of these domain generalization-aimed stereo matching networks in road scenes remains unverified.

B. Sparse Correspondence Matching

Conventional hand-crafted sparse correspondence matching approaches [30], [31], [32] first extract keypoints using explicitly designed local feature detectors and descriptors. Correspondence pairs are then determined using the nearest neighbor search algorithm. Although recent data-driven approaches [33], [34], [35], [36], [37], [38], [39], [40] have demonstrated significant improvements over hand-crafted methods, these supervised methods usually demand a large amount of well-annotated data for model training, resulting in unsatisfactory performance when applied to new domains [19]. Moreover, adopting an independent sparse correspondence matching algorithm, regardless of whether it relies on explicit programming or DCNNs, for seed initialization leads to increased consumption of memory and computational resources. Recent plug-and-play approaches, such as DFM [18] and epipolar-constrained cascade correspondence matching [19], utilize backbone DCNNs pre-trained on the ImageNet database [20] for sparse correspondence matching based on a hierarchical refinement strategy, obviating the necessity for model fine-tuning. Consequently, a preferred solution would be to opt for plug-and-play algorithms that leverage the backbone DCNN incorporated into D3Stereo for seed initialization.

III. Methodology

Based on Markov random field theory [41], stereo matching can be formulated as an energy minimization problem [7]:

$$E = \sum_{\boldsymbol{p}\in\boldsymbol{D}} D(\boldsymbol{p}, d) + \sum_{\boldsymbol{q}\in\mathcal{N}_{\boldsymbol{p}}} V(\boldsymbol{p}, \boldsymbol{q}), \tag{1}$$

where $\boldsymbol{p} = [u, v]^\top$ denotes a 2D pixel within the disparity map $\boldsymbol{D}$, function $D(\cdot)$ measures the stereo matching confidence at a given disparity d, function $V(\cdot)$ quantifies the compatibility between $\boldsymbol{p}$ and its neighborhood system $\mathcal{N}_{\boldsymbol{p}}$ comprising a collection of 2D pixels $\boldsymbol{q}$ adjacent to $\boldsymbol{p}$. As demonstrated in [42], confident disparities tend to have consistent matching costs regardless of scales. Therefore, drawing inspiration from the scale-invariant feature detection introduced in [30], we extend $\mathcal{N}_{\boldsymbol{p}}$ to incorporate neighborhood systems of $\boldsymbol{p}$ across various scales, enabling pyramid stereo matching in this study. Following [18], we perform stereo matching via a hierarchical refinement strategy, as illustrated in Fig. 1. The process of $D(\cdot)$ is accomplished using either conventional

explicit programming-based algorithms or pre-trained DCNN backbones, as detailed in Sect. III-A, while the process of $V(\cdot)$ is achieved through an intra-scale decisive disparity diffusion algorithm, and an inter-scale decisive disparity inheritance algorithm, as detailed in Sects. III-B and III-C, respectively. Our adopted sparse decisive disparity initialization approach obviates the necessity for additional keypoint detection and matching algorithms that are commonly used in conventional seed-and-grow stereo matching methods. Additionally, the combined use of decisive disparity intra-scale diffusion and inter-scale inheritance not only ensures the quantity and distribution uniformity of the estimated disparities but also significantly enhances stereo matching efficiency.

A. Recursive Bilateral Filtering for Cost Volume Initialization

In our earlier research [7], we utilized the normalized cross-correlation (NCC) for stereo matching cost computation. Nevertheless, recent data-driven algorithms, generally developed based on DCNNs, have demonstrated superior performance compared to such explicit programming-based methods. This is attributed to their capabilities of learning more informative hierarchical representations, thereby offering a more effective solution for stereo matching challenges in complex scenarios. Hence, in this paper, we develop D3Stereo for both explicit programming-based and data-driven methods. In this subsection, we detail only the cost volume initialization using pre-trained DCNN backbones (either the backbones pre-trained on the ImageNet [20] database for natural image classification or those embedded in pre-trained deep stereo networks). Nevertheless, this procedure can also be accomplished through explicit block matching.

Given a pair of stereo road images $\boldsymbol{I}^L$ and $\boldsymbol{I}^R$, we first extract a collection of deep feature maps $\mathcal{F}^L = \{\boldsymbol{F}_1^L, \ldots, \boldsymbol{F}_k^L\}$ and $\mathcal{F}^R = \{\boldsymbol{F}_1^R, \ldots, \boldsymbol{F}_k^R\}$ at k different resolutions using a pre-trained DCNN backbone. The feature maps $\boldsymbol{F}_i^{L,R}$ generally possess half the resolution of their shallower adjacent ones $\boldsymbol{F}_{i-1}^{L,R}$. A cost volume pyramid $\mathcal{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_k\}$ can be subsequently obtained by computing the cosine similarity between each pair of left and right deep feature maps, respectively. The matching costs in $\mathcal{C}$ undergo normalization, with a lower matching cost indicating a better match.

As a standard step in stereo matching algorithms, we conduct cost aggregation on the cost volumes to improve the piece-wise disparity coherency across the support region of each pixel [43]. It has been mathematically proven that the function $V(\cdot)$ in (1) can be formulated through an adaptive cost aggregation process using a bilateral filter [44]. A larger kernel size (commonly regarded as the "receptive field") often brings improved disparity estimation results. However, increasing the bilateral filtering kernel size can substantially lead to a notable increase in computational demands, thereby imposing significant memory pressure on parallel computing resources.

A prevalent trend in network architecture design lies in replacing a large convolution kernel with stacked small ones [45]. While possessing the same receptive field size, stacked small kernels exhibit lower computational complexity and greater network depth compared to a single large kernel. Motivated by this network architecture design, we introduce a recursive bilateral filtering algorithm for memory-efficient cost aggregation as follows:

$$\boldsymbol{C}_i^{(t)}(\boldsymbol{p}, d) = \frac{\sum\limits_{\boldsymbol{q} \in \mathcal{N}_{\boldsymbol{p},i} \cup \{\boldsymbol{p}\}} \boldsymbol{K}_i(\boldsymbol{q}) \boldsymbol{C}_i^{(t-1)}(\boldsymbol{p}, d)}{\sum\limits_{\boldsymbol{q} \in \mathcal{N}_{\boldsymbol{p},i} \cup \{\boldsymbol{p}\}} \boldsymbol{K}_i(\boldsymbol{q})}, \tag{2}$$

where $\boldsymbol{p}$ is a 2D pixel in $\boldsymbol{C}_i$, d represents a disparity candidate at $\boldsymbol{p}$, $\boldsymbol{C}_i^{(t)}$ represents the i-th cost volume after the t-th RBF iteration with $\boldsymbol{C}_i^{(0)} = \boldsymbol{C}_i$. In the RBF kernel:

$$\boldsymbol{K}_i(\boldsymbol{q}) = \exp\left\{ -\frac{\|\boldsymbol{p} - \boldsymbol{q}\|_2^2}{{\sigma_1}^2} - \frac{\left(\boldsymbol{I}_i^L(\boldsymbol{p}) - \boldsymbol{I}_i^L(\boldsymbol{q})\right)^2}{{\sigma_2}^2} \right\}, \tag{3}$$

$\mathcal{N}_{\boldsymbol{p},i}$ denotes a neighborhood system of $\boldsymbol{p}$ (the RBF kernel radius $\kappa_a = 1$ corresponds to an eight-connected neighborhood system) at the i-th scale, $\boldsymbol{I}_i^L$ denotes a downsampled $\boldsymbol{I}^L$ with the same resolution as $\boldsymbol{F}_i^L$, σ_1 and σ_2 denote weighting parameters related to spatial distance and color similarity, respectively. As discussed in [46], executing $t_{\max}$ iterations of bilateral filtering with a 3×3 kernel is functionally equivalent, in terms of receptive field size, to performing the filtering process once, but with a $(2t_{\max} + 1) \times (2t_{\max} + 1)$ kernel. The computational consumption ratio of traditional bilateral filtering versus our proposed RBF is $\frac{1}{9}\left(4t_{\max} + \frac{1}{t_{\max}} + 4\right)$, which shows a monotonic increase when $t_{\max} > \frac{1}{2}$. Moreover, it has been mathematically proven in [47] that for a stack of convolutional layers, the weights of each pixel within its theoretical receptive field adhere to a Gaussian distribution. This concept naturally translates to the recursive structure of our proposed RBF. Therefore, with the same computational complexity, our proposed RBF can produce a larger receptive field adhering to a Gaussian distribution, thereby gathering more context information for cost aggregation. In addition, in practical implementations, the GPU memory needs are reduced by a factor of $\frac{1}{9}(4t_{\max}^2 + 4t_{\max} + 1)$ when using our proposed RBF, significantly optimizing the memory resource usage.

