AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.

AQ:1 = Please confirm or add details for any funding or financial support for the research of this article.

AQ:2 = Please provide the publisher location for Ref. [11].

Inter-Scale Similarity Guided Cost Aggregation for Stereo Matching

Pengxiang Li, Chengtang Yao, Yunde Jia, *Member, IEEE*, and Yuwei Wu, *Member, IEEE*

***Abstract*— Stereo matching aims to estimate 3D geometry by computing disparity from a rectified image pair. Most deep learning based stereo matching methods aggregate multi-scale cost volumes computed by downsampling and achieve good performance. However, their effectiveness in fine-grained areas is limited by significant detail loss during downsampling and the use of fixed weights in upsampling. In this paper, we propose an inter-scale similarity-guided cost aggregation method that dynamically upsamples the cost volumes according to the content of images for stereo matching. The method consists of two modules: inter-scale similarity measurement and stereo-content-aware cost aggregation. Specifically, we use inter-scale similarity measurement to generate similarity guidance from feature maps in adjacent scales. The guidance, generated from both reference and target images, is then used to aggregate the cost volumes from low-resolution to high-resolution via stereo-content-aware cost aggregation. We further split the 3D aggregation into 1D disparity and 2D spatial aggregation to reduce the computational cost. Experimental results on various benchmarks (e.g., SceneFlow, KITTI, Middlebury and ETH3D-two-view) show that our method achieves consistent performance gain on multiple models (e.g., PSM-Net, HSM-Net, CF-Net, FastAcv, and FactAcvPlus). The code can be found at https://github.com/Pengxiang-Li/issga-stereo.**

***Index Terms*— Stereo matching, cost aggregation, content-aware upsampling.**

I. Introduction

STEREO matching aims to estimate a pixel-wise disparity map from a rectified image pair. It plays an important role in various applications including 3D reconstruction [1], AR [2], SLAM [3], and autonomous driving [4].

AQ:1 Manuscript received 21 November 2023; revised 17 July 2024; accepted 22 August 2024. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 62176021 and Grant 62172041, in part by the Natural Science Foundation of Shenzhen under Grant JCYJ20230807142703006, and in part by the Key Research Platforms and Projects of Guangdong Provincial Department of Education under Grant 2023ZDZX1034. This article was recommended by Associate Editor X. Sun. *(Pengxiang Li and Chengtang Yao are co-first authors.) (Corresponding authors: Yuwei Wu; Yunde Jia.)*

Pengxiang Li, Chengtang Yao, and Yuwei Wu are with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: pengxiangli@bit.edu.cn; yao.c.t@bit.edu.cn; ~~weyuwei@bit.edu.cn),~~

Yunde Jia is with Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China, and also with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology (BIT), Beijing 100081, China (e-mail: jiayunde@smbu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3453965.

Digital Object Identifier 10.1109/TCSVT.2024.3453965

The well-known pipeline divides stereo matching into four steps: cost computation, cost aggregation, disparity computation, and disparity refinement [5]. Among these four steps, cost aggregation plays a pivotal role in leveraging neighborhood information to rectify the ambiguous matching costs in ill-posed regions such as occluded regions, large textureless areas, repetitive patterns, and thin structures. The cost aggregation is commonly embedded into end-to-end deep neural networks with multi-scale processing to enlarge the receptive field. 3D CNNs [6], [7], GRU [8], [9], and attention mechanism [10] are the most commonly used basic structures for cost aggregation, effectively correcting ambiguous matching costs and substantially enhancing prediction accuracy in ill-posed regions by aggregating multi-scale cost volumes.

However, these cost aggregation methods often struggle in fine-grained areas due to considerable detail loss during downsampling and fixed weight used in upsampling. Many efforts have been targeted at improving the performance of stereo matching in fine-grained areas, including edge information [11], deformable convolutions [12], group-wise correlation [13] and slanted planes [14]. These methods have achieved good performance, but two challenging problems in cost aggregation are still not well solved: (1) the downsampling causes considerable detail loss during the construction of multi-scale cost volumes, and (2) the upsampling fixed in size and weight is prone to data imbalance between large-smooth areas and fine-grained areas. For example, the HSM-Net [7] with multi-scale cost volumes and upsampling fixed in size and weight may lead to poor performance in fine-grained areas, as illustrated in Fig. 1 (b) and Fig. 1 (c).

In this paper, we propose inter-scale similarity-guided cost aggregation that adaptively restores image details by dynamically upsampling cost volumes based on image content. Our method comprises two modules: inter-scale similarity measurement and stereo-content-aware cost aggregation. We utilize inter-scale similarity measurements to generate similarity guidance from the feature maps of adjacent scales. Subsequently, we employ this guidance to aggregate the multi-scale cost volumes through stereo-content-aware cost aggregation.

For the first challenging problem, our idea is to retrieve the fine-grained details lost during the downsampling process. We use inter-scale similarity measurement to measure the similarity between high-resolution and low-resolution features. The similarity explicitly preserves the connection between high-resolution details and low-resolution features,

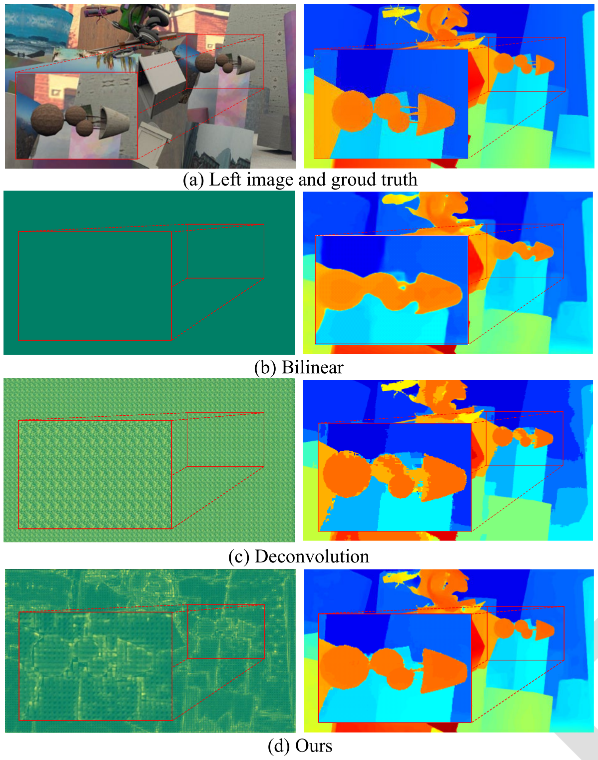


Fig. 1. Predictions and upsampling weights visualizations of HSM-Net [7] using different upsampling strategies.

thereby providing guidance for the upsampling process to restore details. Technically, we first project a point $\boldsymbol{x}_{high}$ from high-resolution to low-resolution features $\boldsymbol{x}_{low}$. Then we compute the similarity between the point in the high-resolution $\boldsymbol{x}_{high}$ and the neighbors of the projected point in low-resolution $\mathcal{N}_{low}$.

For the second challenging problem, a critical factor leading to the suboptimal restoration of fine-grained details is the fixed size and weights of existing upsampling strategies, which are unable to adapt to the complicated fine-grained details. Motivated by this, we replace the fixed upsampling with content-aware upsampling. The content-aware upsampling uses the content information of each point to guide the upsampling process, thereby mitigating the impact of data imbalance between large-smooth and fine-grained areas. In stereo-content-aware cost aggregation, we use similarity guidance (generated from both reference and target images) to guide the aggregation of matching costs in 3D spatial-disparity space. The pair-wise 3D upsampling is computationally expensive. Thus, we split the upsampling in the 3D space into 1D disparity and 2D spatial space. As a result, our method is able to efficiently and adaptively assemble the proper neighbors for cost aggregation and upsampling. Our method generates upsampling weight according to the image content and achieves much finer details, as shown in Fig. 1 (d). Our method can be plugged into any multi-scale cost volume based stereo network and achieve higher accuracy, especially in fine-grained areas.

Our contributions are summarized as follows:

1) We propose an inter-scale similarity guided cost aggregation method to adaptively recover the details of cost volumes under the guidance of similarity generated from images.
2) We introduce an inter-scale similarity measurement to dynamically generate guidance by incorporating information from both low-resolution and high-resolution feature maps. The explicit utilization of high-resolution feature maps ensures the preservation of fine-grained details.
3) We design a decomposition strategy that splits 3D disparity-spatial upsampling into 1D disparity and 2D spatial upsampling, significantly reducing the computational cost of the 3D pair-wise upsampling.

II. Related Work

A. Stereo Matching

Traditional stereo matching methods estimate disparity maps for rectified image pairs using local [15], [16], global [17], [18], and semi-global methods [19], [20], [21]. Deep learning-based stereo matching networks now dominate, delivering state-of-the-art results. Early deep learning methods replaced steps in stereo matching [5]: cost computation [22], [23], [24], cost aggregation [25], [26], [27], disparity computation [28], and disparity refinement [26], [27]. Despite good performance, their non-end-to-end approaches limited data utilization. To overcome this, end-to-end methods compute correlations by warping the target image to the reference image [8], [9], [11], [29]. These achieve excellent results but often lose geometric information. Cost-volume-based models [6], [30], [31], [32], [33], [34], [35] preserve geometric information by concatenating multi-scale cost volumes. State-of-the-art methods use convolution neural networks [36], [37], [38], [39], [40], [41] or attention mechanisms [10], [42], [43], [44] to aggregate these volumes, effectively utilizing image context information.

However, multi-scale cost volume-based stereo matching methods often lose fine-grained details due to downsampling. While cost aggregation usually recovers these details, current fixed-size and fixed-weight schemes struggle with data imbalances between large smooth, and fine-grained areas. To address this, we developed a content-aware cost aggregation method that mitigates detail loss during multi-scale cost volume creation. Our adaptive upsampling approach also remains robust against data imbalances.

B. Cost Aggregation

Multi-scale cost aggregation methods [6], [29] enhance matching cost reliability by optimizing multi-scale cost volumes for precise disparity estimation. Song et al. [11] used edge information to guide cost aggregation, reducing edge mismatches. Zhang et al. [45] improved efficiency by replacing 3D CNNs with semi-global aggregation. Yang et al. [7] proposed a hierarchical feature volume decoder for high-resolution image disparity estimation. Xu et al. [12] utilized deformable convolution for adaptive aggregation. Lipson et al. [8] designed an iterative mixed disparity sampling and aggregation strategy. Liu et al. [46] used local features to address over-smoothing.

Zhang et al. [47] introduced depth-based sampling for balanced density in close and far regions. Xu et al. [48] utilized bilateral grid processing for faster aggregation. Lee et al. [49] introduced a cluster-wise cost aggregation algorithm to parallelized scanline-level disparity computation.

The aforementioned methods demonstrate commendable performance, even in ill-posed areas. However, they still suffer from the loss of details in downsampling, and their strategies for multi-scale cost aggregation are susceptible to data imbalance. These strategies commonly rely on either bilinear interpolation or deconvolution for upsampling. Both bilinear interpolation and deconvolution employ a fixed interpolation rule or deconvolution kernel across all data points, thus failing to exploit the content information of images fully. Constrained by computational memory limitations, these methods are unable to perform direct aggregation at full resolution. Instead, they resort to upsampling to full resolution without introducing additional parameters after aggregating at 1/2 or 1/4 resolution. However, relying solely on parameter-free upsampling is inadequate for recovering lost details.

C. Upsampling

Upsampling is used to transform data from low-resolution to high-resolution. Traditional upsampling strategies fit a curve of a small neighborhood of the upsampled points to compute values for interpolated points, including nearest neighbor interpolation [50], bilinear interpolation [51], trilinear interpolation [52], and bicubic interpolation [53], etc. The advantage of these methods lies in their low computational cost. However, these parameter-free upsampling strategies underutilize image content, resulting in blurred recovery results in fine-grained areas. Deconvolution [54], [55], [56], [57] offers a learning-based approach to upsampling, where weights are optimized through backpropagation. Learning-based upsampling kernels enable the utilization of contextual information learned from extensive data. However, deconvolution has limitations as it struggles in various scenes due to fixed kernel sizes and weights, making it susceptible to data imbalances.

Several works [58], [59], [60] use content-aware upsampling operators to solve the fixed-weight problem. Wang et al. [58], [59] presented a content-aware reassembly approach and argued that traditional feature upsampling methods struggle to capture rich semantic information. While content-aware upsampling mitigates the fixed-weight problem, it relies solely on information from the low-resolution side (i.e., the upsampling process could be regarded as a unary low-resolution to high-resolution mapping). However, the upsampling process inherently consists of both low-resolution and high-resolution components, and relying solely on low-resolution features for upsampling may not suffice. Instead of employing a unary upsampling mapping, we introduce an inter-scale similarity measurement approach to produce a pair-wise upsampling mapping, represented by similarity guidance derived from information gathered across adjacent scales. In other words, we actually model the upsampling process as a binary mapping between low-resolution and high-resolution.

III. Optimization in Multi-Scale Cost Aggregation

In this section, we model the optimization objectives for each layer of multi-scale cost aggregation. Given a cost volume $\boldsymbol{C}_{l-1} \in \boldsymbol{R}^{H_{l-1}\times W_{l-1}\times D_{l-1}}$ at level $l-1$ as input, $\boldsymbol{C}_l$ is computed via a network with learning weights $\boldsymbol{W}_l$. The generation of $\boldsymbol{C}_l$ can be formulated as

$$\begin{aligned} \mathrm{p}(\boldsymbol{C}_l) &= \mathrm{p}(\boldsymbol{C}_l|\boldsymbol{C}_{l-1}, \boldsymbol{W}_l)\mathrm{p}(\boldsymbol{W}_l)\mathrm{p}(\boldsymbol{C}_{l-1}) \\ &= \mathrm{p}(\boldsymbol{C}_l|\boldsymbol{C}_{l-1}, \boldsymbol{W}_l)\mathrm{p}(\boldsymbol{W}_l). \end{aligned} \tag{1}$$

The probability $\mathrm{p}(\boldsymbol{C}_l)$ of cost volume is commonly computed by $\mathrm{p}(\boldsymbol{C}_l) = \mathrm{softmax}(-\boldsymbol{C}_l)$, and $p(\boldsymbol{C}_{l-1})$ is supposed to be 1 as $\boldsymbol{C}_{l-1}$ has already been given. Then, the optimization objective is to find the best $\boldsymbol{W}_l$ that recovers the details lost in $\boldsymbol{C}_{l-1}$, which can be formulated as

$$\begin{aligned} \boldsymbol{W}_l &= \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \mathrm{p}(\boldsymbol{W}_l|\boldsymbol{C}_l, \boldsymbol{C}_{l-1}), \\ &= \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \mathrm{p}(\boldsymbol{W}_l|\boldsymbol{C}_l), \\ &= \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \frac{\mathrm{p}(\boldsymbol{C}_l|\boldsymbol{W}_l)\cdot \mathrm{p}(\boldsymbol{W}_l)}{\sum_{\boldsymbol{W}_l} \mathrm{p}(\boldsymbol{C}_l|\boldsymbol{W}_l)\mathrm{p}(\boldsymbol{W}'_l)\mathrm{d}\boldsymbol{W}'_l}, \\ &\overset{a.s.}{=} \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \mathrm{p}(\boldsymbol{C}_l|\boldsymbol{W}_l)\cdot \mathrm{p}(\boldsymbol{W}_l), \\ &\overset{a.s.}{=} \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\, \mathrm{p}(\boldsymbol{C}_l). \end{aligned} \tag{2}$$

In the aforementioned cost aggregation process, it becomes impractical to recover the details lost during downsampling using bilinear upsampling or deconvolution. This is because $\boldsymbol{W}_l$ is optimized by cost volumes at level $[0, 1, \ldots, l-1]$, and it doesn't consider the image content at level l. In other words, only minimal details at level l contribute to the optimization of $\boldsymbol{W}_l$. Furthermore, the kernel weights are influenced by the content that appears more frequently in the image. Consequently, it becomes challenging to utilize these fixed kernel weights effectively for recovering details that constitute only a small proportion of the image content such as the fine-grained areas.

IV. Proposed Method

A. Problem Formulation

Detail loss and biased upsampling are two challenging problems that cause poor performance in fine-grained areas. To address these two problems, we optimize cost aggregation with image features at levels l and $l-1$. In our method, the optimization objective of cost aggregation at each level is given by

$$\boldsymbol{W}_l = \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\, \mathrm{p}(\boldsymbol{C}_l)\cdot \mathrm{p}(\boldsymbol{W}_l|\boldsymbol{F}_l, \boldsymbol{F}_{l-1}), \tag{3}$$

where $\boldsymbol{F}_l$ is the feature map at level l.

In particular, the optimization objective of cost aggregation with deconvolution is actually one special case of ours, where $\mathrm{p}(\boldsymbol{W}_l|\boldsymbol{F}_l, \boldsymbol{F}_{l-1}) = \mathrm{p}(\boldsymbol{W}_l)$. Besides, the optimization objective of cost aggregation with bilinear interpolation is one special case of deconvolution, i.e., Eq. (2). With substituting $\mathrm{p}(\boldsymbol{W}_l) =$ 1 into Eq. (2), Eq. (2) can be reformulated as

$$\boldsymbol{W}_l = \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \mathrm{p}(\boldsymbol{C}_l), \tag{4}$$

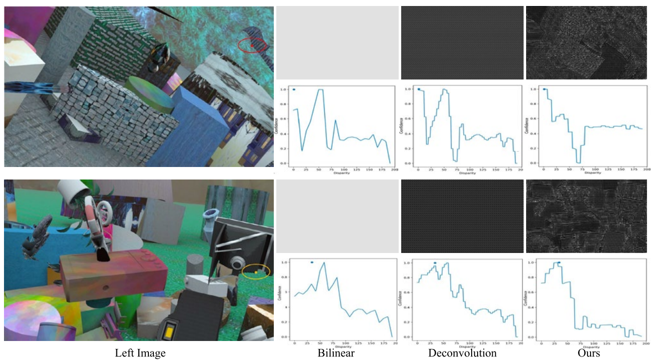


Fig. 2. The visualization of aggregation weight and disparity distribution in cost volume. The upper row shows the aggregation weight and the under row shows the distribution of the cost volume along the disparity dimension for a single point. The point in each distribution map is the ground truth for the point in the reference image. Both the bilinear upsampling and deconvolution predict wrong results, while ours not only predicts the correct disparity but also corrects for multi-modality in the distribution.

which is just the optimization objective of cost aggregation with bilinear interpolation.