B. Intra-Scale Decisive Disparity Diffusion

As illustrated in Fig. 1, D3Stereo strategy is initialized with a collection of coarse decisive disparities, determined by measuring the peak ratio naive (PKRN) [48] scores and checking the left-right disparity consistency (LRDC) [7] at the k-th layer, as employed in DFM [18]. This process results in $\boldsymbol{D}_k^S$, a sparse disparity map with the lowest resolution. The linear hierarchical refinement structure employed in DFM has been proven to dramatically improve the stereo matching efficiency. However, the resulting disparity map is often quasi-dense, comprising a set of disparity clusters originating from a single initial decisive disparity. To address this issue, we introduce an intra-scale decisive disparity diffusion process positioned between two successive inter-scale refinement processes, denoted by an alternating hierarchical refinement structure (ARS). This novel contribution helps densify the sparse depth information initialized by the PKRN dense search process, thereby improving both the quantity and distribution uniformity of decisive disparities while retaining the efficiency

Algorithm 1 Intra-Scale Decisive Disparity Diffusion

Input: Cost volume $\boldsymbol{C}_i$ and decisive disparity map $\boldsymbol{D}_i^S$
Output: Disparity map $\boldsymbol{D}_i$

Initialize an empty set $\mathcal{P}$ to store the candidates for intra-scale decisive disparity diffusion;
$\boldsymbol{D}_i \leftarrow \boldsymbol{D}_i^S$;
for $\boldsymbol{p}$ *whose neighboring pixel* $\boldsymbol{q}$ *is determined to have a decisive disparity* **do**
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{\boldsymbol{p}\}$;
repeat
  for $\boldsymbol{p} \in \mathcal{P}$ **do**
    $\mathcal{P} \leftarrow \mathcal{P} - \{\boldsymbol{p}\}$;
    Calculate its state $s_{\boldsymbol{p},i}^{(t)}$ using (4);
    if *its state satisfies hypotheses (2) and (3), or the adversarial mechanism condition in (5)* **then**
      $\boldsymbol{D}_i(\boldsymbol{p}) \leftarrow s_{\boldsymbol{p},i}^{(t)}$ and $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{N}_{\boldsymbol{p}}$;
until *no more pixels experience state changes*;

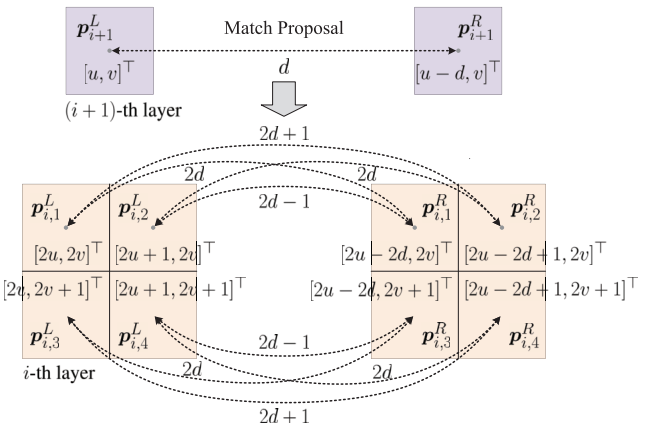


Fig. 2. An illustration of the inter-scale decisive disparity inheritance process. Each match proposal is mapped into a pair of patches with a size of 2×2 pixels at the former layer, comprising eight fine-grained match candidates.

gains achieved by the linear hierarchical refinement strategy in DFM. Our proposed intra-scale decisive disparity diffusion algorithm is developed based on the following hypotheses:

(1) disparities change gradually across continuous regions;
(2) the matching cost of a desired disparity is a local minima;
(3) disparities between stereo images are consistent.

We define a disparity state variable $s_{\boldsymbol{p},i}^{(t)}$ for $\boldsymbol{p}$ in the t-th iteration of decisive disparity diffusion when estimating the i-th disparity map $\boldsymbol{D}_i$ ($i \in [1,k] \cap \mathbb{Z}$) as follows:

$$s_{\boldsymbol{p},i}^{(t)} = \arg\min_{s} \left\{ \boldsymbol{C}_i(\boldsymbol{p}, s) | s \in \Phi\Big(\bigcup_{\boldsymbol{q} \in \mathcal{N}_{\boldsymbol{p},i}} \{ \boldsymbol{D}_i^{(t-1)}(\boldsymbol{q}) + r \} \Big) \right\}, \quad (4)$$

where $r \in [-\tau, \tau] \cap \mathbb{Z}$ denotes the disparity search tolerance, in which $\tau \in \mathbb{Z}$ represents the disparity search bound, the neighborhood system $\mathcal{N}_{\boldsymbol{p},i}$ for disparity diffusion has a radius κ_d, $\Phi(\cdot)$ represents an operation to uniquify a given set, and $\boldsymbol{D}_i^{(t-1)}$ denotes the disparity map obtained after the $(t-1)$-th iteration with $\boldsymbol{D}_k^{(0)} = \boldsymbol{D}_k^S$, and $\boldsymbol{D}_i^{(0)} = \boldsymbol{D}_i^S$ $(i < k)$ and are derived from the inter-scale refinement process at the i-th layer. The disparity state variable $s_{\boldsymbol{p},i}^{(t)}$ that fulfills the hypotheses (2) and (3) mentioned above is considered to be decisive. Afterwards, the newly generated decisive disparities are identically utilized to propagate disparity ranges to their neighborhood systems in the next iteration, corresponding to an ongoing process of depth information completion. To improve computational efficiency, we confine the intra-scale decisive disparity diffusion process only to the pixels whose neighborhood system has experienced state changes in the previous iteration. Moreover, unlike the conventional unidirectional seed-growing process, we also incorporate an adversarial mechanism into our intra-scale decisive disparity diffusion process to update disparities that may have been determined incorrectly in the previous iterations. Specifically, if a pixel satisfies the following condition:

$$\begin{aligned} &\min\left\{ \boldsymbol{C}_i(\boldsymbol{p}, s_{\boldsymbol{p},i}^{(j)}) | j \in [0, t-2] \cap \mathbb{Z} \right\} < \\ &\min\left\{ \boldsymbol{C}_i(\boldsymbol{p}, s) | s \in \Phi\Big(\bigcup_{\boldsymbol{q} \in \mathcal{N}_{\boldsymbol{p},i}} \{ \boldsymbol{D}_i^{(t-1)}(\boldsymbol{q}) + r \} \Big) \right\}, \end{aligned} \quad (5)$$

its disparity will be updated accordingly. This mechanism helps reduce the occurrence of incorrect disparities in the subsequent inter-scale decisive disparity inheritance process. Intra-scale decisive disparity diffusion terminates when no more pixels experience state changes (namely $s_{\boldsymbol{p},i}^{(t)} = s_{\boldsymbol{p},i}^{(t-1)}$), resulting in a dense disparity map $\boldsymbol{D}_i$. Additional details on our proposed intra-scale decisive disparity diffusion strategy are provided in Algorithm 1. Moreover, perspective transformation is performed using the k-th dense disparity map $\boldsymbol{D}_k$, thereby narrowing the disparity search range and enhancing the block matching similarity.

C. Inter-Scale Decisive Disparity Inheritance

The robustness of stereo matching using feature maps from deeper layers has been demonstrated in effectively addressing ambiguities, *e.g.*, repetitive patterns and texture-less regions [49]. This can be attributed to the richer semantic information within these feature maps generated through larger receptive fields [19]. In contrast, feature maps from shallower layers focus on capturing local texture information and fine-grained details, and thus are more sensitive to pixel intensity changes. They help in identifying small details, *e.g.*, edges and surface textures [14]. Therefore, in this study, after estimating decisive disparities at the lowest resolution, we perform a series of inter-scale decisive disparity inheritance and intra-scale decisive disparity diffusion operations alternately until we obtain a dense disparity map $\boldsymbol{D}_1$ at the highest resolution. This strategy incrementally introduces fine-grained details into the disparity map.

In Patch2Pix [50], each coarse match proposal $\mathcal{M}_k^{L,R} = \{\boldsymbol{p}_k^L, \boldsymbol{p}_k^R\}$ (determined by a decisive disparity in $\boldsymbol{D}_k^S$) is expanded into a pair of patches with a resolution of $2^{k-1} \times 2^{k-1}$ pixels at the 1-st layer. All features from the layers $\{1, \ldots, k\}$ that correspond to $\boldsymbol{p}_1^{L,R}$ within the patches are concatenated into a single feature vector, which is then fed to two regressors to determine the correspondence matching confidence. However, the simple concatenation operation may not fully exploit the fine-grained details present in the shallower layers, as features at deeper layers often have higher dimensions. To overcome this limitation, DFM [18] employs a hierarchical

Algorithm 2 Inter-Scale Decisive Disparity Inheritance

Input: Cost volume $\boldsymbol{C}_i$ and disparity map $\boldsymbol{D}_{i+1}$
Output: Disparity map $\boldsymbol{D}_i^S$
for $\boldsymbol{p}_{i+1}^{L,R}$ *corresponding to a decisive disparity in* $\boldsymbol{D}_{i+1}$ **do**
  Initialize $\mathcal{S}_i^{L,R}$ and $\mathcal{G}_i^{L,R}$;
  if $\mathcal{S}_i^{L,R}$ *and* $\mathcal{G}_i^{L,R}$ *satisfy the patch reliability constraint stated in* (6) **then**
    for *co-row pixels* $\boldsymbol{p}_i^L \in \mathcal{S}_i^L$ *and* $\boldsymbol{p}_i^R \in \mathcal{S}_i^R$ **do**
      if $\boldsymbol{p}_i^L$ *and* $\boldsymbol{p}_i^R$ *satisfy the local minima constraint stated in* (8) **then**
        $\boldsymbol{D}_i^S(\boldsymbol{p}_i^L) \leftarrow \|\boldsymbol{p}_i^L - \boldsymbol{p}_i^R\|_1$

refinement strategy to incrementally introduce fine-grained details into sparse correspondence matching. Specifically, in DFM, a pair of correspondences $\boldsymbol{p}_{i+1}^L$ and $\boldsymbol{p}_{i+1}^R$ matched at the $(i+1)$-th layer ($i < k$) are mapped into a pair of patches $\mathcal{S}_i^{L,R} = \{\boldsymbol{p}_{i,1}^{L,R}, \ldots, \boldsymbol{p}_{i,4}^{L,R}\}$ with a size of 2×2 pixels at the i-th layer. Matches between $\mathcal{S}_i^{L,R}$ are subsequently determined via PKRN and LRDC, as introduced in Sect. III-B. However, the PKRN-based hierarchical refinement constraint in DFM has two significant drawbacks: (1) determining satisfactory matches with PKRN and LRDC is not always effective and robust due to the limited matching candidates, resulting in error accumulation and propagation from earlier stages to later stages; (2) PKRN requires a manually set threshold for each layer, which decreases its applicability and requires a manual algorithm tuning step.

This study focuses entirely on stereo matching, which is a 1D search problem. Therefore, the hierarchical refinement process aims at determining accurate matches between pixels in two co-row sets: $\mathcal{S}_{i,t}^{L,R} = \{\boldsymbol{p}_{i,1}^{L,R}, \boldsymbol{p}_{i,2}^{L,R}\}$ and $\mathcal{S}_{i,b}^{L,R} = \{\boldsymbol{p}_{i,3}^{L,R}, \boldsymbol{p}_{i,4}^{L,R}\}$, as illustrated in Fig. 2. The subscripts t and b denote the sets on the top and bottom rows, respectively. Following the intra-scale decisive disparity diffusion criteria stated in Sect. III-B, we first analyze the reliability of the given patch pair based on the following hypotheses:

(1) disparities and matching costs within the patch are similar in continuous regions;
(2) the average of matching costs within the patch is lower than the minimum matching costs on the two sides of the patch, which is inherited from the match proposals that satisfy the local minima constraint.

We denote two sets that store the pixels on the left and right sides of $\mathcal{S}_{i,t}^{L,R}$ and $\mathcal{S}_{i,b}^{L,R}$ as $\mathcal{G}_{i,t}^{L,R} = \{\boldsymbol{p}_{i,1}^{L,R} - [1,0]^\top, \boldsymbol{p}_{i,2}^{L,R} + [1,0]^\top\}$ and $\mathcal{G}_{i,b}^{L,R} = \{\boldsymbol{p}_{i,3}^{L,R} - [1,0]^\top, \boldsymbol{p}_{i,4}^{L,R} + [1,0]^\top\}$, respectively, as visualized in Fig. 1. $\mathcal{S}_i^{L,R}$ is considered reliable when it satisfies the following patch reliability constraint (PRC):

$$\begin{aligned}\text{mean}\{\Theta(\mathcal{S}_{i,t}^L, \mathcal{S}_{i,t}^R), \Theta(\mathcal{S}_{i,b}^L, \mathcal{S}_{i,b}^R)\} < \\ \min\{\Theta(\mathcal{S}_{i,t}^L, \mathcal{G}_{i,t}^R), \Theta(\mathcal{S}_{i,b}^L, \mathcal{G}_{i,b}^R)\},\end{aligned} \tag{6}$$

where

$$\Theta(\mathcal{V}_i^L, \mathcal{V}_i^R) = \bigcup_{\boldsymbol{p}_i^L \in \mathcal{V}_i^L} \min\{\boldsymbol{C}_i(\boldsymbol{p}_i^L, \|\boldsymbol{p}_i^L - \boldsymbol{p}_i^R\|_1) | \boldsymbol{p}_i^R \in \mathcal{V}_i^R\}, \tag{7}$$

in which $\mathcal{V}_i^{L,R}$ denotes two input pixel sets. Otherwise, the given match proposal is considered unreliable and all pixels within $\mathcal{S}_i^{L,R}$ are abandoned. We then identify decisive disparities within the reserved $\mathcal{S}_i^{L,R}$, if it satisfies the following local minima constraint:

$$\begin{aligned}\boldsymbol{C}_i(\boldsymbol{p}_i^L, d) < \min\{\{\boldsymbol{C}_i(\boldsymbol{p}_i^L, d+s) | s \in \{-1, 1\}\} \cup \\ \min\{\Theta(\mathcal{S}_{i,t}^L, \mathcal{G}_{i,t}^R), \Theta(\mathcal{S}_{i,b}^L, \mathcal{G}_{i,b}^R)\}\}.\end{aligned} \tag{8}$$

The combined use of patch-based local minima constraint and search range propagation [7] in (8) ensures a more critical determination for inter-scale decisive disparity inheritance, resulting in improved accuracy at the cost of reduced quantity. A sparse disparity map $\boldsymbol{D}_i^S$ is then obtained. It is noteworthy that both the patch reliability and local minima constraints are bidirectional as in the LRDC [7], and we only provide the left-to-right expressions in (6) and (8) for brevity. Additional details on our proposed inter-scale decisive disparity inheritance strategy are provided in Algorithm 2.

IV. Experimental Results

A. Datasets, Implementation Details, and Evaluation Metrics

Three datasets are used in our experiments:

(1) **UDTIRI-Stereo**: Our proposed UDTIRI-Stereo dataset consists of 3,000 pairs of stereo images (resolution: $720 \times 1{,}280$ pixels), along with their disparity ground truth, collected across 12 scenarios under different illumination conditions (middle sunlight, intense sunlight, and street lighting at dark), weather conditions (tidy and watered), and road materials (asphalt and cement). We introduce random 2D Perlin noise and digital twins of real-world potholes to the road data.
(2) **Stereo-Road** [7]: This dataset provides 71 pairs of well-rectified stereo road images (resolution: $609 \times 1{,}240$ pixels), collected in Bristol, UK.
(3) **Middlebury** [51]: This dataset comprises 15 pairs of stereo images along with their corresponding disparity ground truth, collected across various indoor scenes.