In our method, $\boldsymbol{W}_l$ is automatically adjusted from the change of $\boldsymbol{F}_l$ and $\boldsymbol{F}_{l-1}$ during inference, whereas the weights of deconvolution or bilinear interpolation remain static. Our method generates aggregation weight related to the image content and achieves unimodal distribution results, while others get multimodal distribution or wrong distribution. Fig. 2 provides a visual representation of aggregation using various upsampling strategies. As shown in Fig. 2, the weights for bilinear interpolation remain constant, the weights for deconvolution are repetitive kernels, while our method's weights are content-aware, closely linked to the image's content. It's also worth noting that our method effectively addresses the issue of multiple peaks in the disparity distribution (see the distribution curves of Deconvolutions vs. Ours in Fig. 2). In our method, the disparity distribution exhibits only a single prominent peak precisely at the ground truth disparity, whereas deconvolution may exhibit multiple peaks, potentially leading to incorrect disparity results.

B. Implementation

Given an image pair, we extract multi-scale feature maps $\boldsymbol{F}_l$ at each level l for reference and target images. We then use the feature maps to construct the cost volume at the lowest level. As for the cost volume at the high level, we iteratively upsample the cost volume from the low level to the high level through two steps, the inter-scale similarity measurement, and the stereo-content-aware cost aggregation. The inter-scale similarity measurement uses feature maps from adjacent scales to generate similarity guidance. The stereo-content-aware cost aggregation uses the similarity guidance from two views to guide the cost volume upsampling. At last, we use the cost volume at the highest level to compute the disparity map as the output of our network. Fig. 3 illustrates the pipeline of our method.

1) Inter-Scale Similarity Measurement: The inter-scale similarity measurement takes the feature maps $\boldsymbol{F}_l$ and $\boldsymbol{F}_{l-1}$ as input. We compute the similarity by the summation of the products of $\boldsymbol{F}_l(h', w')$ and the neighbors of $\boldsymbol{F}_{l-1}(h, w)$ with the formula as

$$\boldsymbol{S}_l(h', w') = \frac{1}{M \cdot M}\phi\Big(\sum_{(h,w)\in\mathcal{N}_F} \boldsymbol{F}_l(h', w')\boldsymbol{F}_{l-1}(h, w)\Big), \quad (5)$$

where (h', w') and (h, w) are the location at high-level and low-level respectively, $(h', w') = (h \cdot s, w \cdot s)$, s is the scale change in resolution from level $l-1$ to level l, and $\cdot$ is the scalar multiplication operation. $\boldsymbol{S}_l \in \mathbb{R}^{H^l \times W^l}$ is the similarity guidance at level l, $\boldsymbol{S}_l(h', w')$ is the value of the pixel at location (h', w'), $\mathcal{N}_F \in \mathbb{R}^{M \times M}$ is a 2D neighborhood of the pixel at location (h, w) with the size of $M \times M$. $\phi(\cdot)$ is a subnetwork composed of convolution layers, relu layers, and batch normalization layers.

2) Stereo-Content-Aware Cost Aggregation: 3D convolution based methods [6], [7] usually perform window based cost aggregation:

$$\boldsymbol{C}_l(h', w', d') = \sum_{(h,w,d)\in\mathcal{N}_c} \boldsymbol{W}_l(h', w', d')\boldsymbol{C}_{l-1}(h, w, d), \quad (6)$$

where $\mathcal{N}_c$ is a 3D neighborhood of the point at $(h'/s, w'/s, d'/s)$.

In our method, we replace the 3D weight $\boldsymbol{W}_l$ with the 2D similarity guidance $\boldsymbol{S}_l$. For each level, we use the feature maps of the stereo images, i.e., reference and target images, to compute the content-aware similarity guidance $\boldsymbol{S}_l^R$ and $\boldsymbol{S}_l^T$ by inter-scale similarity measurement, respectively. Then we perform the cost aggregation guided by $\boldsymbol{S}_l^R$ and $\boldsymbol{S}_l^T$:

$$\boldsymbol{C}_l(h', w', d') = \sum_{(h,w,d)\in\mathcal{N}_c} \boldsymbol{S}_l^R(h', w')\boldsymbol{S}_l^T(h', w'-d') \boldsymbol{C}_{l-1}(h, w, d). \quad (7)$$

The memory and computational cost of 3D cost aggregation are unaffordable. Accordingly, we introduce a decomposition strategy to reduce the computation cost. We split the upsampling in full 3D spatial-disparity space into 1D disparity and 2D spatial upsampling by leveraging the property of cost volume on the disparity dimension. The property is that position (h, w, d) in cost volume represents the (h, w) in the reference image and $(h, w-d)$ in the target image. We warp $\boldsymbol{S}_l^T$ to $\boldsymbol{S}_l^R$, and then split the mapping of cost volume into 1D disparity dimension and 2D spatial dimension. Specifically, we replace Eq. (7) with a two-step decomposed cost aggregation.

In the first step, 1D disparity upsampling, the positions $(h, w, d-\lfloor M/2\rfloor), \ldots, (h, w, d), \ldots, (h, w, d+\lfloor M/2\rfloor)$ in cost volume along disparity dimension correspond to (h, w) in the reference image and $(h, w-d+\lfloor M/2\rfloor), \ldots, (h, w-d), \ldots, (h, w-d-\lfloor M/2\rfloor)$ in the target image. Formally, the updating along the disparity dimension is given by

$$\boldsymbol{C}_l(h, w, d') = \sum_{d\in\mathcal{N}_d} \boldsymbol{S}_l^R(h', w')\boldsymbol{S}_l^T(h', w'-d')\boldsymbol{C}_{l-1}(h, w, d), \quad (8)$$

where $\mathcal{N}_d = \{d'/s - \lfloor M/2\rfloor, \ldots, d'/s, \ldots, d'/s + \lfloor M/2\rfloor\}$. In the second step, 2D spatial upsampling, all voxels with location $(h', w', :)$ in cost volume correspond to the pixel with

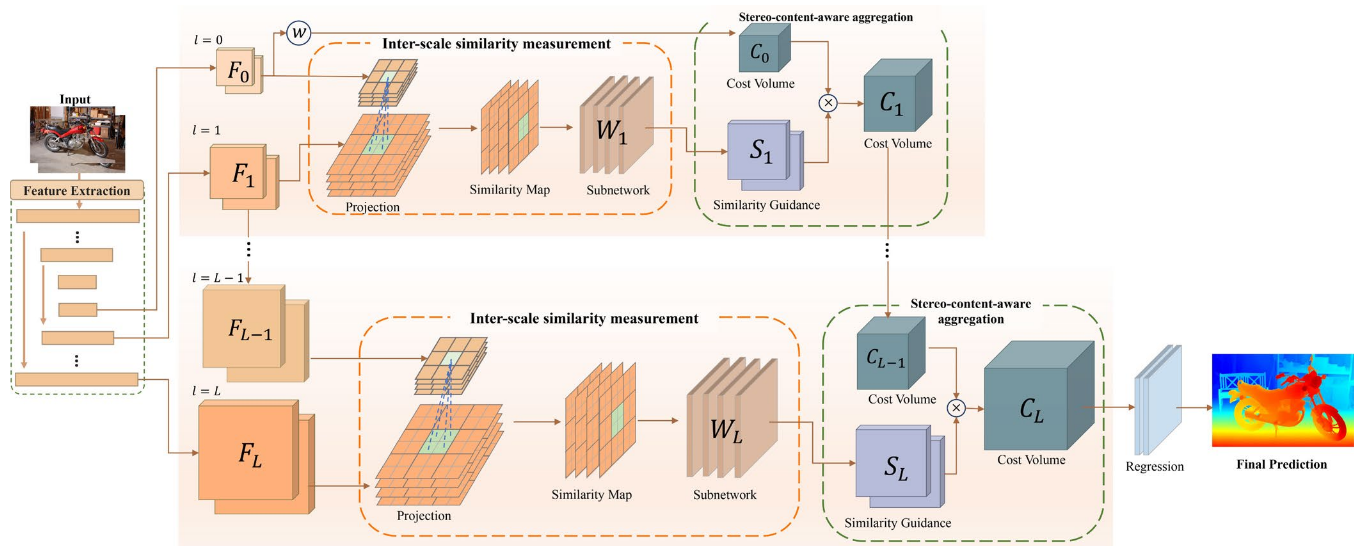


Fig. 3. The overall architecture. Given an image pair, our method extracts multi-scale features at each level l. L is the total number of levels. Ⓦ is the warping operation used for generating the initial cost volume in the lowest resolution. $\otimes$ represents element-wise multiplication.

location (h', w') in the reference image. The update along the spatial dimension is given by

$$\boldsymbol{C}_l(h', w', d') = \sum_{(h,w)\in\mathcal{N}_{sp}} \boldsymbol{S}_l^R(h', w')\boldsymbol{C}_l(h, w, d'), \tag{9}$$

where $\mathcal{N}_{sp} \in \mathbb{R}^{\mathbb{M}\times\mathbb{M}}$ is a 2D neighborhood of the pixel with location $(h'/s, w'/s)$ at level $l-1$.

After all these operations, we complete the transformation from the shape of $H_{l-1} \times W_{l-1} \times D_{l-1}$ to $H_{l-1} \times W_{l-1} \times D_l$ and then to $H_l \times W_l \times D_l$, where $H_l = H_{l-1} \cdot s$, $W_l = W_{l-1} \cdot s$ and $D_l = D_{l-1} \cdot s$.

3) Loss Function: We use a multi-scale loss function that applies smooth L_1 loss to each level. The smooth L_1 loss function is not sensitive to outliers or noises. The loss function is defined as

$$\boldsymbol{D}_l = \sum_{d\in\{d_n\}_{n=1}^N} d \cdot \sigma(-\boldsymbol{C}_l), \tag{10}$$

$$\mathcal{L} = \sum_{l=0} \lambda_l \cdot \mathcal{L}_l(\boldsymbol{D}_l - \boldsymbol{G}_l), \tag{11}$$

$$\mathcal{L}_l(x) = \begin{cases} 0.5\ x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}, \tag{12}$$

where $\{d_n^l\}_{n=1}^N$ is the disparity hypothesis at level l, $\sigma(\cdot)$ is the softmax operation, $\boldsymbol{D}_l$ is the predicted disparity map at level l, λ_l denotes the coefficients for the disparity prediction at level l, and $\boldsymbol{G}_l$ is the ground-truth disparity map at level l.

C. Computational Cost Analysis

To further demonstrate the superiority of our decomposition strategy in computational complexity, we conducted the following analysis and complexity experiments (in Section V). We separate the 3D upsampling into 1D upsampling plus 2D upsampling, reducing the parameters and calculations.

1) Parameters: For deconvolution, the number of parameters per layer is given by $C \times 1 \times k^3 = Ck^3$, where k is the kernel size, C is the number of input channels, and the output channel is set to 1. In contrast, our method requires Ck^2 parameters per layer. Both our method and deconvolution utilize the same number of layers.

2) Calculations: For the computational complexity of 3D upsampling, comparing 3D deconvolution with our method for a feature volume of size $C \times D \times H \times W$ to be upsampled by a scale of s, the computational cost for deconvolution is $\mathcal{O}(s^3k^3)CDHW$, while ours is $\mathcal{O}(s^2k^2 + sk)CDHW = \mathcal{O}(s^2k^2)CDHW$.

V. Experiments

A. Datasets

1) SceneFlow Dataset: SceneFlow [29] is a large synthetic dataset containing 34896 training images and 4248 testing images with the size of 540×960. This dataset has three rendered sub-datasets: FlyingThings3D, Monkaa, and Driving. FlyingThings3D is rendered from the ShapeNet dataset and has 21828 training data and 4248 testing data. Monkaa is rendered from the animated film Monkaa and has 8666 training data. The Driving is constructed by the naturalistic, dynamic street scene from the viewpoint of a driving car and has 4402 training samples.

2) KITTI 2015 Dataset: KITTI 2015 [61] is a real-world dataset with street views from a driving car. It contains 200 training stereo image pairs with sparse ground-truth disparities obtained using LiDAR and another 200 testing image pairs without ground-truth disparities. During the training process, we take 160 images for training and reference 40 images for validation.

3) Middlebury-v3 Dataset: Middlebury-v3 is a subset of the 2014 dataset [62] and is collected in the real world with static indoor scenes containing complicated and rich details. There are 15 stereo pairs for training and 15 stereo pairs for testing.

Each pair is provided in 3 kinds of resolution, full, half, and quarter resolution; where we used the quarter resolution in the experiment.

4) ETH3D-Two-Iew Dataset: ETH3D (two view) [63] comprises grayscale stereo pairs captured from diverse indoor and outdoor scenes. The dataset includes 27 training and 20 testing image pairs with sparsely labeled ground truth. Disparities range from 0 to 64 pixels, and bad 1.0 (percentage of pixels with errors larger than 1 pixel) are reported.

B. Evaluation Metrics

The end-point error (EPE) is the mean disparity error in pixels. The 3-px error refers to the proportion of points in the full map where the absolute value of the difference between the predicted disparity and the true value is greater than 3 pixels. The percentage of disparity outliers in the background (*D1-bg*), foreground (*D1-fg*), or all pixels (*D1-all*) for both noc regions and all regions are applied for evaluation. Disparity outliers are the pixels if their disparity EPE is more than 3 or 5% pixels. *avgerr* is the average absolute error in pixels. *RMS* represents the root mean square of the differences between the estimated and actual disparities. *A90* and *A95* are the 90% and 95% error quantile in pixels, respectively. *Bad 1.0* and *Bad 2.0* percentage of pixels with errors larger than 1 pixel or 2 pixels, respectively.

C. Model Details

To prove the effectiveness of our method, we extend five stereo baseline networks with our method, CF-Net [30], HSM-Net [7] and PSM-Net [6], FastAcv [44] and FastAcvPlus [44]. All networks are implemented via PyTorch and tested on NVIDIA RTX 3090 GPU. For all baselines, the neighborhood size M is set to the scale change s at each level.

For PSM-Net+ours, the model is optimized using Adam [64] with β_1 of 0.9, β_2 of 0.999. During training, the batch size is fixed to 8, and we perform color normalization to each input image and crop them into 256×512 resolution. We train our network on SceneFlow for 10 epochs and change the learning rate from 0.001 to 0.0001 in the 7th epoch. We then fine-tune the network on KITTI 2015 and set the learning rate to 0.001, 0.0001, and 0.00003 in the first 200 epochs, the next 400 epochs, and the final 600 epochs, respectively. As for Middlebury-v3, we also fine-tune the model pre-trained on SceneFlow. The learning rate is set to 0.001 for 300 epochs and then changed to 0.0001 for the rest of 600 epochs. For HSM-Net+ours, we use AdamW [65] with β_1 of 0.9, β_2 of 0.999. During training, the batch size is fixed to 12, and we perform the same data augmentation [7] of the original HSM-Net and crop the images into 256×512 resolution. We train our network for 10 epochs using the same dataset as HSM-Net and change the learning rate from 0.001 to 0.0001 in the 9th epoch. For CF-Net, FastAcv, and FastAcvPlus, we follow all the training strategies of the official repositories.

Furthermore, we downsample the ground truth for our multi-scale loss. We use bilinear downsampling in SceneFlow and nearest downsampling in KITTI 2015 and Middlebury-v3.

TABLE I

EVALUATION RESULTS OF CURRENT STEREO MATCHING ALGORITHMS ON THE SCENEFLOW TEST SET

Model	PSM-Net [6] (2018)	Gwc-Net [13] (2019)	HSM-Net [7] (2019)
EPE(px)	1.09	0.98	2.07
Model	Acf-Net (2020)	LEAStereo (2020)	CF-Net (2021)
EPE(px)	0.87	0.78	0.99
Model	LaC+ GwcNet (2022)	LaC+ GANet (2022)	FastAcv (2023)
EPE(px)	0.75	0.72	0.64
Model	FastAcvPlus (2023)	PSM-Net+ours	HSM-Net+ours
EPE(px)	0.59	0.63	1.39
Model	CF-Net+ours	FastAcv + ours	FastAcvPlus + ours
EPE(px)	0.72	0.59	**0.57**

Additionally, we reduced the computational cost without sacrificing accuracy by moving the averaging operation before the aggregation at each layer. Although we observe better results of bilinear downsampling in the experiment on SceneFlow, the ground truth disparities of the two real-world datasets contain invalid values, like 0 and INF, which will lead to wrong disparity results after bilinear downsampling. In all experiments, no post-processing or unsupervised learning methods are used.

D. Comparison With Stereo Matching Methods

Our method mainly focuses on recovering the fine-grained details lost during cost volume downsampling. Therefore, we conduct experiments on the SceneFlow dataset, specifically targeting fine-grained areas, and we compare the results against mainstream baseline methods. Additionally, we perform experiments on real datasets, including KITTI and Middlebury, to validate the effectiveness of our approach.

1) SceneFlow: The experimental results in Table I show that our proposed method significantly improves the performance of stereo matching algorithms, with the FastAcvPlus+ours achieving the lowest EPE of 0.57. The consistent reduction in EPE across various models demonstrates the robustness and efficacy of our method.

a) Fine-grained areas: We test different baselines in the fine-grained region on the SceneFlow dataset to verify the accuracy improvement of our method in the fine-grained (FG) areas and full areas, as shown in Table II. We use the calculated HOG [66] descriptor of the reference image as a mask of fine-grained areas. The results in Table II show the superiority of our method in fine-grained areas. Our method can improve the accuracy significantly in fine-grained areas, and 37.6%, 32.9%, 16.2%, 11.4% and 10.4% EPE reduction in PSM-Net, HSM-Net, CF-Net, FastAcv and FastAcvPlus, respectively. Our method is effective for different baselines with good universality. Our method also brings no or small increase in runtime. For PSM-Net, we remove the time-consuming 3D convolution layers in the hourglass modules at RES 1/16 and RES 1/8. For the rest baselines, we directly plug our method into them without additional model modification.

The visualization results for the fine-grained regions are depicted in Fig. 4. Our method successfully recovers more details, notably improving estimation results for fine-grained areas like the spokes of the wheel and plant spikes in the left column of Fig. 4 compared to the baseline. Furthermore, our approach enhances results in less refined regions, such as inside the bounding box in the right column of Fig. 4.