We first validate the effectiveness of adapting stereo matching DCNNs (without any model fine-tuning), including PSMNet [10], AANet [52], BGNet [53], LacGwc [54], GMStereo [55], CreStereo [23], and two domain generalization-aimed networks, GraftNet [28] and HVTStereo [26], pre-trained on the KITTI Stereo dataset [56] and SceneFlow dataset [57], using our proposed D3Stereo strategy for road disparity estimation. Additionally, we compare our proposed D3Stereo matching algorithm (abbreviated as PT-D3Stereo), which employs NCC along with PT for cost volume pyramid construction, with three state-of-the-art (SoTA) explicit programming-based stereo matching algorithms: PT-SRP [7], PT-FBS [13], and PT-SGM [14], developed specifically for road surface 3D reconstruction. The above two experiments were conducted on the UDTIRI-Stereo and Stereo-Road datasets. Moreover, we employ the same experimental setups to conduct additional experiments on the Middlebury dataset to further demonstrate the effectiveness of our proposed D3Stereo strategy for general stereo matching. Finally, D3Stereo is applied to backbone DCNNs pre-trained on the ImageNet

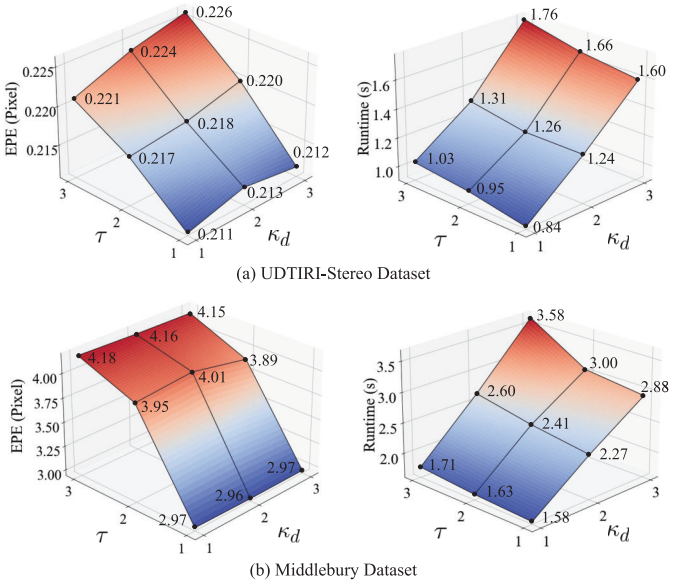


Fig. 3. Experimental results regarding hyperparameter selection for decisive disparity diffusion.

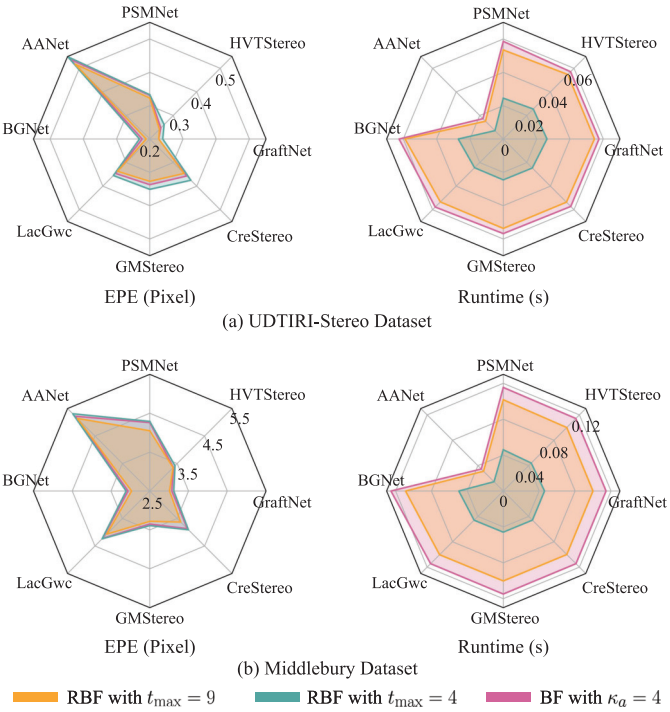


Fig. 4. Comparisons between RBF and RF when having identical computational complexity or an identical receptive field.

database for demonstrating its compatibility with general-purpose backbone DCNNs. All experiments were conducted on an NVIDIA RTX 4090 GPU.

Since the UDTIRI-Stereo and Middlebury datasets provide disparity ground truth, we adopt end-point error (EPE) and percentage of error pixels (PEP) with a tolerance of δ pixels for performance quantification. For experiments conducted on the Stereo-Road dataset, we first warp the right images into the left view using the estimated disparity maps, and then calculate the structural similarity index measure (SSIM), mean squared error (MSE), and peak signal-to-noise ratio (PSNR) metrics between the original left images and generated images to quantify the accuracy of the stereo matching algorithms.

B. Hyperparameter Selection in Decisive Disparity Diffusion

We provide details on the selection of disparity search bound τ and diffusion neighborhood radius κ_d used in the intra-scale decisive disparity diffusion process. The EPE and runtime of stereo matching with respect to different τ and κ_d are given in Fig. 3. It can be observed that setting $\tau = \kappa_d = 1$ results in the best overall performance in stereo matching, and increasing τ and κ_d leads to a noticeable increase in both the EPE and runtime. This phenomenon can be attributed to the introduction of more unreliable disparity candidates when using higher τ and κ_d during decisive disparity diffusion.

C. Ablation Study

We first conduct an ablation study to evaluate the cost aggregation efficiency between RBF and BF, and their impacts on stereo matching accuracy. As discussed in Sect. III-A, BF with $\kappa_a = 4$ has an identical receptive field compared to RBF with $t_{\max} = 4$, and has identical theoretical computational complexity compared to RBF with $t_{\max} = 9$. Therefore, we compare the performance of eight stereo matching networks using RBF and BF with these three parameter settings, as presented in Fig. 4. It can be observed that with an identical receptive field, our RBF with $t_{\max} = 4$ witnesses significantly improved cost aggregation efficiency while leading to slightly decreased stereo matching accuracy, compared to BF with $\kappa_a = 4$. We attribute this accuracy gap to the low effectiveness of small filtering kernels in RBF in aggregating matching costs between long-distance pixels. Specifically, the small kernels are unable to establish direct interactions between pixels with similar disparities and intensities but at long distances, thus making RBF less effective in areas with intense texture variations. However, with identical theoretical computational complexity, BRF with $t_{\max} = 9$ exhibits superiority in both cost aggregation efficiency and stereo matching accuracy. These improvements can be attributed to RBF's recursive filtering process, which expands the receptive field to aggregate matching costs from more related pixels. Additionally, the smaller filtering kernel in RBF results in decreased computations in the weighting initialization process compared to BF. In general, our proposed RBF strikes a better balance between stereo matching accuracy and cost aggregation efficiency compared to BF.

We conduct another ablation study to determine the essential components of D3Stereo strategy. The baseline setup includes the PKRN-based hierarchical refinement constraint and linear hierarchical refinement structure used in the DFM [18], and the traditional BF. As shown in Table I, D3Stereo with all these components achieves the highest stereo matching accuracy on both datasets. Secondly, PT significantly improves both the stereo matching accuracy and efficiency, being consistent with the findings in [7]. Additionally, RBF improves stereo matching when combined with any of the components. Specifically, EPE decreases by 23.7-40.9% on our UDTIRI-Stereo dataset and 5.7-16.2% on the Middlebury dataset. Although

TABLE I

ABLATION STUDY ON THE ESSENTIAL COMPONENTS OF D3STEREO STRATEGY. ↓ REPRESENTS THAT LOWER VALUES CORRESPOND TO BETTER PERFORMANCE. ↑ REPRESENTS THAT HIGHER VALUES CORRESPOND TO BETTER PERFORMANCE. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

(a) Experimental results on the UDTIRI-Stereo dataset.