TABLE II

RESULTS OF DIFFERENT BASELINE IN FULL AREAS (FULL) AND FINE-GRAINED (FG) AREAS ON THE SCENEFLOW DATASET. FOR PSM-NET, WE REMOVE THE TIME-CONSUMING 3D CNNS AT RES 1/16 AND RES 1/8

Method	EPE in Full	EPE in FG	Time (s)
PSM-Net [6] (2018)	1.09	1.01	0.41
PSM-Net+ours	0.60	0.63	0.37
HSM-Net [7] (2019)	1.88	2.07	0.05
HSM-Net+ours	1.25	1.39	0.09
CF-Net [30] (2021)	1.06	0.99	0.18
CF-Net+ours	0.72	0.83	0.22
FastAcv [44] (2023)	0.64	0.70	0.05
FastAcv+ours	0.59	0.62	0.08
FastAcvPlus [44] (2023)	0.59	0.67	0.05
FastAcvPlus+ours	0.57	0.60	0.08

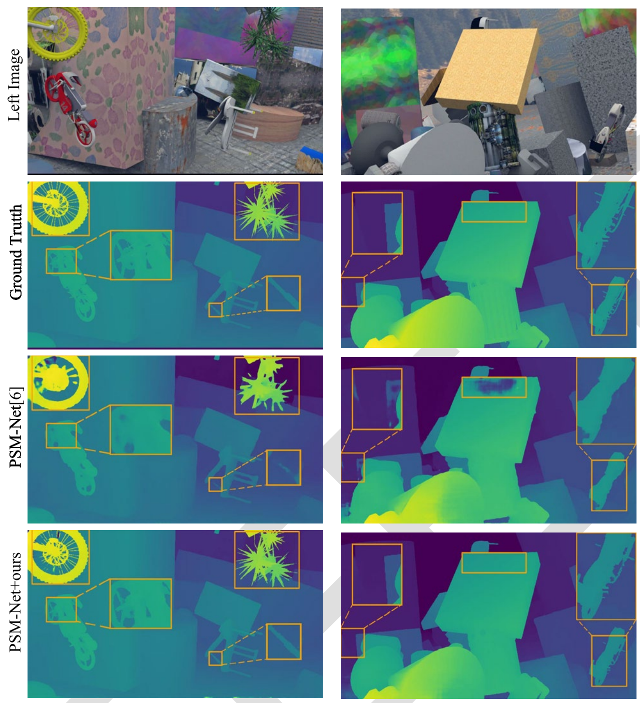


Fig. 4. The results of PSM-Net [6] and PSM-Net+ours on the SceneFlow dataset.

b) Full areas: Our method brings improvement for full areas across different baselines. Visualization of the results (Fig. 5) reveals that our method exhibits certain corrective effects on large-scale weakly-textured regions as well. The experimental results demonstrate that our approach achieves significant accuracy improvement when applied to datasets with complete depth information as ground truth.

2) Kitti: Table III displays the performance and runtime of various algorithms evaluated on the KITTI stereo2015 benchmark. Across different baselines, our method enhances the accuracy of the original baselines with only a marginal increase in processing time. Notably, CF-Net combined with our method surpasses other competing methods in the Noc D1-all and All D1-all. Next, we would like to provide a more detailed explanation of the comparison between our method and stereo matching methods based on attention mechanisms, as well as methods based on decomposition strategies.

TABLE III

EVALUATION ON KITTI 2015 BENCHMARK. THE BEST RESULTS FOR EACH EVALUATION METRIC ARE SHOWN IN BOLD

Models	Noc (%)			All (%)			Time (s)
	bg	fg	all	bg	fg	all	
PSM-Net [6] (2018)	1.71	4.31	2.14	1.86	4.62	2.32	0.41
GwcNet-g [13] (2019)	1.61	3.49	1.92	1.74	3.93	2.11	0.32
HSM-Net [7] (2019)	1.63	3.40	1.92	1.80	3.85	2.14	0.05
RAFT-Stereo [8] (2021)	-	-	-	2.89	**1.75**	1.96	-
HDA-Net [67] (2021)	1.55	3.32	1.84	1.69	3.76	2.03	0.42
BG-Net [48] (2021)	1.91	4.34	2.31	2.07	4.74	2.51	0.02
Dec-Net [31] (2021)	1.89	3.53	2.16	2.07	3.87	2.37	0.05
CF-Net [30] (2021)	**1.43**	3.25	1.73	**1.54**	3.56	1.88	0.18
ChiT-12 [42] (2022)	2.11	3.79	2.38	2.34	4.05	2.60	-
FC-PSMNet [39](2022)	1.73	4.19	2.13	1.86	4.61	2.32	-
HTSGM [49] (2022)	-	-	5.66	-	-	5.84	-
FastACV-Plus [44] (2023)	-	-	-	1.70	3.53	2.01	0.05
PSM-Net+ours	1.50	3.42	1.82	1.73	3.78	2.07	0.38
HSM-Net+ours	1.64	3.21	1.90	1.77	3.60	2.09	0.10
CF-Net+ours	1.46	**2.95**	**1.70**	1.58	3.30	**1.87**	0.22

a) Compared with the attention-based method: HDA-Net [67] proposes an efficient horizontal attention module to adaptively capture the global correspondence clues. Our method uses inter-scale information to generate similarity guidance to improve cost aggregation. As shown in Table III, our method has lower D1-all (HDA-Net 2.03 *vs.* CF-Net+Ours 1.87) with faster running time (HDA-Net 0.42ms *vs.* CF-Net+Ours 0.22ms) on the KITTI 2015 dataset.

b) Compared with the decomposition method: DecNet [31] decomposes the original stereo matching into a dense matching at the lowest resolution and a series of sparse matching at higher resolutions. Unlike DecNet, our method decomposes the 3D upsampling of cost volume into a 2D-spatial and 1D-disparity upsampling. Our method outperforms DecNet in D1-all (Dec-Net 2.37 *vs.* HSM-Net+Ours 2.09) but is slower in runtime (Dec-Net 0.05ms *vs.* HSM-Net+Ours 0.09ms), as shown in Table III.

c) Visualization: Fig. 6 presents the experimental results on the KITTI 2015 dataset, showcasing images from top to bottom. Our method excels in recovering slender structures, as seen in the iron chain at the center of the first row and the fence in the lower left corner of the third row. Moreover, our approach accurately estimates depth-mutation areas such as signboards and utility poles. For instance, unlike PSM-Net and HSM-Net in the first row's bounding box around the signboard, our method produces correct results. In rows two, five, and six, the other methods misidentify parts of the background as utility poles, which our method avoids.

3) Middlebury: We compare our method with several approaches using different aggregation strategies on the Middlebury stereo dataset v3, as shown in Table IV. We outperform these 3D aggregation based approaches on most of the metrics. The result also demonstrates the effectiveness of our content-aware upsampling method. Based on the visualizations in Fig. 7, we can draw the following conclusions: **1) Improved depth estimation for fine-grained regions**: Our method shows superior performance in depth estimation for fine-grained regions, demonstrating the effectiveness of explicitly integrating high-resolution and low-resolution information. This is evident in almost all cases, such as the

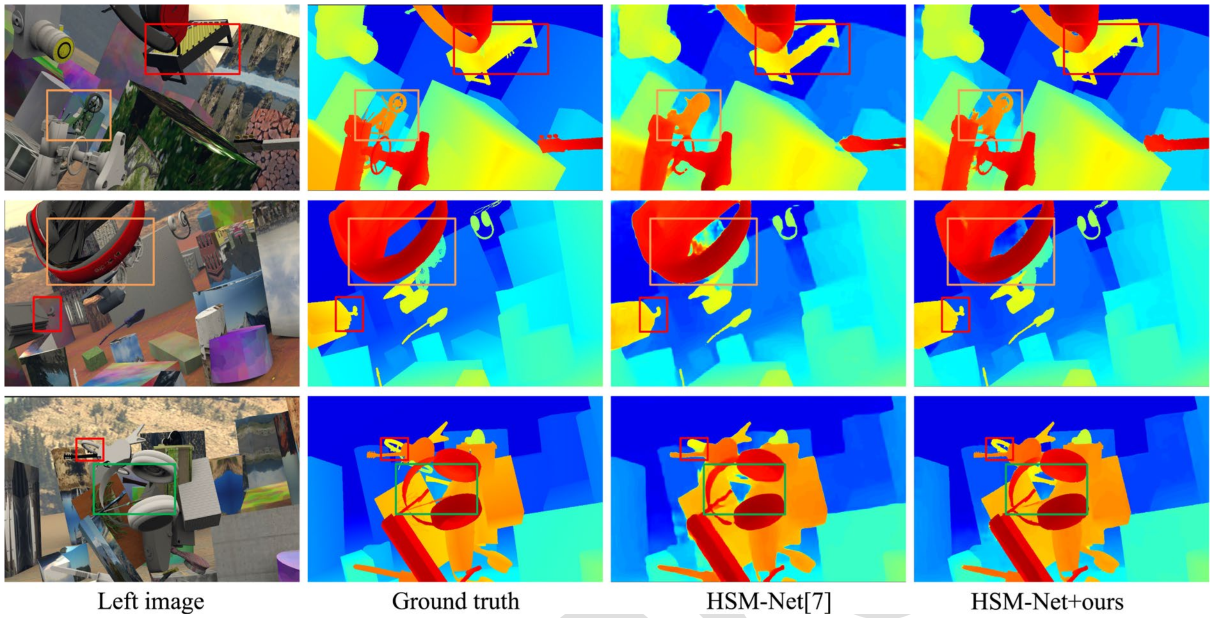


Fig. 5. The visualization of results on HSM-Net baseline. The first and second columns are the reference input images and ground truth. The rest columns are results from HSM-Net and HSM-Net+ours.

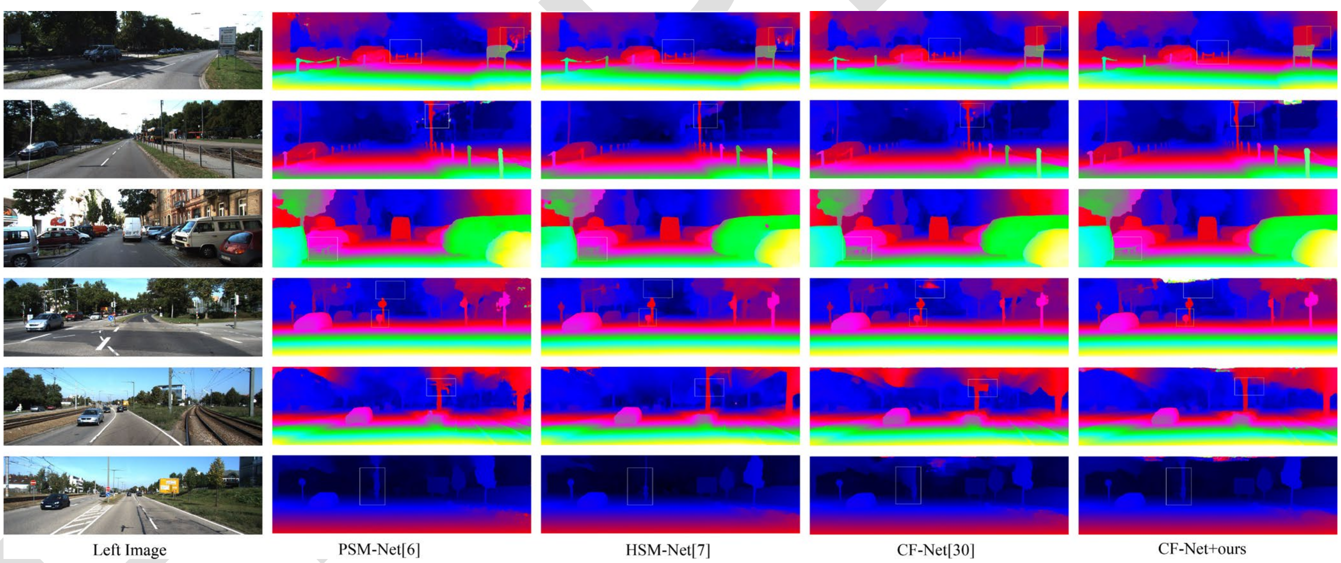


Fig. 6. The results of different deep stereo networks on KITTI 2015 dataset. Our method performs better in fine-grained areas than other methods, especially for the region denoted with the boxes. Please zoom in to check the details.

detailed areas in “DjembL” and the water cup on the table in “Crusade” (Line 3, PSM-Net vs. PSM-Net + ours), as well as the small figurine (Line 3, HSM-Net vs. HSM-Net + ours). **2) Enhanced foreground and background decoupling**: Our method has a stronger ability to decouple the foreground from the background. Retaining low-resolution information effectively enhances this capability. Examples include the depth estimation of the potted plants and background in “Plants” (Line 5) and the estimation of the hollow part of the staircase handrail in “Staircase” (Line 4, HSM-Net vs. HSM-Net + ours; CF-Net vs. CF-Net + ours). **3) Competitive performance in flat regions**: Our method also shows competitive performance in flat regions. For instance, the wall in the upper left of “Staircase” (Line 4, PSM-Net vs. PSM-Net + ours) and the restoration of the table corner in “Crusade” (Line 3, CF-Net vs. CF-Net + ours). However, our method has some shortcomings in certain areas, such as the seats in the PSM-Net case of “Classroom2E” (Line 1, PSM-Net vs. PSM-Net + ours). We will systematically discuss these limitations in the Limitation Analysis section.

E. Ablation Studies

We conduct all the analysis in ablation studies mainly on the HSM-Net baseline. Ablation studies are performed on the SceneFlow dataset and the KITTI 2015 dataset.

1) Effectiveness of Stereo-Content-Aware Cost Aggregation: During Stereo-Content-Aware Cost Aggregation, we use both

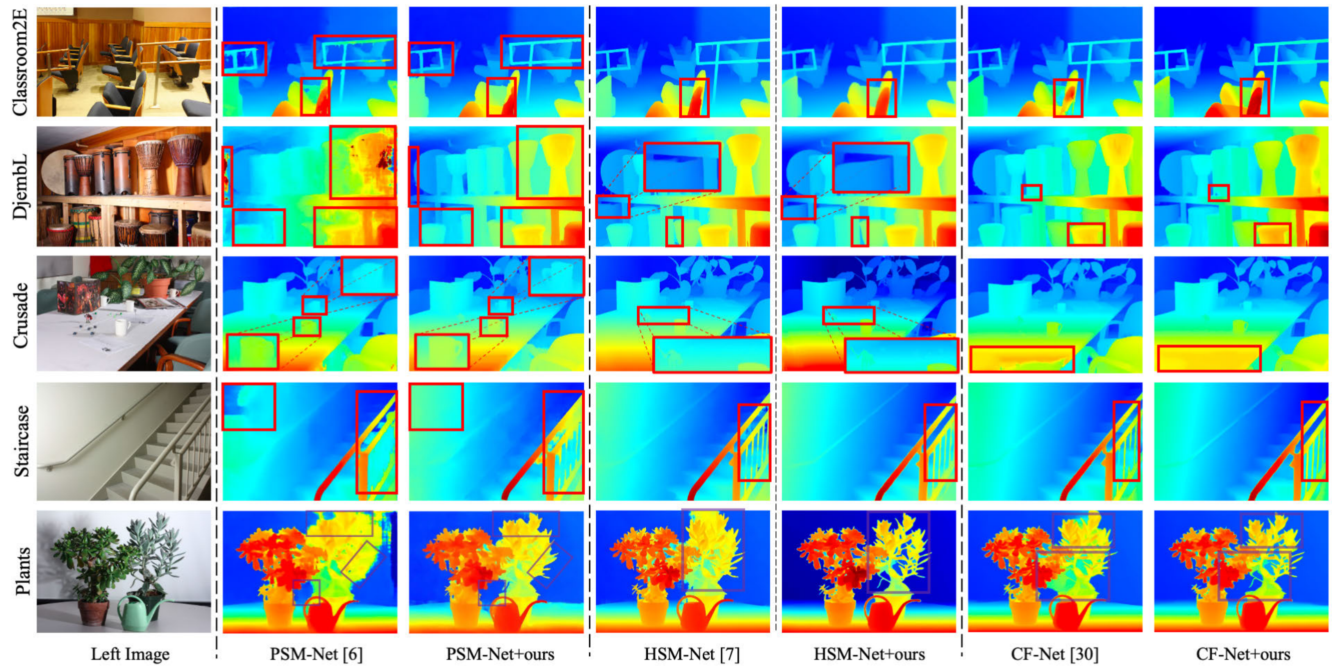


Fig. 7. The visualization of results on Middlebury-v3 test set. In the first column are the reference input images. The rest columns are results from PSM-Net [6], PSM-Net + ours, HSM-Net [7], + ours, CF-Net [30] and CF-Net + ours respectively.

TABLE IV
EVALUATION ON MIDDLEBURY-V3. THE BEST RESULTS FOR EACH EVALUATION METRIC ARE SHOWN IN BOLD

Models	Res	Avgerr	Rms	A90	A95
PSM-Net_ROB (2018) [6]	Q	8.78	23.3	22.8	43.4
DeepPruner (2019) [68]	Q	6.56	18	17.9	33.1
FADNet++ (2021) [69]	Q	11.9	27.7	34.3	61.2
MCP-HA-VQ (2022) [70]	Q	6.01	37.5	40.6	85.9
H-CENST (2022) [71]	Q	10.2	29.1	24.3	59.0
FM-DT (2023) [72]	Q	11.7	31.4	33.4	67.1
PSM-Net+ours	Q	**5.43**	**17.3**	**8.11**	**25.2**

reference and target images to extract similarity guidance and separate the 3D spatial-disparity upsampling into 1D disparity / 2D spatial upsampling. We evaluate the effectiveness of our method at different resolutions through two experiments: i. Training on the SceneFlow dataset and testing on the SceneFlow dataset. ii. Training on the SceneFlow dataset and testing on the validation set of the KITTI 2015 dataset. Table V demonstrates that our decomposition strategy reduces the running time by nearly half compared to full 3D upsampling at the setting of "RES 1/16 to RES 1/8" and "RES 1/8 to RES 1" on the SceneFlow dataset and KITTI 2015 dataset. Our decomposition strategy not only proves to be faster but also more accurate than full 3D upsampling. When integrating our method at "RES 1/16 to 1/8," HSM-Net+ours experiences a decrease in EPE of 18.09% and 15.86% compared to the original HSM-Net on the SceneFlow dataset and the KITTI 2015 dataset, respectively. Plugging our method at higher resolutions, i.e., "RES 1/8 to 1", the EPE of HSM-Net+ours is 33.51% and 26.21% lower than the original HSM-Net on the SceneFlow dataset and the KITTI 2015 dataset, respectively. Our method is effective and the higher the resolution at which we employ our method, the greater the improvement it brings.

2) Effectiveness of Inter-Scale Similarity Measurement: We utilize inter-scale similarity measurement to generate a similarity guidance map for cost aggregation. Each pixel in the similarity map corresponds to the content information at the same location. Our method calculates the similarity between high-resolution feature points and their corresponding $M \times M$ points in the low-resolution counterpart. Visualizations of similarity maps of a 3 size neighborhood are shown in Fig. 8.

We confirm the effectiveness of our inter-scale policy on the SceneFlow dataset through a series of experiments. These experiments are conducted in three settings: without similarity guidance, with single-scale similarity guidance, and with inter-scale similarity guidance. The results presented in Table VI clearly demonstrate that the use of inter-scale similarity guidance results in higher accuracy when compared to single-scale similarity guidance. The inter-scale similarity guidance transforms the unary mapping inherent in single-scale similarity guidance into a pair-wise mapping, consequently leading to improved accuracy. Furthermore, we verify the significance of employing stereo information, which includes both reference and target images, to achieve favorable results. In Table VI, it is evident that the EPE when using stereo information is significantly lower than when not using stereo information. Utilizing stereo information to model the mapping relationship between cost volumes of different resolutions proves to be more reliable than relying solely on reference images.