PT	PRC	RBF	ARS	PEP (%) ↓		EPE (pixel)↓	Runtime (s)↓
				δ=0.5	δ=1		
				10.5	3.57	0.48	0.79
✔				9.15	3.27	0.38	0.62
✔	✔			9.13	3.22	0.37	0.64
✔		✔		9.22	2.51	0.29	0.91
✔			✔	9.60	3.64	0.44	**0.59**
✔	✔	✔		4.54	2.23	0.28	0.94
✔		✔	✔	3.64	1.37	0.26	0.90
✔	✔		✔	8.31	2.75	0.35	0.61
✔	✔	✔	✔	**3.53**	**1.20**	**0.21**	0.84

(b) Experimental results on the Middlebury dataset.

PRC	RBF	ARS	PEP (%) ↓		EPE (pixel)↓	Runtime (s) ↓
			δ=1	δ=2		
			34.3	24.0	4.07	1.11
✔			34.2	23.8	3.75	1.03
	✔		25.2	17.6	3.41	1.89
		✔	34.4	23.8	4.17	**0.94**
✔	✔		25.2	17.4	3.15	1.90
	✔	✔	25.4	17.8	3.46	1.77
✔		✔	33.1	22.8	3.39	0.96
✔	✔	✔	**24.9**	**17.1**	**2.97**	1.58

RBF increases the runtime of D3Stereo by over 50% on the Middlebury dataset, this increase is lower to 35% on our UDTIRI-Stereo dataset when used in conjunction with PT, which significantly reduces the stereo matching search range. Finally, the collaboration between our introduced PRC and the ARS leads to improvements in both the stereo matching accuracy and efficiency. However, when used alone, they either improve only disparity estimation accuracy or efficiency while decreasing the other.

We further validate the effectiveness of our introduced PRC and the alternating hierarchical refinement structure. The percentage of error pixels and invalid pixels of decisive disparities in $\boldsymbol{D}_1^S$ yielded using different hierarchical refinement strategies are provided in Fig. 5, where γ represents the PKRN threshold in the decisive disparity initialization process at the k-th layer. Our observations indicate that the alternating structure considerably increases the quantity of decisive disparities while leading to decreased accuracy. In contrast, our introduced PRC significantly enhances the accuracy of the decisive disparities, albeit with a decrease in their quantity. However, when these two components are used jointly, both the quantity and reliability of decisive disparities are significantly improved. While the quantity of decisive disparities obtained using our alternating hierarchical refinement strategy is lower than that achieved by DFM, our structure ensures a more uniform distribution of decisive disparities, resulting in improvements in both disparity estimation accuracy and efficiency in D3Stereo. Moreover, our hierarchical refinement

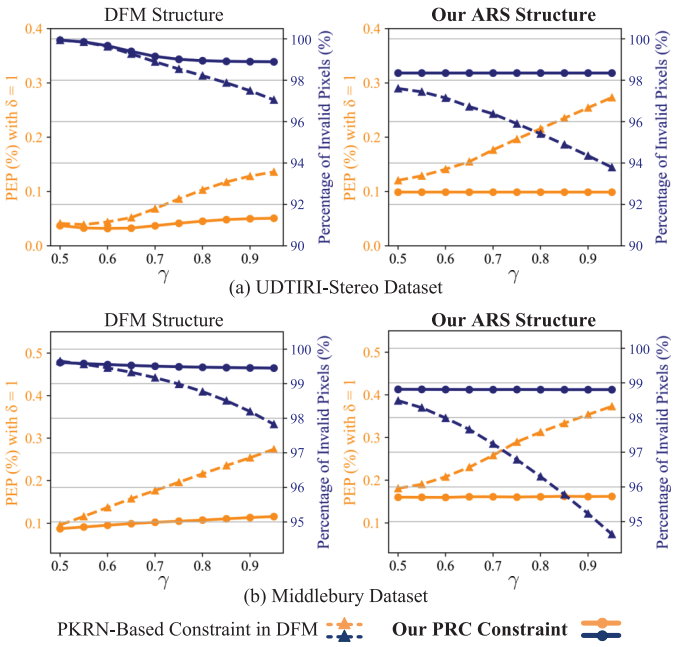


Fig. 5. Quantitative results of $\boldsymbol{D}_1^S$ yielded with different hierarchical refinement structures and constraints.

strategy is significantly less sensitive to the impact of γ, demonstrating better adaptivity compared to DFM.

D. Comparisons With SoTA Algorithms for Road Surface 3D Reconstruction

The quantitative and qualitative experimental results on our created UDTIRI-Stereo dataset are presented in Table II and Fig. 6, respectively. It is noticeable that when applying our proposed D3Stereo strategy to the existing stereo matching algorithms, EPE and PEP decrease by up to 63.10% and 83.26%, respectively. Although CreStereo demonstrates comparable performance in PEP with $\delta = 1$ when adapted to the UDTIRI-Stereo dataset using D3Stereo strategy, it shows dramatic improvement in PEP with $\delta = 0.5$ and EPE. Moreover, PT-D3Stereo diffuses decisive disparities across the entire image in a multi-directional fashion. This results in significantly improved disparity estimation results and a more uniform distribution of errors compared to other explicit programming-based stereo matching algorithms.

Secondly, the quantitative and qualitative experimental results on the Stereo-Road dataset are presented in Table III and Fig. 7, respectively. It is observed that the D3Stereo strategy can improve all algorithms across different evaluation metrics, with increases ranging from 0.41% to 10.63% in PSNR and 0.97% to 13.15% in SSIM, as well as a decrease ranging from 4.19% to 54.98% in MSE. Additionally, we observe that by using the D3Stereo strategy, disparity estimation near or on discontinuities can be significantly improved, as highlighted with green dashed boxes in Fig. 7.

It is noteworthy that GraftNet and HVTStereo achieve similar stereo matching accuracy on road scenes compared to stereo matching networks without additional domain generalization designs. Additionally, applying D3Stereo to both GraftNet and HVT-Stereo yields improved stereo match-

TABLE II

EXPERIMENTAL RESULTS ON THE UDTIRI-STEREO DATASET. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

(a) Comparison among SoTA explicit programming-based disparity estimation algorithms developed specifically for road surface 3D reconstruction.

Method	PEP (%) ↓		EPE (pixel) ↓	PSNR (dB) ↑	MSE ↓	SSIM ↑
	δ=0.5	δ=1				
PT-SRP [7]	43.2	33.0	1.27	33.01	52.12	0.867
PT-FBS [13]	33.7	10.2	0.75	32.87	48.80	0.895
PT-SGM [14]	7.17	4.13	0.72	32.50	**41.12**	0.932
PT-D3Stereo (Ours)	**5.90**	**2.99**	**0.65**	**33.14**	42.58	**0.951**

(b) Comparisons of SoTA stereo matching networks without and with our proposed D3Stereo strategy applied.

Method	PEP (%) ↓		EPE (pixel) ↓	PSNR (dB) ↑	MSE ↓	SSIM ↑
	δ=0.5	δ=1				
PSMNet [10]	46.7	13.2	0.84	32.36	55.51	0.894
PSMNet+D3Stereo (Ours)	**4.81**	**2.21**	**0.31**	**34.75**	**32.80**	**0.953**
AANet [52]	35.4	8.69	0.56	34.01	38.14	0.932
AANet+D3Stereo (Ours)	**11.1**	**1.79**	**0.43**	**34.21**	**36.49**	**0.948**
BGNet [53]	12.7	1.51	0.26	34.59	36.10	0.948
BGNet+D3Stereo (Ours)	**3.53**	**1.20**	**0.21**	**34.72**	**34.07**	**0.954**
LacGwc [54]	18.6	2.87	**0.36**	34.45	36.91	0.945
LacGwc+D3Stereo (Ours)	**4.61**	**2.01**	0.37	**34.61**	**34.28**	**0.953**
GMStereo [55]	23.6	4.12	0.43	32.66	47.39	0.937
GMStereo+D3Stereo (Ours)	**3.91**	**1.46**	**0.31**	**34.66**	**33.48**	**0.953**
CreStereo [23]	6.02	**1.41**	0.40	**34.77**	34.85	0.950
CreStereo+D3Stereo (Ours)	**4.58**	1.50	**0.34**	34.60	**33.87**	**0.952**
GraftNet [28]	20.9	2.78	0.33	34.72	36.69	0.943
GraftNet+D3Stereo (Ours)	**4.25**	**1.34**	**0.22**	**35.17**	**32.35**	**0.951**
HVTStereo [26]	6.83	1.95	0.36	34.75	35.65	0.937
HVTStereo+D3Stereo (Ours)	**4.91**	**1.67**	**0.26**	**34.79**	**31.11**	**0.951**

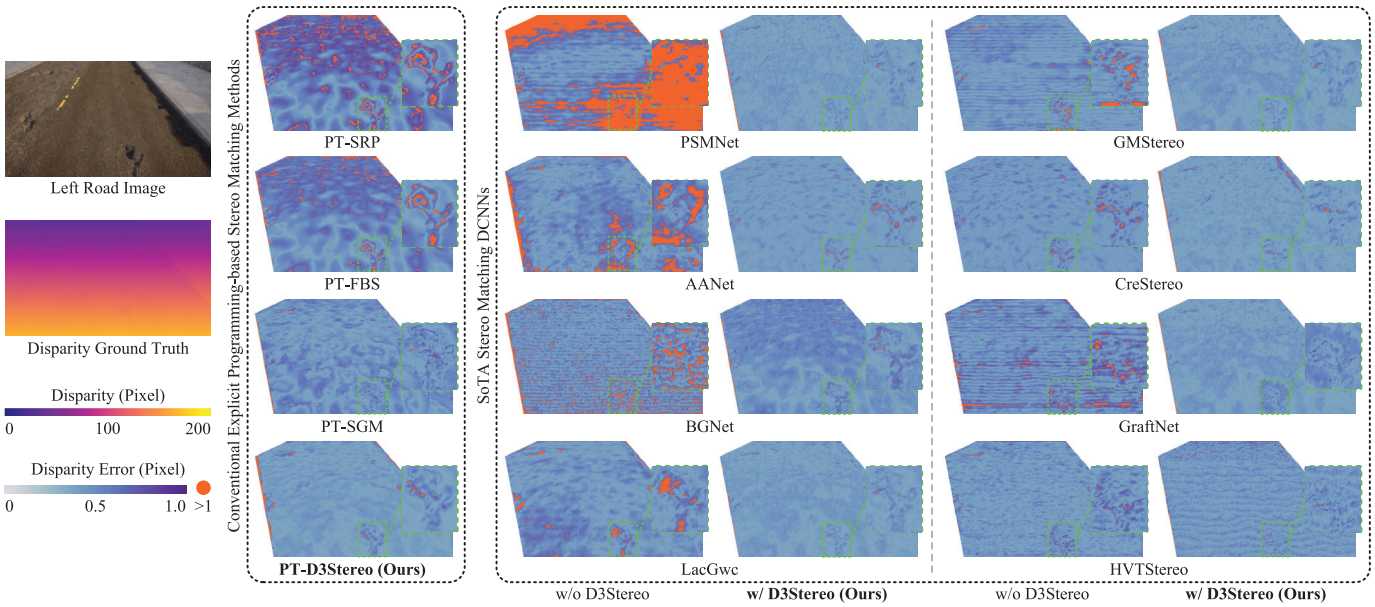


Fig. 6. Examples of disparity estimation results on our created UDTIRI-Stereo dataset.

ing accuracy in all metrics on both the UDTIRI-Stereo and Stereo-Road datasets. These results indicate a considerable domain gap between road scenes and other common indoor/outdoor scenes, and further demonstrate the superiority of D3Stereo in solving the road surface 3D reconstruction task. Additionally, we notice from the above experimental results that explicit programming-based algorithms developed specifically for road surface 3D reconstruction yield comparable performance to the DCNNs pre-trained on the KITTI Stereo 2015 dataset. Specifically, PT-D3Stereo outperforms most DCNNs without applying D3Stereo strategy in the majority of cases. These results suggest that the generalizability of SoTA

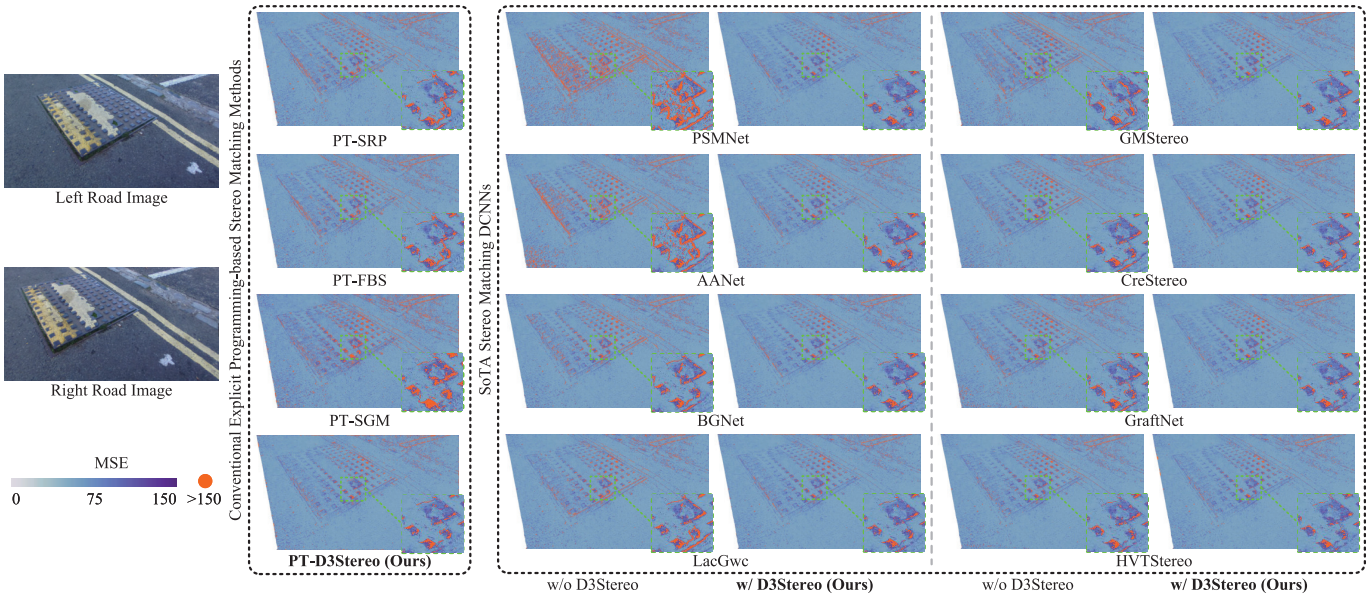


Fig. 7. Examples of stereo image reconstruction results on the Stereo-Road dataset.

TABLE III

EXPERIMENTAL RESULTS ON THE STEREO-ROAD DATASET [7]. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

(a) Comparison among SoTA explicit programming-based disparity estimation algorithms developed specifically for road surface 3D reconstruction.

Method	PSNR (dB) ↑	MSE ↓	SSIM ↑
PT-SRP [7]	30.82	68.91	0.909
PT-FBS [13]	30.83	68.14	0.911
PT-SGM [14]	30.08	74.80	0.906
PT-D3Stereo (Ours)	**31.00**	**63.37**	**0.930**

(b) Comparisons of SoTA stereo matching networks without and with our proposed D3Stereo strategy applied.

Method	PSNR (dB) ↑	MSE ↓	SSIM ↑
PSMNet [10]	28.68	122.5	0.829
PSMNet+D3Stereo (Ours)	**31.73**	**55.15**	**0.938**
AANet [52]	29.79	90.49	0.885
AANet+D3Stereo (Ours)	**31.34**	**60.50**	**0.931**
BGNet [53]	31.46	59.67	0.928
BGNet+D3Stereo (Ours)	**31.65**	**56.17**	**0.937**
LacGwc [54]	31.32	62.02	0.923
LacGwc+D3Stereo (Ours)	**31.70**	**55.43**	**0.938**
GMStereo [55]	30.98	67.10	0.915
GMStereo+D3Stereo (Ours)	**31.61**	**56.34**	**0.937**
CreStereo [23]	31.44	60.25	0.927
CreStereo+D3Stereo (Ours)	**31.57**	**57.05**	**0.936**
GraftNet [28]	30.88	65.91	0.917
GraftNet+D3Stereo (Ours)	**31.71**	**55.50**	**0.937**
HVTStereo [26]	31.43	59.10	0.933
HVTStereo+D3Stereo (Ours)	**31.63**	**56.32**	**0.938**

DCNNs is still not sufficiently satisfactory for road disparity estimation. When applying our proposed D3Stereo strategy to DCNNs, a SoTA performance is achieved.