3) Effectiveness of Our Method in Different Resolution: We further provide visualizations of the results obtained from HSM-Net and HSM-Net+ours at different resolutions on the SceneFlow dataset. These visualizations help us understand how our model enhances the baseline at various resolutions, as shown in Fig. 9. At a resolution of 1/32, HSM-Net

TABLE V

RESULTS OF USING THE GUIDANCE IN MULTIPLE STEPS OF MULTI-SCALE COST AGGREGATION ON SYNTHETIC AND REAL DATASETS. RES 1/16, 1/8, 1 REPRESENTS THE ORIGINAL IMAGE'S 1/16, 1/8, AND 1 RESOLUTION. RES 1/16 TO 1/8 INDICATES WHETHER THE BASELINE IS PLUGGED WITH OUR METHOD IN COST AGGREGATION FROM RESOLUTION 1/16 TO RESOLUTION 1/8, SO AS RES 1/8 TO 1

Models	Inter-scale		3D upsampling		SceneFlow		KITTI 15	
	RES 1/16 to 1/8	RES 1/8 to 1	Full 3D	2D + 1D	EPE	Times (s)	EPE	Times (s)
HSM	-	-	-	-	1.88	0.05	1.45	0.05
HSM+ours	✓	-	✓	-	1.81	0.17	1.37	0.18
HSM+ours	✓	-	-	✓	1.54	0.06	1.22	0.08
HSM+ours	✓	✓	✓	-	1.67	0.27	1.32	0.24
HSM+ours	✓	✓	-	✓	1.25	0.09	1.07	0.10

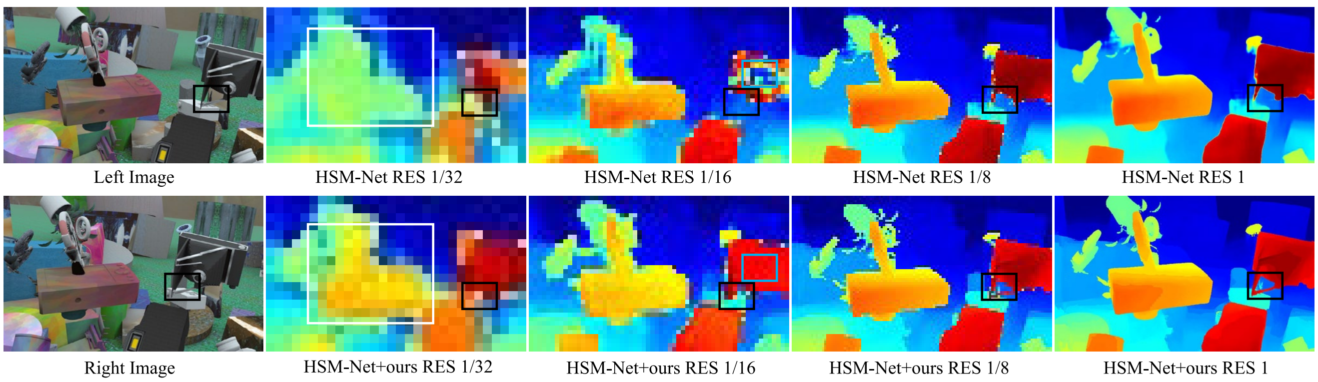


Fig. 8. The visualization of similarity. (a) and (b) are the similarity of the two images. The three columns on the right are visualizations of the similarity, representing the similarity of points in high resolution to their corresponding neighbors in low-resolution projection points. In each map, the brightness indicates the similarity, which corresponds to the upsampling kernel weight. It can be seen that the aggregation weight is directly related to the image content and that each weight in the global picture uniquely adapts the content information of the corresponding points.

TABLE VI

THE RESULTS OF USING DIFFERENT SCALES OF GUIDANCE TO GUIDE COST AGGREGATION. "INTER-SCALE" AND "SINGLE-SCALE" REPRESENT THAT THE GUIDANCE MAPS ARE GENERATED FROM ADJACENT SCALES OR A SINGLE SCALE, RESPECTIVELY. THE "STEREO INFO" INDICATES WHETHER THE GUIDANCE MAPS ARE GENERATED WITH STEREO INFORMATION INCLUDING BOTH REFERENCE AND TARGET IMAGE FEATURES, OR ONLY FEATURES OF THE REFERENCE IMAGES

Models	Guidance	Stereo Info	EPE	>3-px	Time (s)
HSM	None	-	1.88	7.51%	0.05
HSM+ours	Single-scale	-	1.88	7.19%	0.09
HSM+ours	Single-scale	✓	1.83	6.40%	0.09
HSM+ours	Inter-scale	-	1.73	6.25%	0.09
HSM+ours	Inter-scale	✓	1.25	4.21%	0.09

exhibits a failure in recovering the objects within the white bounding box, but our method successfully rectifies this error. Additionally, our method corrects the gaps within the blue bounding box at a resolution of 1/16. From a resolution of 1/32 to 1, our method effectively recovers the triangular area within the black bounding box. It is evident that high-resolution cost aggregation is markedly influenced by low-resolution cost aggregation. Our method systematically addresses errors in the original method at each resolution, commencing with the lowest resolution.

F. Generalization Evaluation

1) Universality of Cost Aggregation Method on Different Baseline: We apply our method to five stereo networks, i.e., PSM-Net [6], HSM-Net [7], and CF-Net [30], FastAcv [44] and FastAcvPlus [44] to verify the university of our method. The results on the SceneFlow dataset are shown in Table II, and the results on the KITTI 2015 dataset are shown in Table III.

For PSM-Net, HSM-Net, and CF-Net, our methods have improved by 44.5%, 33.5%, and 32.1% on the SceneFlow dataset, respectively. Moreover, our method has achieved reductions in D-all metrics for all three baselines on the KITTI 2015 dataset. Our method consistently enhances various baselines on both synthetic and real datasets.

2) Zero-Shot Generalization Ability: Obtaining large-scale real-world datasets for training is challenging, making the generalization capability of stereo models crucial. To this end, we evaluate the generalization performance of our methods from synthetic datasets to unseen real-world scenes. In this evaluation, we train various baseline models augmented with our approach on the Scene Flow dataset and directly evaluate them on the Middlebury 2014 and ETH3D training sets. As shown in Table VII, our method consistently outperforms all baselines, demonstrating its strong generalization capability.

G. Comparison With Content-Aware Upsamping Methods

To demonstrate our superiority over conventional content-aware upsampling operators, we directly applied CARAFE++ [59] to the HSM-Net baseline for comparative analysis. The content-aware operators were implemented at resolutions of $\frac{1}{32}$, $\frac{1}{16}$, and $\frac{1}{8}$ of full resolution, aligning

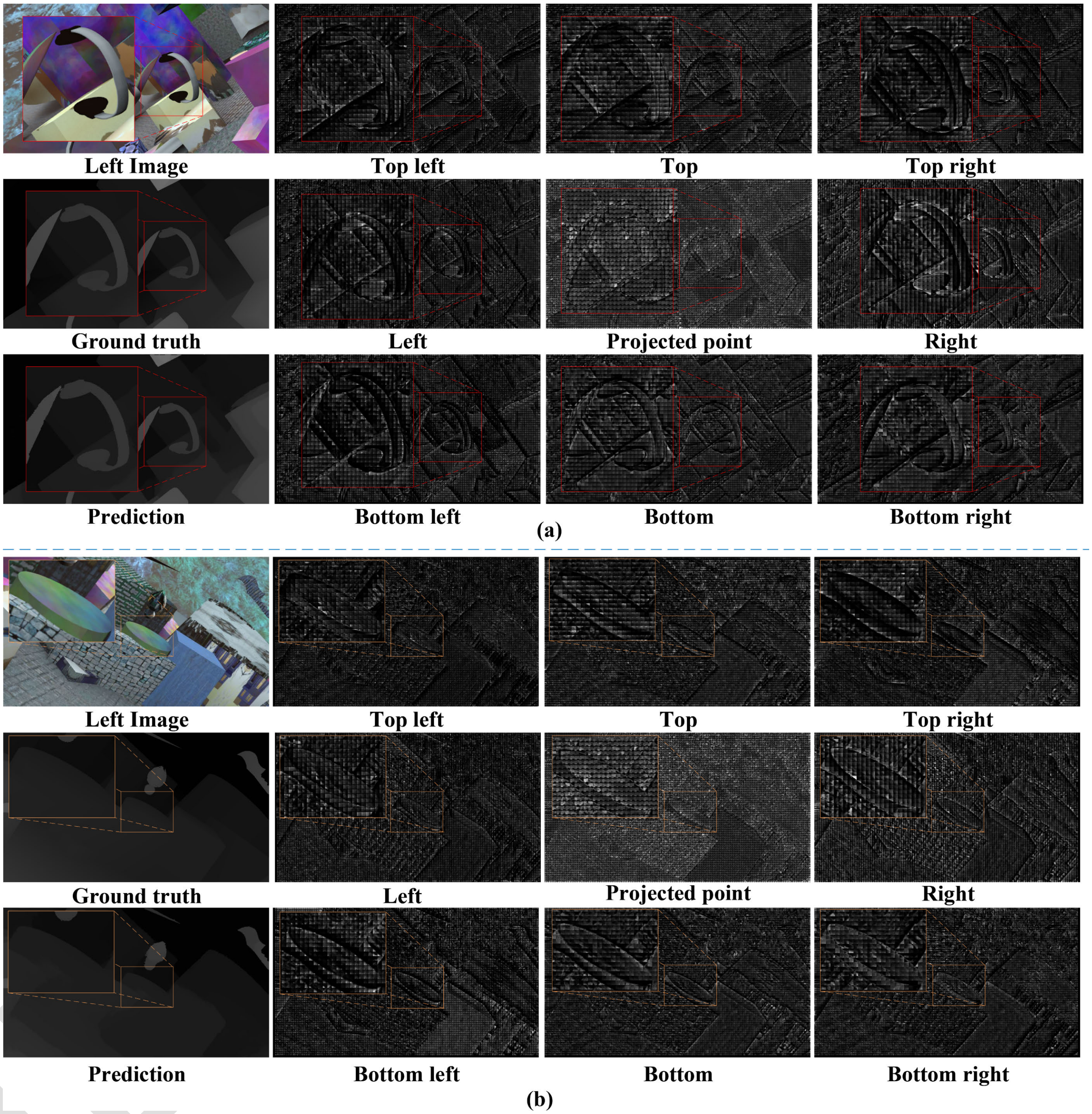


Fig. 9. Results of HSM-Net and HSM-Net+ours at different resolutions. We obtain the disparity map by regressing the cost volume at each resolution.

with the settings of our method. We conducted training and testing of HSM-Net with CARAFE++ on the SceneFlow dataset, using EPE as the measurement metric. The results presented in Table VIII clearly indicate that our method outperforms CARAFE++ in terms of accuracy and speed. Inter-scale information provides us with a broader receptive field for aggregation and access to more content information. Furthermore, our approach involves separating the 3D upsampling process into 1D and 2D upsampling, resulting in a significant reduction in computational cost.

1) Complexity Analysis: To further demonstrate the superiority of our decomposition strategy in computational complexity, we conducted the following analysis and complexity experiments. We separate the 3D upsampling into 1D upsampling plus 2D upsampling, reducing the parameters and calculations. We test the memory cost of different cost aggregation methods of HSM-Net in $\frac{1}{8}$ and 1 of the full resolution (540×960) of the SceneFlow dataset and the results are shown in Table IX. At the same resolution, our module exhibits lower memory and time consumption compared to the other two upsampling methods.

TABLE VII

TABLE IX SYNTHETIC TO REAL GENERALIZATION EXPERIMENTS. ALL MODELS ARE TRAINED ON SCENE FLOW. THE BAD 2.0 ERROR RATE IS USED FOR MIDDLEBURY-V3, AND THE BAD 1.0 ERROR RATE FOR ETH3D

Model	Middlebury H-res	Middlebury Q-res	ETH3D
PSM-Net [6](2018)	15.8	9.8	10.2
GA-Net [45](2019)	13.5	8.5	6.5
HSM-Net [7] (2019)	11.9	7.9	6.7
DSM-Net [73] (2020)	13.8	8.1	6.2
CF-Net [30] (2021)	15.3	9.8	5.8
FC-GANet [39](2022)	10.2	7.8	5.8
FastAcv [44](2023)	12.0	10.6	11.8
FastAcvPlus [44] (2023)	12.4	10.2	11.8
PSM-Net + ours	13.5	7.8	7.1
HSM-Net + ours	**9.8**	**6.2**	5.6
CF-Net + ours	12.3	7.2	**4.6**
FastAcv + ours	11.0	10.1	10.2
FastAcvPlus + ours	10.7	8.9	9.9

TABLE VIII

RESULT OF COMPARISON BETWEEN CARAFE++ [59] AND OURS IN BASELINE HSM-NET [7]. BOTH CARAFE++ AND OURS ONLY REPLACE THE UPSAMPLING MODULE AT RES 1/16 TO 1/8

Experiments	Raw [7]	CARAFE++ [59]	Ours	EPE	Time (s)
Baseline	✓			1.88	0.05
Baseline		✓		1.81	0.15
Baseline			✓	1.54	0.06

TABLE IX

(COMPLEXITY AND EFFICIENCY ANALYSIS OF DIFFERENT COST AGGREGATION STRATEGIES (THE BASELINE MODEL IS HSM-NET). DUE TO HARDWARE LIMITATIONS, WE DO NOT RUN CARAFE++ AT 1/8 TO 1 RESOLUTION. THE BEST RESULTS FOR EACH EVALUATION METRIC ARE SHOWN IN BOLD

Upsampling	Resolution (RES) $\frac{1}{16}$ to $\frac{1}{8}$	$\frac{1}{8}$ to 1	Memory (MB)	Extra Parameter (KB)	Times (s)
HSM + CARAFE++[52]	✓		5153.81	467	0.15
HSM + 3D Deconv	✓		1574.71	216	0.08
HSM + Ours	✓		**1094.86**	**113**	**0.06**
HSM + CARAFE++[52]		-	-	-	-
HSM + 3D Deconv		✓	10936.72	216	0.20
HSM + Ours		✓	**7988.72**	**113**	**0.09**

H. Limitation

1) Lack of Dense Outdoor Data: The performance gains for outdoor scenes are smaller compared to those in virtual and indoor datasets. Additionally, in the CF-Net baseline, our method still fails to completely correct the erroneous depth estimation for the sky, as shown in Fig. 10 (a). We believe there are two main reasons for this: 1) Poor ground truth quality. Outdoor datasets like KITTI use LiDAR scanning, resulting in sparse depth maps. Ground truth is missing in areas beyond the LiDAR scan range, as shown in Fig. 10 (b). This sparsity affects model training. 2) Lack of fine-grained regions. Our method focuses on fine-grained areas, but the coarse nature of LiDAR scans in outdoor datasets means many details are overlooked. For this scenario, we believe that employing some advanced depth completion methods to refine sparse areas in the ground truth could be a reasonable approach.

2) Future Work: In future work, we aim to delve into super-resolution techniques to augment the detail information

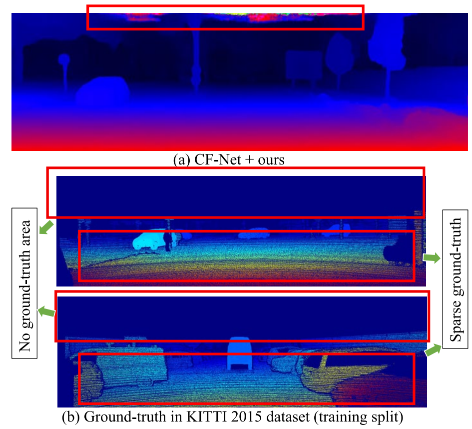


Fig. 10. Failure case and ground-truth in the outdoor scenarios.

within extensive textureless areas, which will significantly bolster the performance in outdoor environments. Furthermore, the present study has adopted a distinct spatial domain modeling strategy to address the issue of detail loss. Yet, the utilization of high-frequency components in the frequency domain for such fine-grained information presents itself as an inherently viable alternative. Moving forward, we intend to experiment with frequency domain analysis techniques, including wavelet transformations, to facilitate the restoration of fine-grained regional information.

VI. CONCLUSION

We have presented an inter-scale similarity guided cost aggregation method designed to adaptively recover details in fine-grained areas. By leveraging both low-resolution and high-resolution information, our approach effectively exploits detail while generating inter-scale similarity measurements. Additionally, our stereo-content-aware cost aggregation method employs a decomposition strategy that divides the 3D disparity-spatial space into 1D disparity space and 2D spatial space, significantly reducing computational costs associated with 3D cost volumes. Experimental results across three benchmarks demonstrate the effectiveness of our method with various models.

REFERENCES

[1] R. Chabra, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, “StereoDRNet: Dilated residual StereoNet,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11778–11787.

[2] Q. Chen et al., “Virtual blood vessels in complex background using stereo X-ray images,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 99–106.

[3] T. Pire, T. Fischer, J. Civera, P. De Cristóforis, and J. J. Berlles, “Stereo parallel tracking and mapping for robot localization,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 1373–1378.

[4] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, “DeepDriving: Learning affordance for direct perception in autonomous driving,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2722–2730.

[5] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

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

[7] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5510–5519.

[8] 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.

[9] 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.

[10] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12981–12990.

AQ:2 [11] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A context integrated residual pyramid network for stereo matching," in *Proc. Asian Conf. Comput. Vis.* Springer, 2018, pp. 20–35.

[12] 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.

[13] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3268–3277.

[14] V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "HITNet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14362–14372.

[15] H. Hirschmüller, P. R. Innocent, and J. Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 229–246, 2002.

[16] K.-J. Yoon and I. So Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.

[17] C. Xu, C. Wu, D. Qu, F. Xu, H. Sun, and J. Song, "Accurate and efficient stereo matching by log-angle and pyramid-tree," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4007–4019, Oct. 2021.

[18] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, vol. 2, Jul. 2001, pp. 508–515.

[19] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[20] 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.

[21] X. Zhang, H. Dai, H. Sun, and N. Zheng, "Algorithm and VLSI architecture co-design on efficient semi-global stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4390–4403, Nov. 2020.

[22] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A deep visual correspondence embedding model for stereo matching costs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 972–980.

[23] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proc. BMVC*, 2016, vol. 2, no. 3, p. 4.

[24] J. L. Schonberger, S. N. Sinha, and M. Pollefeys, "Learning to fuse proposals from multiple scanline optimizations in semi-global matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 739–755.

[25] J. Žbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1592–1599.

[26] A. Seki and M. Pollefeys, "SGM-Nets: Semi-global matching with neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 231–240.

[27] K. Batsos, C. Cai, and P. Mordohai, "CBMV: A coalesced bidirectional matching volume for disparity estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2060–2069.

[28] Y. Zhang et al., "Adaptive unimodal cost volume filtering for deep stereo matching," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12926–12934.

[29] 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.

[30] 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.

[31] 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.

[32] H. Dai, X. Zhang, Y. Zhao, H. Sun, and N. Zheng, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3099–3110, May 2022.

[33] T. Song, S. Kim, and K. Sohn, "Unsupervised deep asymmetric stereo matching with spatially-adaptive self-similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13672–13680.

[34] J. Jing et al., "Uncertainty guided adaptive warping for robust and efficient stereo matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, vol. 29, Oct. 2023, pp. 3295–3304.

[35] J. Zeng, C. Yao, L. Yu, Y. Wu, and Y. Jia, "Parameterized cost volume for stereo matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18347–18357.

[36] C. Yao and L. Yu, "FoggyStereo: Stereo matching with fog volume representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13033–13042.

[37] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Depth-aware multi-grid deep homography estimation with contextual correlation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4460–4472, Jul. 2022.

[38] J. Cheng, X. Yang, Y. Pu, and P. Guo, "Region separable stereo matching," *IEEE Trans. Multimedia (TMM)*, vol. 26, pp. 4880–4893, 2022.

[39] 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.

[40] X. Ye, J. Zhang, Y. Yuan, R. Xu, Z. Wang, and H. Li, "Underwater depth estimation via stereo adaptation networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 5089–5101, Sep. 2023.

[41] 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.

[42] Q. Su and S. Ji, "ChiTransformer: Towards reliable stereo from cues," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1929–1939.

[43] H. Zhao, H. Zhou, Y. Zhang, J. Chen, Y. Yang, and Y. Zhao, "High-frequency stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1327–1336.

[44] G. Xu, Y. Wang, J. Cheng, J. Tang, and X. Yang, "Accurate and efficient stereo matching via attention concatenation volume," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2461–2474, Apr. 2024.

[45] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 185–194.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] O. Rukundo and H. Cao, "Nearest neighbor value interpolation," 2012, *arXiv:1211.1768*.

[51] M. Mastyło, "Bilinear interpolation theorems and applications," *J. Funct. Anal.*, vol. 265, no. 2, pp. 185–207, Jul. 2013.

[52] D. A. Rajon and W. E. Bolch, "Marching cube algorithm: Review and trilinear interpolation adaptation for image-based dosimetric models," *Computerized Med. Imag. Graph.*, vol. 27, no. 5, pp. 411–435, Sep. 2003.

[53] R. E. Carlson and F. N. Fritsch, "Monotone piecewise bicubic interpolation," *SIAM J. Numer. Anal.*, vol. 22, no. 2, pp. 386–400, Apr. 1985.

[54] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[55] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[56] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 269–284.

[57] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, 2016, pp. 1–9.

[58] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3007–3016.

[59] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE++: Unified content-aware reassembly of features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4674–4687, Sep. 2022.

[60] D. Mazzini, "Guided upsampling network for real-time semantic segmentation," 2018, *arXiv:1807.07466*.

[61] 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.

[62] D. Scharstein et al., "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.

[63] T. Schöps, T. Sattler, and M. Pollefeys, "BAD SLAM: Bundle adjusted direct RGB-D SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 134–144.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[65] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[66] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[67] Q. Zhang, X. Zhang, B. Li, Y. Chen, and A. Ming, "HDA-net: Horizontal deformable attention network for stereo matching," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 32–40.

[68] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "DeepPruner: Learning efficient stereo matching via differentiable PatchMatch," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4384–4393.

[69] Q. Wang, S. Shi, S. Zheng, K. Zhao, and X. Chu, "FADNet++: Real-time and accurate disparity estimation with configurable networks," 2021, *arXiv:2110.02582*.

[70] A. F. Kadmin, R. A. Hamzah, M. N. A. Manap, M. S. Hamid, and T. F. T. Wook, "Local stereo matching algorithm using modified dynamic cost computation," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 22, no. 3, pp. 1312–1319, Jun. 2021.

[71] M. N. Z. Azali, R. A. Hamzah, and Z. M. Noh, "Disparity map algorithm using census transform and hierarchical segment-tree from stereo image," in *Proc. Eng. Technol. Int. Conf. (ETIC)*, Sep. 2022, pp. 244–249.

[72] M. Zahari, R. A. Hamzah, N. A. Manap, and A. I. Herman, "Stereo matching algorithm for autonomous vehicle navigation using integrated matching cost and non-local aggregation," *Bull. Electr. Eng. Informat.*, vol. 12, no. 1, pp. 328–337, Feb. 2023.

[73] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, and P. Torr, "Domain-invariant stereo matching networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 420–439.

Pengxiang Li received the B.S. degree from the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China, in 2021, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and machine learning.

Chengtang Yao received the B.S. and M.S. degrees in computer science from Beijing Institute of Technology (BIT), China, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include computer vision, pattern recognition, and machine learning.

Yunde Jia (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Beijing Institute of Technology (BIT), China, in 1983, 1986, and 2000, respectively. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University (CMU), from 1995 to 1997. He is currently a Professor with the Department of Engineering, Shenzhen MSU-BIT University (SMBU). He is also the Director of Guangdong Provincial Key Laboratory of Machine Perception and Intelligent Computing. His research interests include computer vision, vision-based HCI and HRI, and intelligent robotics.

Yuwei Wu (Member, IEEE) received the Ph.D. degree in computer science from Beijing Institute of Technology (BIT), Beijing, China, in 2014. From August 2014 to August 2016, he was a Post-Doctoral Research Fellow with the Rapid-Rich Object Search (ROSE) Laboratory, School of Electrical and Electronic Engineering (EEE), Nanyang Technological University (NTU), Singapore. He is currently a tenured Associate Professor with the School of Computer Science, BIT. He has strong research interests include computer vision and machine learning.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.

AQ:1 = Please confirm or add details for any funding or financial support for the research of this article.

AQ:2 = Please provide the publisher location for Ref. [11].

Inter-Scale Similarity Guided Cost Aggregation for Stereo Matching

Pengxiang Li, Chengtang Yao, Yunde Jia, *Member, IEEE*, and Yuwei Wu, *Member, IEEE*

***Abstract*— Stereo matching aims to estimate 3D geometry by computing disparity from a rectified image pair. Most deep learning based stereo matching methods aggregate multi-scale cost volumes computed by downsampling and achieve good performance. However, their effectiveness in fine-grained areas is limited by significant detail loss during downsampling and the use of fixed weights in upsampling. In this paper, we propose an inter-scale similarity-guided cost aggregation method that dynamically upsamples the cost volumes according to the content of images for stereo matching. The method consists of two modules: inter-scale similarity measurement and stereo-content-aware cost aggregation. Specifically, we use inter-scale similarity measurement to generate similarity guidance from feature maps in adjacent scales. The guidance, generated from both reference and target images, is then used to aggregate the cost volumes from low-resolution to high-resolution via stereo-content-aware cost aggregation. We further split the 3D aggregation into 1D disparity and 2D spatial aggregation to reduce the computational cost. Experimental results on various benchmarks (e.g., SceneFlow, KITTI, Middlebury and ETH3D-two-view) show that our method achieves consistent performance gain on multiple models (e.g., PSM-Net, HSM-Net, CF-Net, FastAcv, and FactAcvPlus). The code can be found at https://github.com/Pengxiang-Li/issga-stereo.**

***Index Terms*— Stereo matching, cost aggregation, content-aware upsampling.**

I. Introduction

STEREO matching aims to estimate a pixel-wise disparity map from a rectified image pair. It plays an important role in various applications including 3D reconstruction [1], AR [2], SLAM [3], and autonomous driving [4]. The well-known pipeline divides stereo matching into four steps: cost computation, cost aggregation, disparity computation, and disparity refinement [5]. Among these four steps, cost aggregation plays a pivotal role in leveraging neighborhood information to rectify the ambiguous matching costs in ill-posed regions such as occluded regions, large textureless areas, repetitive patterns, and thin structures. The cost aggregation is commonly embedded into end-to-end deep neural networks with multi-scale processing to enlarge the receptive field. 3D CNNs [6], [7], GRU [8], [9], and attention mechanism [10] are the most commonly used basic structures for cost aggregation, effectively correcting ambiguous matching costs and substantially enhancing prediction accuracy in ill-posed regions by aggregating multi-scale cost volumes.

However, these cost aggregation methods often struggle in fine-grained areas due to considerable detail loss during downsampling and fixed weight used in upsampling. Many efforts have been targeted at improving the performance of stereo matching in fine-grained areas, including edge information [11], deformable convolutions [12], group-wise correlation [13] and slanted planes [14]. These methods have achieved good performance, but two challenging problems in cost aggregation are still not well solved: (1) the downsampling causes considerable detail loss during the construction of multi-scale cost volumes, and (2) the upsampling fixed in size and weight is prone to data imbalance between large-smooth areas and fine-grained areas. For example, the HSM-Net [7] with multi-scale cost volumes and upsampling fixed in size and weight may lead to poor performance in fine-grained areas, as illustrated in Fig. 1 (b) and Fig. 1 (c).

In this paper, we propose inter-scale similarity-guided cost aggregation that adaptively restores image details by dynamically upsampling cost volumes based on image content. Our method comprises two modules: inter-scale similarity measurement and stereo-content-aware cost aggregation. We utilize inter-scale similarity measurements to generate similarity guidance from the feature maps of adjacent scales. Subsequently, we employ this guidance to aggregate the multi-scale cost volumes through stereo-content-aware cost aggregation.

For the first challenging problem, our idea is to retrieve the fine-grained details lost during the downsampling process. We use inter-scale similarity measurement to measure the similarity between high-resolution and low-resolution features. The similarity explicitly preserves the connection between high-resolution details and low-resolution features,

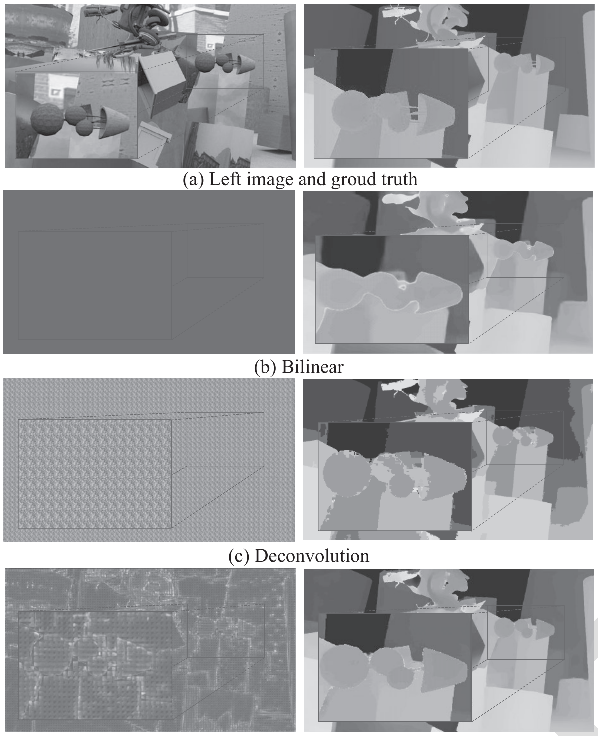
AQ:1 Manuscript received 21 November 2023; revised 17 July 2024; accepted 22 August 2024. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 62176021 and Grant 62172041, in part by the Natural Science Foundation of Shenzhen under Grant JCYJ20230807142703006, and in part by the Key Research Platforms and Projects of Guangdong Provincial Department of Education under Grant 2023ZDZX1034. This article was recommended by Associate Editor X. Sun. *(Pengxiang Li and Chengtang Yao are co-first authors.) (Corresponding authors: Yuwei Wu; Yunde Jia.)*

Pengxiang Li, Chengtang Yao, and Yuwei Wu are with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: pengxiangli@bit.edu.cn; yao.c.t@bit.edu.cn; ~~weyuwei@bit.edu.cn),~~

Yunde Jia is with Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China, and also with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology (BIT), Beijing 100081, China (e-mail: jiayunde@smbu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3453965.

Digital Object Identifier 10.1109/TCSVT.2024.3453965



(a) Left image and groud truth

(b) Bilinear

(c) Deconvolution

(d) Ours

Fig. 1. Predictions and upsampling weights visualizations of HSM-Net [7] using different upsampling strategies.

thereby providing guidance for the upsampling process to restore details. Technically, we first project a point $\boldsymbol{x}_{high}$ from high-resolution to low-resolution features $\boldsymbol{x}_{low}$. Then we compute the similarity between the point in the high-resolution $\boldsymbol{x}_{high}$ and the neighbors of the projected point in low-resolution $\mathcal{N}_{low}$.

For the second challenging problem, a critical factor leading to the suboptimal restoration of fine-grained details is the fixed size and weights of existing upsampling strategies, which are unable to adapt to the complicated fine-grained details. Motivated by this, we replace the fixed upsampling with content-aware upsampling. The content-aware upsampling uses the content information of each point to guide the upsampling process, thereby mitigating the impact of data imbalance between large-smooth and fine-grained areas. In stereo-content-aware cost aggregation, we use similarity guidance (generated from both reference and target images) to guide the aggregation of matching costs in 3D spatial-disparity space. The pair-wise 3D upsampling is computationally expensive. Thus, we split the upsampling in the 3D space into 1D disparity and 2D spatial space. As a result, our method is able to efficiently and adaptively assemble the proper neighbors for cost aggregation and upsampling. Our method generates upsampling weight according to the image content and achieves much finer details, as shown in Fig. 1 (d). Our method can be plugged into any multi-scale cost volume based stereo network and achieve higher accuracy, especially in fine-grained areas.

Our contributions are summarized as follows:

1) We propose an inter-scale similarity guided cost aggregation method to adaptively recover the details of cost volumes under the guidance of similarity generated from images.
2) We introduce an inter-scale similarity measurement to dynamically generate guidance by incorporating information from both low-resolution and high-resolution feature maps. The explicit utilization of high-resolution feature maps ensures the preservation of fine-grained details.
3) We design a decomposition strategy that splits 3D disparity-spatial upsampling into 1D disparity and 2D spatial upsampling, significantly reducing the computational cost of the 3D pair-wise upsampling.

II. Related Work

A. Stereo Matching

Traditional stereo matching methods estimate disparity maps for rectified image pairs using local [15], [16], global [17], [18], and semi-global methods [19], [20], [21]. Deep learning-based stereo matching networks now dominate, delivering state-of-the-art results. Early deep learning methods replaced steps in stereo matching [5]: cost computation [22], [23], [24], cost aggregation [25], [26], [27], disparity computation [28], and disparity refinement [26], [27]. Despite good performance, their non-end-to-end approaches limited data utilization. To overcome this, end-to-end methods compute correlations by warping the target image to the reference image [8], [9], [11], [29]. These achieve excellent results but often lose geometric information. Cost-volume-based models [6], [30], [31], [32], [33], [34], [35] preserve geometric information by concatenating multi-scale cost volumes. State-of-the-art methods use convolution neural networks [36], [37], [38], [39], [40], [41] or attention mechanisms [10], [42], [43], [44] to aggregate these volumes, effectively utilizing image context information.

However, multi-scale cost volume-based stereo matching methods often lose fine-grained details due to downsampling. While cost aggregation usually recovers these details, current fixed-size and fixed-weight schemes struggle with data imbalances between large smooth, and fine-grained areas. To address this, we developed a content-aware cost aggregation method that mitigates detail loss during multi-scale cost volume creation. Our adaptive upsampling approach also remains robust against data imbalances.

B. Cost Aggregation

Multi-scale cost aggregation methods [6], [29] enhance matching cost reliability by optimizing multi-scale cost volumes for precise disparity estimation. Song et al. [11] used edge information to guide cost aggregation, reducing edge mismatches. Zhang et al. [45] improved efficiency by replacing 3D CNNs with semi-global aggregation. Yang et al. [7] proposed a hierarchical feature volume decoder for high-resolution image disparity estimation. Xu et al. [12] utilized deformable convolution for adaptive aggregation. Lipson et al. [8] designed an iterative mixed disparity sampling and aggregation strategy. Liu et al. [46] used local features to address over-smoothing.

Zhang et al. [47] introduced depth-based sampling for balanced density in close and far regions. Xu et al. [48] utilized bilateral grid processing for faster aggregation. Lee et al. [49] introduced a cluster-wise cost aggregation algorithm to parallelized scanline-level disparity computation.

The aforementioned methods demonstrate commendable performance, even in ill-posed areas. However, they still suffer from the loss of details in downsampling, and their strategies for multi-scale cost aggregation are susceptible to data imbalance. These strategies commonly rely on either bilinear interpolation or deconvolution for upsampling. Both bilinear interpolation and deconvolution employ a fixed interpolation rule or deconvolution kernel across all data points, thus failing to exploit the content information of images fully. Constrained by computational memory limitations, these methods are unable to perform direct aggregation at full resolution. Instead, they resort to upsampling to full resolution without introducing additional parameters after aggregating at 1/2 or 1/4 resolution. However, relying solely on parameter-free upsampling is inadequate for recovering lost details.

C. Upsampling

Upsampling is used to transform data from low-resolution to high-resolution. Traditional upsampling strategies fit a curve of a small neighborhood of the upsampled points to compute values for interpolated points, including nearest neighbor interpolation [50], bilinear interpolation [51], trilinear interpolation [52], and bicubic interpolation [53], etc. The advantage of these methods lies in their low computational cost. However, these parameter-free upsampling strategies underutilize image content, resulting in blurred recovery results in fine-grained areas. Deconvolution [54], [55], [56], [57] offers a learning-based approach to upsampling, where weights are optimized through backpropagation. Learning-based upsampling kernels enable the utilization of contextual information learned from extensive data. However, deconvolution has limitations as it struggles in various scenes due to fixed kernel sizes and weights, making it susceptible to data imbalances.

Several works [58], [59], [60] use content-aware upsampling operators to solve the fixed-weight problem. Wang et al. [58], [59] presented a content-aware reassembly approach and argued that traditional feature upsampling methods struggle to capture rich semantic information. While content-aware upsampling mitigates the fixed-weight problem, it relies solely on information from the low-resolution side (i.e., the upsampling process could be regarded as a unary low-resolution to high-resolution mapping). However, the upsampling process inherently consists of both low-resolution and high-resolution components, and relying solely on low-resolution features for upsampling may not suffice. Instead of employing a unary upsampling mapping, we introduce an inter-scale similarity measurement approach to produce a pair-wise upsampling mapping, represented by similarity guidance derived from information gathered across adjacent scales. In other words, we actually model the upsampling process as a binary mapping between low-resolution and high-resolution.