E. Generalizability Evaluation for General Stereo Matching

We further evaluate the generalizability of D3Stereo strategy for general stereo matching using the Middlebury dataset. The quantitative and qualitative experimental results are presented in Table IV and Fig. 1 in the supplementary material, respectively. The quantitative results suggest that domain generalization-aimed networks, GraftNet and HVTStereo, achieve higher stereo matching accuracy on the Middlebury dataset. However, for stereo matching networks without any domain generalization functionality incorporated, applying D3Stereo strategy to these DCNNs results in significantly improved stereo matching accuracy. Specifically, the EPE decreases by 3.54-33.69%, while the PEP with $\delta = 1$ and $\delta = 2$ decreases by 2.59-36.26% except for LacGwc, which achieves slightly lower PEP with $\delta = 2$ compared with LacGwc without using D3Stereo strategy. The qualitative results indicate that these pre-trained DCNNs perform poorly in ambiguous regions where color intensities are similar,

TABLE IV

EXPERIMENTAL RESULTS OF SOTA DCNNS ON THE MIDDLEBURY DATASET. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Method	PEP (%) ↓		EPE (pixel) ↓
	δ=1	δ=2	
PSMNet [10]	53.5	25.7	5.33
PSMNet+D3Stereo (Ours)	**34.1**	**23.6**	**4.01**
AANet [52]	38.9	27.2	6.62
AANet+D3Stereo (Ours)	**31.8**	**21.2**	**4.39**
BGNet [53]	27.1	19.8	4.28
BGNet+D3Stereo (Ours)	**24.9**	**17.1**	**2.97**
LacGwc [54]	27.4	**17.1**	4.35
LacGwc+D3Stereo (Ours)	**25.9**	18.2	**4.07**
GMStereo [55]	29.7	19.3	3.39
GMStereo+D3Stereo (Ours)	**27.3**	**18.8**	**3.27**
CreStereo [23]	31.2	22.1	3.77
CreStereo+D3Stereo (Ours)	**28.6**	**19.5**	**3.33**
GraftNet [28]	**22.7**	**12.3**	**2.67**
GraftNet+D3Stereo (Ours)	23.1	15.6	3.01
HVTStereo [26]	**21.1**	**11.8**	**2.16**
HVTStereo+D3Stereo (Ours)	30.5	21.4	3.35

TABLE V

EXPERIMENTAL RESULTS OF FOUR BACKBONE DCNNS PRE-TRAINED ON THE IMAGENET [20] DATABASE. H AND W DENOTE THE HEIGHT AND WIDTH OF THE INPUT IMAGE, RESPECTIVELY

Backbones	Sizes of the selected feature maps	UDTIRI-Stereo dataset		Middlebury dataset	
		EPE (pixel)↓	Runtime (s)↓	EPE (pixel)↓	Runtime (s)↓
VGG [45]	$[(\frac{H}{2},\frac{W}{2},128),(\frac{H}{4},\frac{W}{4},256)]$	0.261	0.92	3.95	1.50
	$[(\frac{H}{2},\frac{W}{2},128),(\frac{H}{4},\frac{W}{4},256),(\frac{H}{8},\frac{W}{8},512)]$	**0.241**	0.96	**3.75**	1.52
	$[(\frac{H}{2},\frac{W}{2},128),(\frac{H}{4},\frac{W}{4},256),(\frac{H}{8},\frac{W}{8},512),(\frac{H}{16},\frac{W}{16},512)]$	0.242	**0.91**	3.77	**1.37**
ResNet [58]	$[(\frac{H}{2},\frac{W}{2},64),(\frac{H}{4},\frac{W}{4},64)]$	0.314	0.84	4.02	1.22
	$[(\frac{H}{2},\frac{W}{2},64),(\frac{H}{4},\frac{W}{4},64),(\frac{H}{8},\frac{W}{8},128)]$	**0.309**	**0.81**	**3.88**	**1.06**
	$[(\frac{H}{2},\frac{W}{2},64),(\frac{H}{4},\frac{W}{4},64),(\frac{H}{8},\frac{W}{8},128),(\frac{H}{16},\frac{W}{16},256)]$	0.309	0.82	4.07	1.08
MobileNetV3 [59]	$[(\frac{H}{2},\frac{W}{2},16),(\frac{H}{4},\frac{W}{4},24)]$	0.672	0.87	4.67	1.17
	$[(\frac{H}{2},\frac{W}{2},16),(\frac{H}{4},\frac{W}{4},24),(\frac{H}{8},\frac{W}{8},40)]$	0.573	0.84	4.54	1.12
	$[(\frac{H}{2},\frac{W}{2},16),(\frac{H}{4},\frac{W}{4},24),(\frac{H}{8},\frac{W}{8},40),(\frac{H}{16},\frac{W}{16},112)]$	**0.562**	**0.81**	**4.48**	**1.08**
MobileNetV3-S [59]	$[(\frac{H}{2},\frac{W}{2},16),(\frac{H}{4},\frac{W}{4},16)]$	1.412	0.89	7.23	1.05
	$[(\frac{H}{2},\frac{W}{2},16),(\frac{H}{4},\frac{W}{4},16),(\frac{H}{8},\frac{W}{8},24)]$	**1.357**	0.82	6.44	**0.96**
	$[(\frac{H}{2},\frac{W}{2},16),(\frac{H}{4},\frac{W}{4},16),(\frac{H}{8},\frac{W}{8},24),(\frac{H}{16},\frac{W}{16},48)]$	1.366	**0.81**	**6.29**	1.01

without any further model fine-tuning. In contrast, their performance improves with the use of D3Stereo strategy.

However, our proposed D3Stereo strategy suffers from the edge-fattening issue [52], [60] and exhibits limited stereo matching accuracy near/on extensive disparity discontinuities. Specifically, large disparities from foreground objects are diffused to background objects in the intra-scale decisive disparity diffusion process. This phenomenon primarily occurs at the overlaps between different objects. Therefore, we are motivated to explicitly leverage semantic segmentation results to constrain the intra-scale disparity propagation process within each spatially continuous region in our future work, thereby exploring broader applications of our proposed D3Stereo for general stereo matching.

F. Performance Evaluation on Backbone DCNNs

Additional experiments are conducted on the UDTIRI-Stereo and Middlebury datasets with backbone DCNNs pre-trained on the ImageNet database. The quantitative and qualitative experimental results are presented in Table V and Fig. 2 in the supplementary material, respectively. Surprisingly, these backbone DCNNs demonstrate the capability to perform accurate dense correspondence matching when utilizing our proposed D3Stereo strategy. Specifically, when employing the D3Stereo strategy, VGG [45] and ResNet [58] achieve competitive results in comparison to stereo matching networks on both UDTIRI-Stereo and Middlebury datasets and outperform explicit programming-based algorithms on the UDTIRI-Stereo dataset. These results strongly suggest that our approach is effective in extending the capabilities of backbone DCNNs for solving the dense correspondence matching problem. It can also be observed that MobileNetV3 [59] has fewer feature channels compared to VGG and ResNet, resulting in lower stereo matching accuracy. We further validate this viewpoint using MobileNetV3-S [59], a lighter version of MobileNetV3. As expected, MobileNetV3-S underperforms MobileNetV3 on both UDTIRI-Stereo and Middlebury datasets, confirming that the number of feature channels is a crucial factor that significantly impacts the stereo matching accuracy.

Moreover, it is noteworthy that D3Stereo achieves improved efficiency and accuracy when employing three or four feature layers instead of two feature layers. These results suggest that deep features obtained at 1/8 and 1/16 of the full image resolution exhibit similar performance in aggregating global information and eliminating stereo matching ambiguities. Consequently, these comparable intermediate results enable the subsequent decisive disparity diffusion and hierarchical refinement processes in D3Stereo to ultimately produce dense disparity maps with similar accuracy. Finally, performing nearest neighbor search at 1/16 of the full image resolution significantly reduces the computational complexity compared to 1/8 of the full image resolution, which offsets the additional computational demands of decisive disparity diffusion and hierarchical refinement processes at 1/16 of the full image resolution, thereby maintaining the overall efficiency of D3Stereo.