III. Optimization in Multi-Scale Cost Aggregation

In this section, we model the optimization objectives for each layer of multi-scale cost aggregation. Given a cost volume $\boldsymbol{C}_{l-1} \in \boldsymbol{R}^{H_{l-1}\times W_{l-1}\times D_{l-1}}$ at level $l-1$ as input, $\boldsymbol{C}_l$ is computed via a network with learning weights $\boldsymbol{W}_l$. The generation of $\boldsymbol{C}_l$ can be formulated as

$$\begin{aligned}\mathrm{p}(\boldsymbol{C}_l) &= \mathrm{p}(\boldsymbol{C}_l|\boldsymbol{C}_{l-1}, \boldsymbol{W}_l)\mathrm{p}(\boldsymbol{W}_l)\mathrm{p}(\boldsymbol{C}_{l-1})\\ &= \mathrm{p}(\boldsymbol{C}_l|\boldsymbol{C}_{l-1}, \boldsymbol{W}_l)\mathrm{p}(\boldsymbol{W}_l). \qquad (1)\end{aligned}$$

The probability $\mathrm{p}(\boldsymbol{C}_l)$ of cost volume is commonly computed by $\mathrm{p}(\boldsymbol{C}_l) = \mathrm{softmax}(-\boldsymbol{C}_l)$, and $p(\boldsymbol{C}_{l-1})$ is supposed to be 1 as $\boldsymbol{C}_{l-1}$ has already been given. Then, the optimization objective is to find the best $\boldsymbol{W}_l$ that recovers the details lost in $\boldsymbol{C}_{l-1}$, which can be formulated as

$$\begin{aligned}\boldsymbol{W}_l &= \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \mathrm{p}(\boldsymbol{W}_l|\boldsymbol{C}_l, \boldsymbol{C}_{l-1}),\\ &= \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \mathrm{p}(\boldsymbol{W}_l|\boldsymbol{C}_l),\\ &= \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \frac{\mathrm{p}(\boldsymbol{C}_l|\boldsymbol{W}_l)\cdot \mathrm{p}(\boldsymbol{W}_l)}{\sum_{\boldsymbol{W}_l}\mathrm{p}(\boldsymbol{C}_l|\boldsymbol{W}_l)\mathrm{p}(\boldsymbol{W}_l')\mathrm{d}\boldsymbol{W}_l'},\\ &\overset{a.s.}{=} \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \mathrm{p}(\boldsymbol{C}_l|\boldsymbol{W}_l)\cdot\mathrm{p}(\boldsymbol{W}_l),\\ &\overset{a.s.}{=} \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\,\mathrm{p}(\boldsymbol{C}_l). \qquad (2)\end{aligned}$$

In the aforementioned cost aggregation process, it becomes impractical to recover the details lost during downsampling using bilinear upsampling or deconvolution. This is because $\boldsymbol{W}_l$ is optimized by cost volumes at level $[0, 1, \dots, l-1]$, and it doesn't consider the image content at level l. In other words, only minimal details at level l contribute to the optimization of $\boldsymbol{W}_l$. Furthermore, the kernel weights are influenced by the content that appears more frequently in the image. Consequently, it becomes challenging to utilize these fixed kernel weights effectively for recovering details that constitute only a small proportion of the image content such as the fine-grained areas.

IV. Proposed Method

A. Problem Formulation

Detail loss and biased upsampling are two challenging problems that cause poor performance in fine-grained areas. To address these two problems, we optimize cost aggregation with image features at levels l and $l-1$. In our method, the optimization objective of cost aggregation at each level is given by

$$\boldsymbol{W}_l = \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\,\mathrm{p}(\boldsymbol{C}_l)\cdot\mathrm{p}(\boldsymbol{W}_l|\boldsymbol{F}_l, \boldsymbol{F}_{l-1}), \qquad (3)$$

where $\boldsymbol{F}_l$ is the feature map at level l.

In particular, the optimization objective of cost aggregation with deconvolution is actually one special case of ours, where $\mathrm{p}(\boldsymbol{W}_l|\boldsymbol{F}_l, \boldsymbol{F}_{l-1}) = \mathrm{p}(\boldsymbol{W}_l)$. Besides, the optimization objective of cost aggregation with bilinear interpolation is one special case of deconvolution, i.e., Eq. (2). With substituting $\mathrm{p}(\boldsymbol{W}_l) = 1$ into Eq. (2), Eq. (2) can be reformulated as

$$\boldsymbol{W}_l = \underset{\boldsymbol{W}_l}{\mathrm{argmax}}\ \mathrm{p}(\boldsymbol{C}_l), \qquad (4)$$

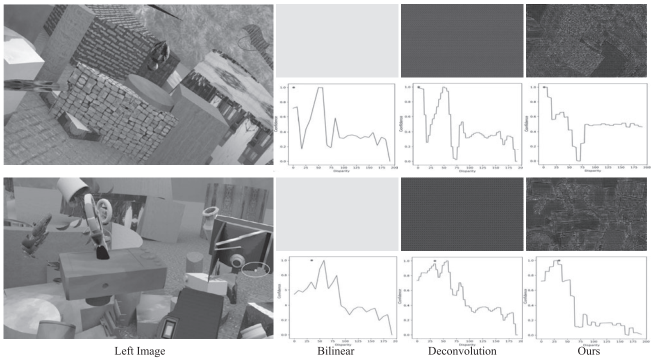


Fig. 2. The visualization of aggregation weight and disparity distribution in cost volume. The upper row shows the aggregation weight and the under row shows the distribution of the cost volume along the disparity dimension for a single point. The point in each distribution map is the ground truth for the point in the reference image. Both the bilinear upsampling and deconvolution predict wrong results, while ours not only predicts the correct disparity but also corrects for multi-modality in the distribution.

which is just the optimization objective of cost aggregation with bilinear interpolation.

In our method, $\boldsymbol{W}_l$ is automatically adjusted from the change of $\boldsymbol{F}_l$ and $\boldsymbol{F}_{l-1}$ during inference, whereas the weights of deconvolution or bilinear interpolation remain static. Our method generates aggregation weight related to the image content and achieves unimodal distribution results, while others get multimodal distribution or wrong distribution. Fig. 2 provides a visual representation of aggregation using various upsampling strategies. As shown in Fig. 2, the weights for bilinear interpolation remain constant, the weights for deconvolution are repetitive kernels, while our method's weights are content-aware, closely linked to the image's content. It's also worth noting that our method effectively addresses the issue of multiple peaks in the disparity distribution (see the distribution curves of Deconvolutions vs. Ours in Fig. 2). In our method, the disparity distribution exhibits only a single prominent peak precisely at the ground truth disparity, whereas deconvolution may exhibit multiple peaks, potentially leading to incorrect disparity results.

B. Implementation

Given an image pair, we extract multi-scale feature maps $\boldsymbol{F}_l$ at each level l for reference and target images. We then use the feature maps to construct the cost volume at the lowest level. As for the cost volume at the high level, we iteratively upsample the cost volume from the low level to the high level through two steps, the inter-scale similarity measurement, and the stereo-content-aware cost aggregation. The inter-scale similarity measurement uses feature maps from adjacent scales to generate similarity guidance. The stereo-content-aware cost aggregation uses the similarity guidance from two views to guide the cost volume upsampling. At last, we use the cost volume at the highest level to compute the disparity map as the output of our network. Fig. 3 illustrates the pipeline of our method.

1) Inter-Scale Similarity Measurement: The inter-scale similarity measurement takes the feature maps $\boldsymbol{F}_l$ and $\boldsymbol{F}_{l-1}$ as input. We compute the similarity by the summation of the products of $\boldsymbol{F}_l(h', w')$ and the neighbors of $\boldsymbol{F}_{l-1}(h, w)$ with the formula as

$$\boldsymbol{S}_l(h', w') = \frac{1}{M \cdot M}\phi\Big(\sum_{(h,w)\in\mathcal{N}_F} \boldsymbol{F}_l(h', w')\boldsymbol{F}_{l-1}(h, w)\Big), \quad (5)$$

where (h', w') and (h, w) are the location at high-level and low-level respectively, $(h', w') = (h \cdot s, w \cdot s)$, s is the scale change in resolution from level $l-1$ to level l, and $\cdot$ is the scalar multiplication operation. $\boldsymbol{S}_l \in \mathbb{R}^{H^l \times W^l}$ is the similarity guidance at level l, $\boldsymbol{S}_l(h', w')$ is the value of the pixel at location (h', w'), $\mathcal{N}_F \in \mathbb{R}^{M \times M}$ is a 2D neighborhood of the pixel at location (h, w) with the size of $M \times M$. $\phi(\cdot)$ is a subnetwork composed of convolution layers, relu layers, and batch normalization layers.

2) Stereo-Content-Aware Cost Aggregation: 3D convolution based methods [6], [7] usually perform window based cost aggregation:

$$\boldsymbol{C}_l(h', w', d') = \sum_{(h,w,d)\in\mathcal{N}_c} \boldsymbol{W}_l(h', w', d')\boldsymbol{C}_{l-1}(h, w, d), \quad (6)$$

where $\mathcal{N}_c$ is a 3D neighborhood of the point at $(h'/s, w'/s, d'/s)$.

In our method, we replace the 3D weight $\boldsymbol{W}_l$ with the 2D similarity guidance $\boldsymbol{S}_l$. For each level, we use the feature maps of the stereo images, i.e., reference and target images, to compute the content-aware similarity guidance $\boldsymbol{S}_l^R$ and $\boldsymbol{S}_l^T$ by inter-scale similarity measurement, respectively. Then we perform the cost aggregation guided by $\boldsymbol{S}_l^R$ and $\boldsymbol{S}_l^T$:

$$\boldsymbol{C}_l(h', w', d') = \sum_{(h,w,d)\in\mathcal{N}_c} \boldsymbol{S}_l^R(h', w')\boldsymbol{S}_l^T(h', w' - d') \boldsymbol{C}_{l-1}(h, w, d). \quad (7)$$

The memory and computational cost of 3D cost aggregation are unaffordable. Accordingly, we introduce a decomposition strategy to reduce the computation cost. We split the upsampling in full 3D spatial-disparity space into 1D disparity and 2D spatial upsampling by leveraging the property of cost volume on the disparity dimension. The property is that position (h, w, d) in cost volume represents the (h, w) in the reference image and $(h, w-d)$ in the target image. We warp $\boldsymbol{S}_l^T$ to $\boldsymbol{S}_l^R$, and then split the mapping of cost volume into 1D disparity dimension and 2D spatial dimension. Specifically, we replace Eq. (7) with a two-step decomposed cost aggregation.

In the first step, 1D disparity upsampling, the positions $(h, w, d - \lfloor M/2 \rfloor), \ldots, (h, w, d), \ldots, (h, w, d + \lfloor M/2 \rfloor)$ in cost volume along disparity dimension correspond to (h, w) in the reference image and $(h, w - d + \lfloor M/2 \rfloor), \ldots, (h, w - d), \ldots, (h, w - d - \lfloor M/2 \rfloor)$ in the target image. Formally, the updating along the disparity dimension is given by

$$\boldsymbol{C}_l(h, w, d') = \sum_{d\in\mathcal{N}_d} \boldsymbol{S}_l^R(h', w')\boldsymbol{S}_l^T(h', w' - d')\boldsymbol{C}_{l-1}(h, w, d), \quad (8)$$

where $\mathcal{N}_d = \{d'/s - \lfloor M/2 \rfloor, \ldots, d'/s, \ldots, d'/s + \lfloor M/2 \rfloor\}$. In the second step, 2D spatial upsampling, all voxels with location $(h', w', :)$ in cost volume correspond to the pixel with

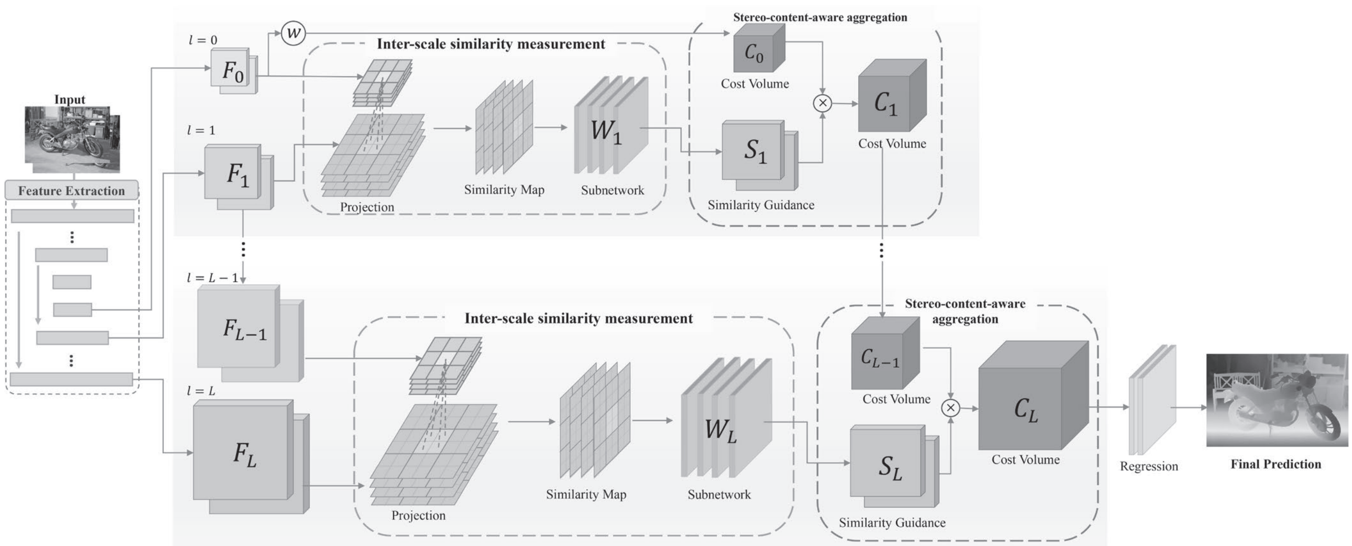


Fig. 3. The overall architecture. Given an image pair, our method extracts multi-scale features at each level l. L is the total number of levels. Ⓦ is the warping operation used for generating the initial cost volume in the lowest resolution. ⊗ represents element-wise multiplication.

location (h', w') in the reference image. The update along the spatial dimension is given by

$$\boldsymbol{C}_l(h', w', d') = \sum_{(h,w)\in\mathcal{N}_{sp}} \boldsymbol{S}_l^R(h', w')\boldsymbol{C}_l(h, w, d'), \tag{9}$$

where $\mathcal{N}_{sp} \in \mathbb{R}^{\mathbb{M}\times\mathbb{M}}$ is a 2D neighborhood of the pixel with location $(h'/s, w'/s)$ at level $l-1$.

After all these operations, we complete the transformation from the shape of $H_{l-1} \times W_{l-1} \times D_{l-1}$ to $H_{l-1} \times W_{l-1} \times D_l$ and then to $H_l \times W_l \times D_l$, where $H_l = H_{l-1} \cdot s$, $W_l = W_{l-1} \cdot s$ and $D_l = D_{l-1} \cdot s$.

3) Loss Function: We use a multi-scale loss function that applies smooth L_1 loss to each level. The smooth L_1 loss function is not sensitive to outliers or noises. The loss function is defined as

$$\boldsymbol{D}_l = \sum_{d\in\{d_n\}_{n=1}^N} d \cdot \sigma(-\boldsymbol{C}_l), \tag{10}$$

$$\mathcal{L} = \sum_{l=0} \lambda_l \cdot \mathcal{L}_l(\boldsymbol{D}_l - \boldsymbol{G}_l), \tag{11}$$

$$\mathcal{L}_l(x) = \begin{cases} 0.5\ x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}, \tag{12}$$

where $\{d_n^l\}_{n=1}^N$ is the disparity hypothesis at level l, $\sigma(\cdot)$ is the softmax operation, $\boldsymbol{D}_l$ is the predicted disparity map at level l, λ_l denotes the coefficients for the disparity prediction at level l, and $\boldsymbol{G}_l$ is the ground-truth disparity map at level l.

C. Computational Cost Analysis

To further demonstrate the superiority of our decomposition strategy in computational complexity, we conducted the following analysis and complexity experiments (in Section V). We separate the 3D upsampling into 1D upsampling plus 2D upsampling, reducing the parameters and calculations.

1) Parameters: For deconvolution, the number of parameters per layer is given by $C \times 1 \times k^3 = Ck^3$, where k is the kernel size, C is the number of input channels, and the output channel is set to 1. In contrast, our method requires Ck^2 parameters per layer. Both our method and deconvolution utilize the same number of layers.

2) Calculations: For the computational complexity of 3D upsampling, comparing 3D deconvolution with our method for a feature volume of size $C \times D \times H \times W$ to be upsampled by a scale of s, the computational cost for deconvolution is $\mathcal{O}(s^3k^3)CDHW$, while ours is $\mathcal{O}(s^2k^2 + sk)CDHW = \mathcal{O}(s^2k^2)CDHW$.

V. Experiments

A. Datasets

1) SceneFlow Dataset: SceneFlow [29] is a large synthetic dataset containing 34896 training images and 4248 testing images with the size of 540×960. This dataset has three rendered sub-datasets: FlyingThings3D, Monkaa, and Driving. FlyingThings3D is rendered from the ShapeNet dataset and has 21828 training data and 4248 testing data. Monkaa is rendered from the animated film Monkaa and has 8666 training data. The Driving is constructed by the naturalistic, dynamic street scene from the viewpoint of a driving car and has 4402 training samples.

2) KITTI 2015 Dataset: KITTI 2015 [61] is a real-world dataset with street views from a driving car. It contains 200 training stereo image pairs with sparse ground-truth disparities obtained using LiDAR and another 200 testing image pairs without ground-truth disparities. During the training process, we take 160 images for training and reference 40 images for validation.

3) Middlebury-v3 Dataset: Middlebury-v3 is a subset of the 2014 dataset [62] and is collected in the real world with static indoor scenes containing complicated and rich details. There are 15 stereo pairs for training and 15 stereo pairs for testing.

Each pair is provided in 3 kinds of resolution, full, half, and quarter resolution; where we used the quarter resolution in the experiment.

4) ETH3D-Two-Iew Dataset: ETH3D (two view) [63] comprises grayscale stereo pairs captured from diverse indoor and outdoor scenes. The dataset includes 27 training and 20 testing image pairs with sparsely labeled ground truth. Disparities range from 0 to 64 pixels, and bad 1.0 (percentage of pixels with errors larger than 1 pixel) are reported.

B. Evaluation Metrics

The end-point error (EPE) is the mean disparity error in pixels. The 3-px error refers to the proportion of points in the full map where the absolute value of the difference between the predicted disparity and the true value is greater than 3 pixels. The percentage of disparity outliers in the background (*D1-bg*), foreground (*D1-fg*), or all pixels (*D1-all*) for both noc regions and all regions are applied for evaluation. Disparity outliers are the pixels if their disparity EPE is more than 3 or 5% pixels. *avgerr* is the average absolute error in pixels. *RMS* represents the root mean square of the differences between the estimated and actual disparities. *A90* and *A95* are the 90% and 95% error quantile in pixels, respectively. *Bad 1.0* and *Bad 2.0* percentage of pixels with errors larger than 1 pixel or 2 pixels, respectively.

C. Model Details

To prove the effectiveness of our method, we extend five stereo baseline networks with our method, CF-Net [30], HSM-Net [7] and PSM-Net [6], FastAcv [44] and FastAcvPlus [44]. All networks are implemented via PyTorch and tested on NVIDIA RTX 3090 GPU. For all baselines, the neighborhood size M is set to the scale change s at each level.

For PSM-Net+ours, the model is optimized using Adam [64] with β_1 of 0.9, β_2 of 0.999. During training, the batch size is fixed to 8, and we perform color normalization to each input image and crop them into 256×512 resolution. We train our network on SceneFlow for 10 epochs and change the learning rate from 0.001 to 0.0001 in the 7th epoch. We then fine-tune the network on KITTI 2015 and set the learning rate to 0.001, 0.0001, and 0.00003 in the first 200 epochs, the next 400 epochs, and the final 600 epochs, respectively. As for Middlebury-v3, we also fine-tune the model pre-trained on SceneFlow. The learning rate is set to 0.001 for 300 epochs and then changed to 0.0001 for the rest of 600 epochs. For HSM-Net+ours, we use AdamW [65] with β_1 of 0.9, β_2 of 0.999. During training, the batch size is fixed to 12, and we perform the same data augmentation [7] of the original HSM-Net and crop the images into 256×512 resolution. We train our network for 10 epochs using the same dataset as HSM-Net and change the learning rate from 0.001 to 0.0001 in the 9th epoch. For CF-Net, FastAcv, and FastAcvPlus, we follow all the training strategies of the official repositories.

Furthermore, we downsample the ground truth for our multi-scale loss. We use bilinear downsampling in SceneFlow and nearest downsampling in KITTI 2015 and Middlebury-v3.

TABLE I

EVALUATION RESULTS OF CURRENT STEREO MATCHING ALGORITHMS ON THE SCENEFLOW TEST SET

Model	PSM-Net [6] (2018)	Gwc-Net [13] (2019)	HSM-Net [7] (2019)
EPE(px)	1.09	0.98	2.07
Model	Acf-Net (2020)	LEAStereo (2020)	CF-Net (2021)
EPE(px)	0.87	0.78	0.99
Model	LaC+ GwcNet (2022)	LaC+ GANet (2022)	FastAcv (2023)
EPE(px)	0.75	0.72	0.64
Model	FastAcvPlus (2023)	PSM-Net+ours	HSM-Net+ours
EPE(px)	0.59	0.63	1.39
Model	CF-Net+ours	FastAcv + ours	FastAcvPlus + ours
EPE(px)	0.72	0.59	**0.57**

Additionally, we reduced the computational cost without sacrificing accuracy by moving the averaging operation before the aggregation at each layer. Although we observe better results of bilinear downsampling in the experiment on SceneFlow, the ground truth disparities of the two real-world datasets contain invalid values, like 0 and INF, which will lead to wrong disparity results after bilinear downsampling. In all experiments, no post-processing or unsupervised learning methods are used.

D. Comparison With Stereo Matching Methods

Our method mainly focuses on recovering the fine-grained details lost during cost volume downsampling. Therefore, we conduct experiments on the SceneFlow dataset, specifically targeting fine-grained areas, and we compare the results against mainstream baseline methods. Additionally, we perform experiments on real datasets, including KITTI and Middlebury, to validate the effectiveness of our approach.

1) SceneFlow: The experimental results in Table I show that our proposed method significantly improves the performance of stereo matching algorithms, with the FastAcvPlus+ours achieving the lowest EPE of 0.57. The consistent reduction in EPE across various models demonstrates the robustness and efficacy of our method.

a) Fine-grained areas: We test different baselines in the fine-grained region on the SceneFlow dataset to verify the accuracy improvement of our method in the fine-grained (FG) areas and full areas, as shown in Table II. We use the calculated HOG [66] descriptor of the reference image as a mask of fine-grained areas. The results in Table II show the superiority of our method in fine-grained areas. Our method can improve the accuracy significantly in fine-grained areas, and 37.6%, 32.9%, 16.2%, 11.4% and 10.4% EPE reduction in PSM-Net, HSM-Net, CF-Net, FastAcv and FastAcvPlus, respectively. Our method is effective for different baselines with good universality. Our method also brings no or small increase in runtime. For PSM-Net, we remove the time-consuming 3D convolution layers in the hourglass modules at RES 1/16 and RES 1/8. For the rest baselines, we directly plug our method into them without additional model modification.

The visualization results for the fine-grained regions are depicted in Fig. 4. Our method successfully recovers more details, notably improving estimation results for fine-grained areas like the spokes of the wheel and plant spikes in the left column of Fig. 4 compared to the baseline. Furthermore, our approach enhances results in less refined regions, such as inside the bounding box in the right column of Fig. 4.

TABLE II

RESULTS OF DIFFERENT BASELINE IN FULL AREAS (FULL) AND FINE-GRAINED (FG) AREAS ON THE SCENEFLOW DATASET. FOR PSM-NET, WE REMOVE THE TIME-CONSUMING 3D CNNS AT RES 1/16 AND RES 1/8

Method	EPE in Full	EPE in FG	Time (s)
PSM-Net [6] (2018)	1.09	1.01	0.41
PSM-Net+ours	0.60	0.63	0.37
HSM-Net [7] (2019)	1.88	2.07	0.05
HSM-Net+ours	1.25	1.39	0.09
CF-Net [30] (2021)	1.06	0.99	0.18
CF-Net+ours	0.72	0.83	0.22
FastAcv [44] (2023)	0.64	0.70	0.05
FastAcv+ours	0.59	0.62	0.08
FastAcvPlus [44] (2023)	0.59	0.67	0.05
FastAcvPlus+ours	0.57	0.60	0.08

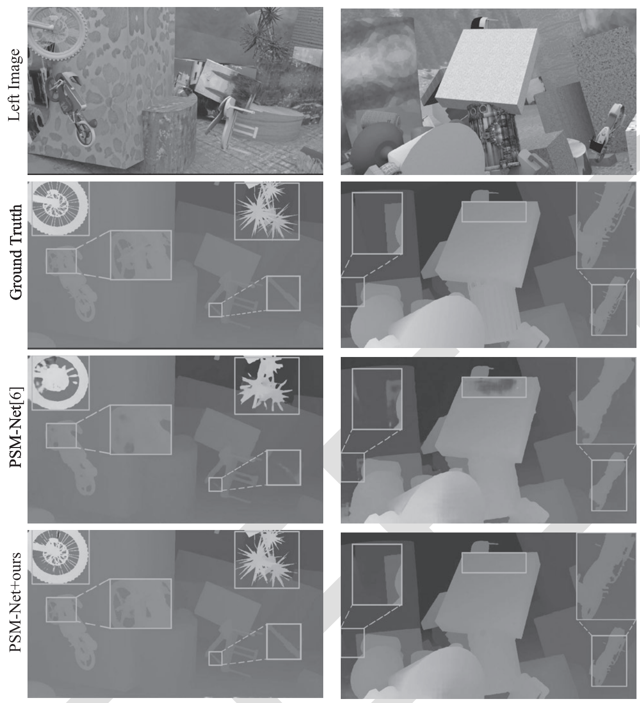


Fig. 4. The results of PSM-Net [6] and PSM-Net+ours on the SceneFlow dataset.

b) Full areas: Our method brings improvement for full areas across different baselines. Visualization of the results (Fig. 5) reveals that our method exhibits certain corrective effects on large-scale weakly-textured regions as well. The experimental results demonstrate that our approach achieves significant accuracy improvement when applied to datasets with complete depth information as ground truth.

2) Kitti: Table III displays the performance and runtime of various algorithms evaluated on the KITTI stereo2015 benchmark. Across different baselines, our method enhances the accuracy of the original baselines with only a marginal increase in processing time. Notably, CF-Net combined with our method surpasses other competing methods in the Noc D1-all and All D1-all. Next, we would like to provide a more detailed explanation of the comparison between our method and stereo matching methods based on attention mechanisms, as well as methods based on decomposition strategies.

TABLE III

EVALUATION ON KITTI 2015 BENCHMARK. THE BEST RESULTS FOR EACH EVALUATION METRIC ARE SHOWN IN BOLD

Models	Noc (%)			All (%)			Time (s)
	bg	fg	all	bg	fg	all	
PSM-Net [6] (2018)	1.71	4.31	2.14	1.86	4.62	2.32	0.41
GwcNet-g [13] (2019)	1.61	3.49	1.92	1.74	3.93	2.11	0.32
HSM-Net [7] (2019)	1.63	3.40	1.92	1.80	3.85	2.14	0.05
RAFT-Stereo [8] (2021)	-	-	-	2.89	**1.75**	1.96	-
HDA-Net [67] (2021)	1.55	3.32	1.84	1.69	3.76	2.03	0.42
BG-Net [48] (2021)	1.91	4.34	2.31	2.07	4.74	2.51	0.02
Dec-Net [31] (2021)	1.89	3.53	2.16	2.07	3.87	2.37	0.05
CF-Net [30] (2021)	**1.43**	3.25	1.73	**1.54**	3.56	1.88	0.18
ChiT-12 [42] (2022)	2.11	3.79	2.38	2.34	4.05	2.60	-
FC-PSMNet [39](2022)	1.73	4.19	2.13	1.86	4.61	2.32	-
HTSGM [49] (2022)	-	-	5.66	-	-	5.84	-
FastACV-Plus [44] (2023)	-	-	-	1.70	3.53	2.01	0.05
PSM-Net+ours	1.50	3.42	1.82	1.73	3.78	2.07	0.38
HSM-Net+ours	1.64	3.21	1.90	1.77	3.60	2.09	0.10
CF-Net+ours	1.46	**2.95**	**1.70**	1.58	3.30	**1.87**	0.22

a) Compared with the attention-based method: HDA-Net [67] proposes an efficient horizontal attention module to adaptively capture the global correspondence clues. Our method uses inter-scale information to generate similarity guidance to improve cost aggregation. As shown in Table III, our method has lower D1-all (HDA-Net 2.03 *vs.* CF-Net+Ours 1.87) with faster running time (HDA-Net 0.42ms *vs.* CF-Net+Ours 0.22ms) on the KITTI 2015 dataset.

b) Compared with the decomposition method: DecNet [31] decomposes the original stereo matching into a dense matching at the lowest resolution and a series of sparse matching at higher resolutions. Unlike DecNet, our method decomposes the 3D upsampling of cost volume into a 2D-spatial and 1D-disparity upsampling. Our method outperforms DecNet in D1-all (Dec-Net 2.37 *vs.* HSM-Net+Ours 2.09) but is slower in runtime (Dec-Net 0.05ms *vs.* HSM-Net+Ours 0.09ms), as shown in Table III.

c) Visualization: Fig. 6 presents the experimental results on the KITTI 2015 dataset, showcasing images from top to bottom. Our method excels in recovering slender structures, as seen in the iron chain at the center of the first row and the fence in the lower left corner of the third row. Moreover, our approach accurately estimates depth-mutation areas such as signboards and utility poles. For instance, unlike PSM-Net and HSM-Net in the first row's bounding box around the signboard, our method produces correct results. In rows two, five, and six, the other methods misidentify parts of the background as utility poles, which our method avoids.

3) Middlebury: We compare our method with several approaches using different aggregation strategies on the Middlebury stereo dataset v3, as shown in Table IV. We outperform these 3D aggregation based approaches on most of the metrics. The result also demonstrates the effectiveness of our content-aware upsampling method. Based on the visualizations in Fig. 7, we can draw the following conclusions: **1) Improved depth estimation for fine-grained regions**: Our method shows superior performance in depth estimation for fine-grained regions, demonstrating the effectiveness of explicitly integrating high-resolution and low-resolution information. This is evident in almost all cases, such as the detailed areas

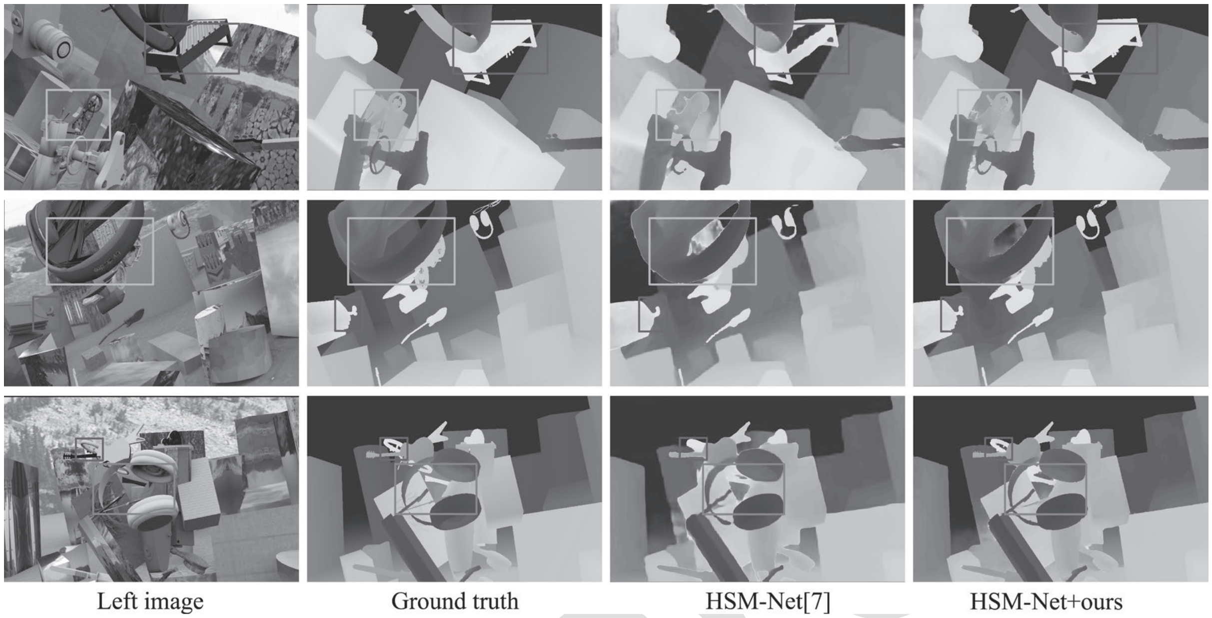


Fig. 5. The visualization of results on HSM-Net baseline. The first and second columns are the reference input images and ground truth. The rest columns are results from HSM-Net and HSM-Net+ours.

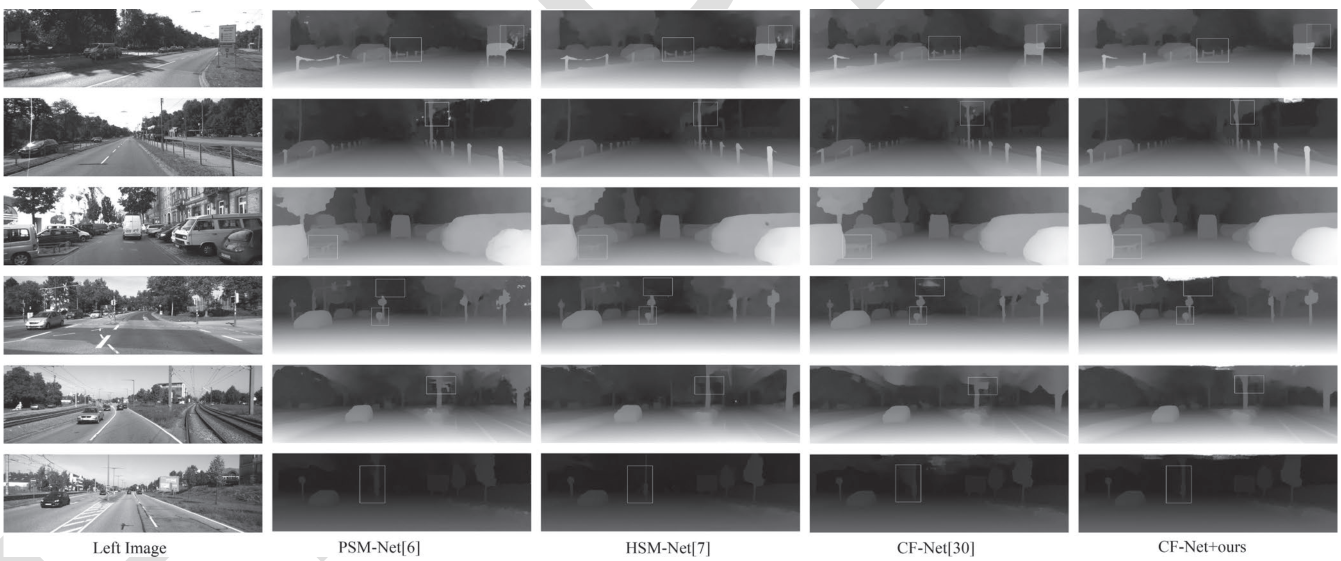


Fig. 6. The results of different deep stereo networks on KITTI 2015 dataset. Our method performs better in fine-grained areas than other methods, especially for the region denoted with the boxes. Please zoom in to check the details.

in "DjembL" and the water cup on the table in "Crusade" (Line 3, PSM-Net vs. PSM-Net + ours), as well as the small figurine (Line 3, HSM-Net vs. HSM-Net + ours). **2) Enhanced foreground and background decoupling**: Our method has a stronger ability to decouple the foreground from the background. Retaining low-resolution information effectively enhances this capability. Examples include the depth estimation of the potted plants and background in "Plants" (Line 5) and the estimation of the hollow part of the staircase handrail in "Staircase" (Line 4, HSM-Net vs. HSM-Net + ours; CF-Net vs. CF-Net + ours). **3) Competitive performance in flat regions**: Our method also shows competitive performance in flat regions. For instance, the wall in the upper left of "Staircase" (Line 4, PSM-Net vs. PSM-Net + ours) and the restoration of the table corner in "Crusade" (Line 3, CF-Net vs. CF-Net + ours). However, our method has some shortcomings in certain areas, such as the seats in the PSM-Net case of "Classroom2E" (Line 1, PSM-Net vs. PSM-Net + ours). We will systematically discuss these limitations in the Limitation Analysis section.

E. Ablation Studies

We conduct all the analysis in ablation studies mainly on the HSM-Net baseline. Ablation studies are performed on the SceneFlow dataset and the KITTI 2015 dataset.