V. CONCLUSION

This article introduces D3Stereo, a novel decisive disparity diffusion strategy. Our technical contributions include (1) a recursive bilateral filtering algorithm for efficient and adaptive cost aggregation, (2) an intra-scale disparity diffusion algorithm for sparse disparity map completion, and (3) an inter-scale disparity inheritance algorithm for fine-grained disparity estimation at higher resolution. Additionally, we also developed a new dataset to address the need for comprehensive performance quantification of stereo matching-based road surface 3D reconstruction algorithms. Comprehensive experiments demonstrate the effectiveness of D3Stereo in adapting pre-trained deep learning models to address the stereo matching task in both road and general scenes. In the future, we aim to further improve the accuracy of estimated disparities near/on disparity discontinuities and explore broader applications of D3Stereo for general stereo matching. This includes restraining the disparity diffusion process with each

instance and investigating the integration of D3Stereo for self-supervised stereo matching.

ACKNOWLEDGMENT

Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union-European Commission. Neither the European Commission nor the European Union can be held responsible for them.

REFERENCES

[1] N. Ma et al., "Computer vision for road imaging and pothole detection: A state-of-the-art review of systems and algorithms," *Transp. Saf. Environ.*, vol. 4, no. 4, Nov. 2022, Art. no. tdac026.

[2] X. Liang, X. Yu, C. Chen, Y. Jin, and J. Huang, "Automatic classification of pavement distress using 3D ground-penetrating radar and deep convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22269–22277, Nov. 2022.

[3] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4906–4911, Nov. 2020.

[4] M. U. U. Haq, M. Ashfaque, S. Mathavan, K. Kamal, and A. Ahmed, "Stereo-based 3D reconstruction of potholes by a hybrid, dense matching scheme," *IEEE Sensors J.*, vol. 19, no. 10, pp. 3807–3817, May 2019.

[5] A. Ahmed, M. Ashfaque, M. U. Ulhaq, S. Mathavan, K. Kamal, and M. Rahman, "Pothole 3D reconstruction with a novel imaging system and structure from motion techniques," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4685–4694, May 2022.

[6] 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.

[7] 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.

[8] 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.

[9] 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. [Online]. Available: https://ieeexplore.ieee.org/document/10693503

[10] J. R. Chang and Y. S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[11] L. Lipson, Z. Teed, and J. Deng, "RAFT-stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2021, pp. 218–227.

[12] 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*. Cham, Switzerland: Springer, 2023, pp. 63–100.

[13] R. Fan, J. Jiao, J. Pan, H. Huang, S. Shen, and M. Liu, "Real-time dense stereo embedded in a UAV for road inspection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 535–543.

[14] 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.

[15] S. Roy, "Stereo without epipolar lines: A maximum-flow formulation," *Int. J. Comput. Vis.*, vol. 34, nos. 2–3, pp. 147–161, 1999.

[16] O. Miksik, Y. Amar, V. Vineet, P. Pérez, and P. H. S. Torr, "Incremental dense multi-modal 3D scene reconstruction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 908–915.

[17] S. Pillai, S. Ramalingam, and J. J. Leonard, "High-performance and tunable stereo reconstruction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 3188–3195.

[18] U. Efe, K. G. Ince, and A. A. Alatan, "DFM: A performance baseline for deep feature matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4284–4293.

[19] C. Zhou, S. Su, Q. Chen, and R. Fan, "E3CM: Epipolar-constrained cascade correspondence matching," *Neurocomputing*, vol. 559, Nov. 2023, Art. no. 126788.

[20] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[21] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn. (CoRL)*, 2017, pp. 1–16.

[22] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," 2020, *arXiv:2001.10773*.

[23] 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.

[24] Z. Zhang, X. Ai, C. K. Chan, and N. Dahnoun, "An efficient algorithm for pothole detection using stereo vision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 564–568.

[25] Z. Rao et al., "Masked representation learning for domain generalized stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5435–5444.

[26] T. Chang, X. Yang, T. Zhang, and M. Wang, "Domain generalized stereo matching via hierarchical visual transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9559–9568.

[27] J. Zhang et al., "Revisiting domain generalized stereo matching networks from a feature consistency perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12991–13001.

[28] B. Liu, H. Yu, and G. Qi, "GraftNet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13002–13011.

[29] Z. Shen, Y. Dai, and Z. Rao, "CFNet: Cascade and fused cost volume for robust stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 13906–13915.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[31] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2006, pp. 404–417.

[32] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2548–2555.

[33] A. Barroso-Laguna and K. Mikolajczyk, "Key.Net: Keypoint detection by handcrafted and learned CNN filters revisited," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 698–711, Jan. 2023.

[34] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.

[35] M. Dusmanu et al., "D^2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8092–8101.

[36] J. Revaud, C. R. de Souza, M. Humenberger, and P. Weinzaepfel, "R2D2: Reliable and repeatable detector and descriptor," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, Sep. 2019, pp. 12405–12415.

[37] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4938–4947.

[38] I. Rocco, M. Cimpoi, R. Arandjelovic, A. Torii, T. Pajdla, and J. Sivic, "NCNet: Neighbourhood consensus networks for estimating image correspondences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1020–1034, Feb. 2022.

[39] I. Rocco et al., "Efficient neighborhood consensus networks via submanifold sparse convolutions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 605–621.

[40] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8922–8931.

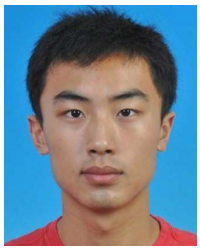
[41] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 2, Oct. 2003, pp. 900–906.

[42] 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.

[43] K. Zhang, Y. Fang, D. Min, L. Sun, S. Yang, and S. Yan, "Cross-scale cost aggregation for stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 5, pp. 965–976, May 2017.

[44] M. G. Mozerov and J. van de Weijer, "Accurate stereo matching by two-step energy minimization," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1153–1163, Mar. 2015.
[45] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," in *Proc. 2015th Int. Conf. Learn. Represent. (ICLR)*, Jan. 2015.
[46] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10428–10436.
[47] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4898–4906.
[48] 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.
[49] 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.
[50] Q. Zhou et al., "Patch2pix: Epipolar-guided pixel-level correspondences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4669–4678.
[51] D. Scharstein et al., "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2014, pp. 31–42.
[52] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1959–1968.
[53] B. Xu, Y. Xu, X. Yang, W. Jia, and Y. Guo, "Bilateral grid learning for stereo matching networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12492–12501.
[54] B. Liu, H. Yu, and Y. Long, "Local similarity pattern and cost self-reassembling for deep stereo matching networks," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1647–1655.
[55] H. Xu et al., "Unifying flow, stereo and depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13941–13958, Nov. 2023.
[56] 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.
[57] 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. (CVPR)*, Jun. 2016, pp. 4040–4048.
[58] 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.
[59] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
[60] 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.

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



Yikang Zhang received the B.Sc. degree in automation from Beijing Institute of Technology in 2017 and the M.S. degree in ECE from UMASS Amherst in 2019, specialized in model predictive control and physical unclonable functions. He is currently pursuing the Ph.D. degree, supervised by Prof. Rui Fan. His research interests include simulation, planning, and control in complex environments.

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, Shanghai. His research interests include robotics control, environmental perception, and understanding of mobile robots and bioinspired control.

Ioannis Pitas (Life Fellow, IEEE) received the Diploma and Ph.D. degrees in electrical engineering from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece. Since 1994, he has been a Professor with the Department of Informatics, AUTH, where he is the Director of the Artificial Intelligence and Information Analysis Laboratory. He was a Visiting Professor with several universities. He leads the big European H2020 Research and Development Project MULTIDRONE. His current interests are in the areas of computer vision, machine learning, autonomous systems, and image/video processing.

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 Postdoctoral Scholar-Employee with 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 with the same college and Shanghai Research Institute for Intelligent Autonomous Systems at Tongji University. 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 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. He is the General Chair of the AVVision Community; and organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, and ECCV'22. 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.