1) Effectiveness of Stereo-Content-Aware Cost Aggregation: During Stereo-Content-Aware Cost Aggregation, we use both

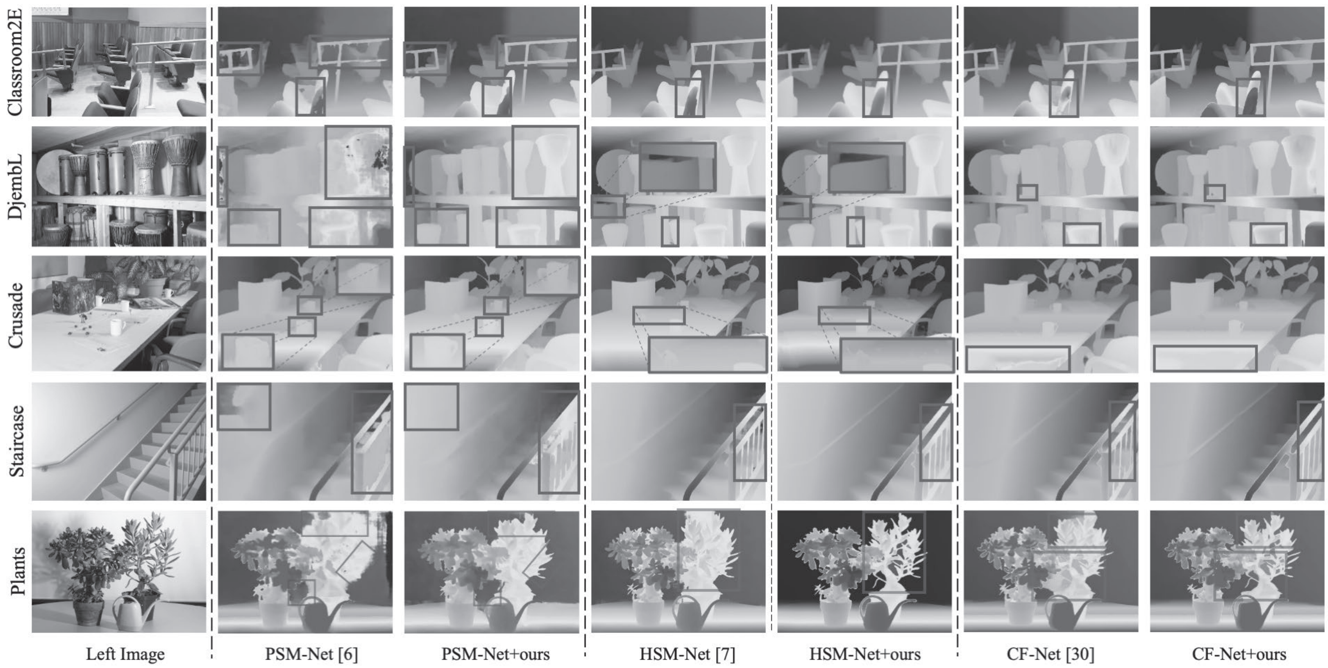


Fig. 7. The visualization of results on Middlebury-v3 test set. In the first column are the reference input images. The rest columns are results from PSM-Net [6], PSM-Net + ours, HSM-Net [7], + ours, CF-Net [30] and CF-Net + ours respectively.

TABLE IV
EVALUATION ON MIDDLEBURY-V3. THE BEST RESULTS FOR EACH EVALUATION METRIC ARE SHOWN IN BOLD

Models	Res	Avgerr	Rms	A90	A95
PSM-Net_ROB (2018) [6]	Q	8.78	23.3	22.8	43.4
DeepPruner (2019) [68]	Q	6.56	18	17.9	33.1
FADNet++ (2021) [69]	Q	11.9	27.7	34.3	61.2
MCP-HA-VQ (2022) [70]	Q	6.01	37.5	40.6	85.9
H-CENST (2022) [71]	Q	10.2	29.1	24.3	59.0
FM-DT (2023) [72]	Q	11.7	31.4	33.4	67.1
PSM-Net+ours	Q	**5.43**	**17.3**	**8.11**	**25.2**

reference and target images to extract similarity guidance and separate the 3D spatial-disparity upsampling into 1D disparity / 2D spatial upsampling. We evaluate the effectiveness of our method at different resolutions through two experiments: i. Training on the SceneFlow dataset and testing on the SceneFlow dataset. ii. Training on the SceneFlow dataset and testing on the validation set of the KITTI 2015 dataset. Table V demonstrates that our decomposition strategy reduces the running time by nearly half compared to full 3D upsampling at the setting of "RES 1/16 to RES 1/8" and "RES 1/8 to RES 1" on the SceneFlow dataset and KITTI 2015 dataset. Our decomposition strategy not only proves to be faster but also more accurate than full 3D upsampling. When integrating our method at "RES 1/16 to 1/8," HSM-Net+ours experiences a decrease in EPE of 18.09% and 15.86% compared to the original HSM-Net on the SceneFlow dataset and the KITTI 2015 dataset, respectively. Plugging our method at higher resolutions, i.e., "RES 1/8 to 1", the EPE of HSM-Net+ours is 33.51% and 26.21% lower than the original HSM-Net on the SceneFlow dataset and the KITTI 2015 dataset, respectively. Our method is effective and the higher the resolution at which we employ our method, the greater the improvement it brings.

2) Effectiveness of Inter-Scale Similarity Measurement: We utilize inter-scale similarity measurement to generate a similarity guidance map for cost aggregation. Each pixel in the similarity map corresponds to the content information at the same location. Our method calculates the similarity between high-resolution feature points and their corresponding $M \times M$ points in the low-resolution counterpart. Visualizations of similarity maps of a 3 size neighborhood are shown in Fig. 8.

We confirm the effectiveness of our inter-scale policy on the SceneFlow dataset through a series of experiments. These experiments are conducted in three settings: without similarity guidance, with single-scale similarity guidance, and with inter-scale similarity guidance. The results presented in Table VI clearly demonstrate that the use of inter-scale similarity guidance results in higher accuracy when compared to single-scale similarity guidance. The inter-scale similarity guidance transforms the unary mapping inherent in single-scale similarity guidance into a pair-wise mapping, consequently leading to improved accuracy. Furthermore, we verify the significance of employing stereo information, which includes both reference and target images, to achieve favorable results. In Table VI, it is evident that the EPE when using stereo information is significantly lower than when not using stereo information. Utilizing stereo information to model the mapping relationship between cost volumes of different resolutions proves to be more reliable than relying solely on reference images.

3) Effectiveness of Our Method in Different Resolution: We further provide visualizations of the results obtained from HSM-Net and HSM-Net+ours at different resolutions on the SceneFlow dataset. These visualizations help us understand how our model enhances the baseline at various resolutions, as shown in Fig. 9. At a resolution of 1/32, HSM-Net

TABLE V

RESULTS OF USING THE GUIDANCE IN MULTIPLE STEPS OF MULTI-SCALE COST AGGREGATION ON SYNTHETIC AND REAL DATASETS. RES 1/16, 1/8, 1 REPRESENTS THE ORIGINAL IMAGE'S 1/16, 1/8, AND 1 RESOLUTION. RES 1/16 TO 1/8 INDICATES WHETHER THE BASELINE IS PLUGGED WITH OUR METHOD IN COST AGGREGATION FROM RESOLUTION 1/16 TO RESOLUTION 1/8, SO AS RES 1/8 TO 1

Models	Inter-scale		3D upsampling		SceneFlow		KITTI 15	
	RES 1/16 to 1/8	RES 1/8 to 1	Full 3D	2D + 1D	EPE	Times (s)	EPE	Times (s)
HSM	-	-	-	-	1.88	0.05	1.45	0.05
HSM+ours	✓	-	✓	-	1.81	0.17	1.37	0.18
HSM+ours	✓	-	-	✓	1.54	0.06	1.22	0.08
HSM+ours	✓	✓	✓	-	1.67	0.27	1.32	0.24
HSM+ours	✓	✓	-	✓	1.25	0.09	1.07	0.10

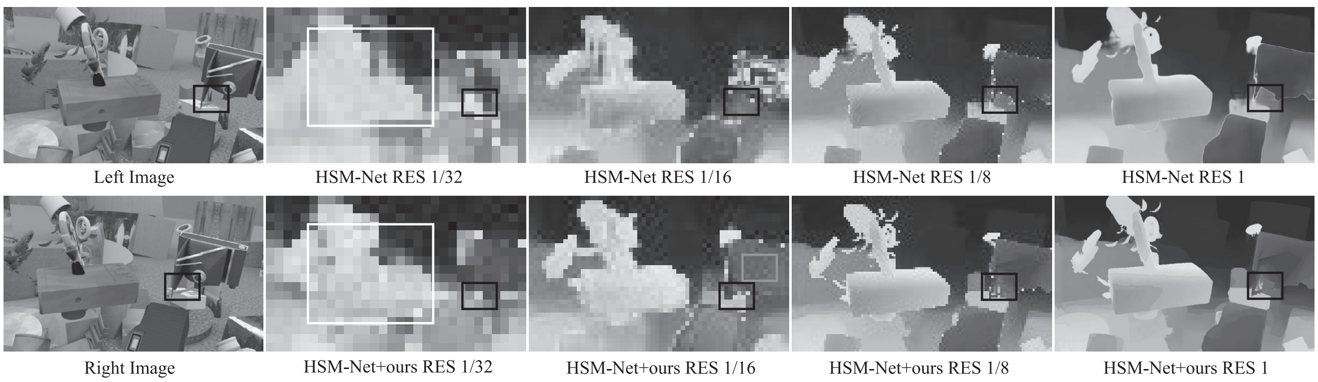


Fig. 8. The visualization of similarity. (a) and (b) are the similarity of the two images. The three columns on the right are visualizations of the similarity, representing the similarity of points in high resolution to their corresponding neighbors in low-resolution projection points. In each map, the brightness indicates the similarity, which corresponds to the upsampling kernel weight. It can be seen that the aggregation weight is directly related to the image content and that each weight in the global picture uniquely adapts the content information of the corresponding points.

TABLE VI

THE RESULTS OF USING DIFFERENT SCALES OF GUIDANCE TO GUIDE COST AGGREGATION. "INTER-SCALE" AND "SINGLE-SCALE" REPRESENT THAT THE GUIDANCE MAPS ARE GENERATED FROM ADJACENT SCALES OR A SINGLE SCALE, RESPECTIVELY. THE "STEREO INFO" INDICATES WHETHER THE GUIDANCE MAPS ARE GENERATED WITH STEREO INFORMATION INCLUDING BOTH REFERENCE AND TARGET IMAGE FEATURES, OR ONLY FEATURES OF THE REFERENCE IMAGES

Models	Guidance	Stereo Info	EPE	>3-px	Time (s)
HSM	None	-	1.88	7.51%	0.05
HSM+ours	Single-scale	-	1.88	7.19%	0.09
HSM+ours	Single-scale	✓	1.83	6.40%	0.09
HSM+ours	Inter-scale	-	1.73	6.25%	0.09
HSM+ours	Inter-scale	✓	1.25	4.21%	0.09

exhibits a failure in recovering the objects within the white bounding box, but our method successfully rectifies this error. Additionally, our method corrects the gaps within the blue bounding box at a resolution of 1/16. From a resolution of 1/32 to 1, our method effectively recovers the triangular area within the black bounding box. It is evident that high-resolution cost aggregation is markedly influenced by low-resolution cost aggregation. Our method systematically addresses errors in the original method at each resolution, commencing with the lowest resolution.

F. Generalization Evaluation

1) Universality of Cost Aggregation Method on Different Baseline: We apply our method to five stereo networks, i.e., PSM-Net [6], HSM-Net [7], and CF-Net [30], FastAcv [44] and FastAcvPlus [44] to verify the university of our method. The results on the SceneFlow dataset are shown in Table II, and the results on the KITTI 2015 dataset are shown in Table III.

For PSM-Net, HSM-Net, and CF-Net, our methods have improved by 44.5%, 33.5%, and 32.1% on the SceneFlow dataset, respectively. Moreover, our method has achieved reductions in D-all metrics for all three baselines on the KITTI 2015 dataset. Our method consistently enhances various baselines on both synthetic and real datasets.

2) Zero-Shot Generalization Ability: Obtaining large-scale real-world datasets for training is challenging, making the generalization capability of stereo models crucial. To this end, we evaluate the generalization performance of our methods from synthetic datasets to unseen real-world scenes. In this evaluation, we train various baseline models augmented with our approach on the Scene Flow dataset and directly evaluate them on the Middlebury 2014 and ETH3D training sets. As shown in Table VII, our method consistently outperforms all baselines, demonstrating its strong generalization capability.

G. Comparison With Content-Aware Upsamping Methods

To demonstrate our superiority over conventional content-aware upsampling operators, we directly applied CARAFE++ [59] to the HSM-Net baseline for comparative analysis. The content-aware operators were implemented at resolutions of $\frac{1}{32}$, $\frac{1}{16}$, and $\frac{1}{8}$ of full resolution, aligning

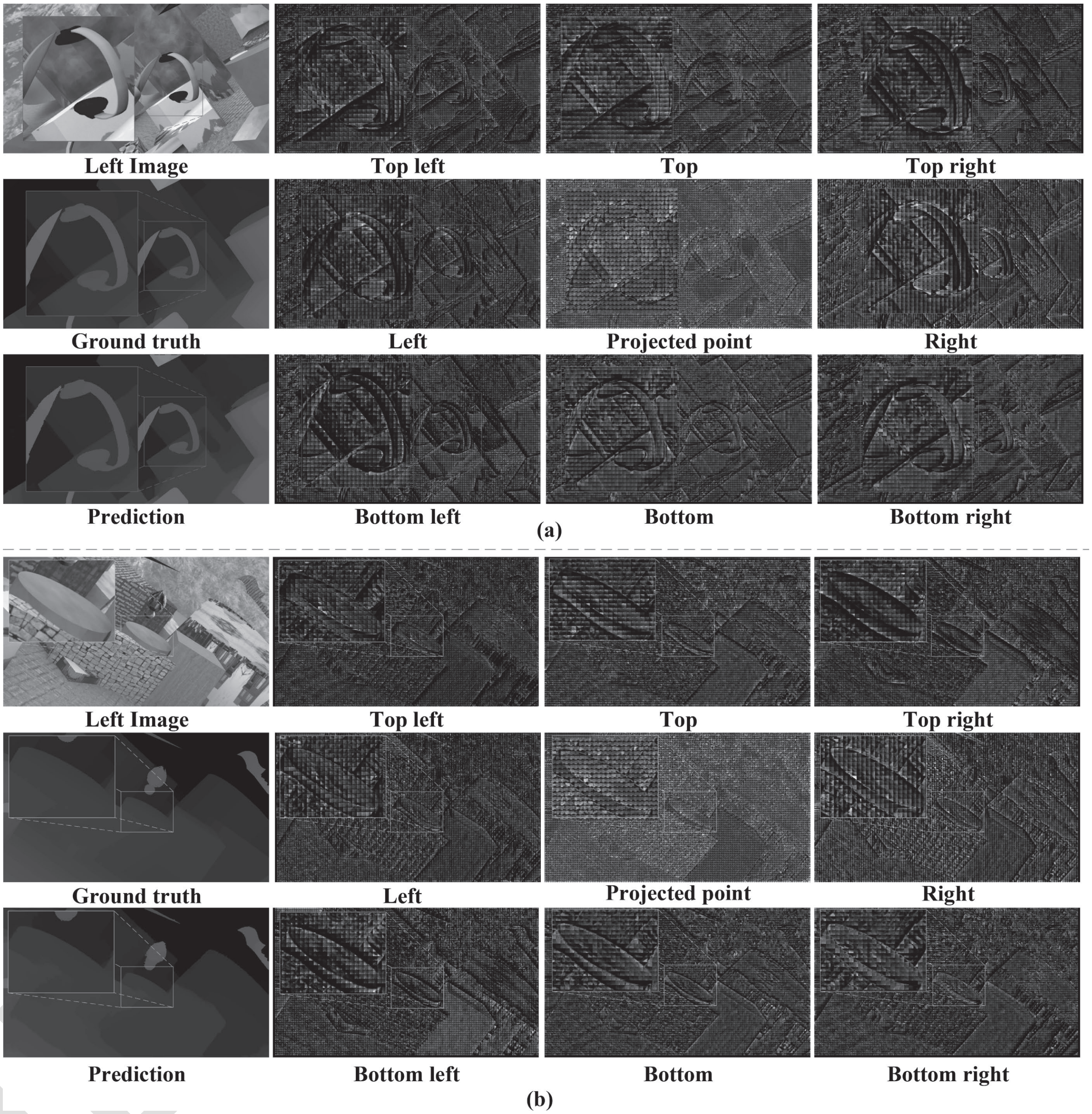


Fig. 9. Results of HSM-Net and HSM-Net+ours at different resolutions. We obtain the disparity map by regressing the cost volume at each resolution.

with the settings of our method. We conducted training and testing of HSM-Net with CARAFE++ on the SceneFlow dataset, using EPE as the measurement metric. The results presented in Table VIII clearly indicate that our method outperforms CARAFE++ in terms of accuracy and speed. Inter-scale information provides us with a broader receptive field for aggregation and access to more content information. Furthermore, our approach involves separating the 3D upsampling process into 1D and 2D upsampling, resulting in a significant reduction in computational cost.

1) Complexity Analysis: To further demonstrate the superiority of our decomposition strategy in computational complexity, we conducted the following analysis and complexity experiments. We separate the 3D upsampling into 1D upsampling plus 2D upsampling, reducing the parameters and calculations. We test the memory cost of different cost aggregation methods of HSM-Net in $\frac{1}{8}$ and 1 of the full resolution (540×960) of the SceneFlow dataset and the results are shown in Table IX. At the same resolution, our module exhibits lower memory and time consumption compared to the other two upsampling methods.

TABLE VII

TABLE IX Synthetic to Real Generalization Experiments. All Models Are Trained on Scene Flow. The Bad 2.0 Error Rate Is Used for Middlebury-v3, and the Bad 1.0 Error Rate for ETH3D

Model	Middlebury		ETH3D
	H-res	Q-res	
PSM-Net [6](2018)	15.8	9.8	10.2
GA-Net [45](2019)	13.5	8.5	6.5
HSM-Net [7] (2019)	11.9	7.9	6.7
DSM-Net [73] (2020)	13.8	8.1	6.2
CF-Net [30] (2021)	15.3	9.8	5.8
FC-GANet [39](2022)	10.2	7.8	5.8
FastAcv [44](2023)	12.0	10.6	11.8
FastAcvPlus [44] (2023)	12.4	10.2	11.8
PSM-Net + ours	13.5	7.8	7.1
HSM-Net + ours	**9.8**	**6.2**	5.6
CF-Net + ours	12.3	7.2	**4.6**
FastAcv + ours	11.0	10.1	10.2
FastAcvPlus + ours	10.7	8.9	9.9

TABLE VIII

Result of Comparison Between CARAFE++ [59] and Ours in Baseline HSM-Net [7]. Both CARAFE++ and Ours Only Replace the Upsampling Module at RES 1/16 to 1/8

Experiments	Raw [7]	CARAFE++ [59]	Ours	EPE	Time (s)
Baseline	✓			1.88	0.05
Baseline		✓		1.81	0.15
Baseline			✓	1.54	0.06

TABLE IX

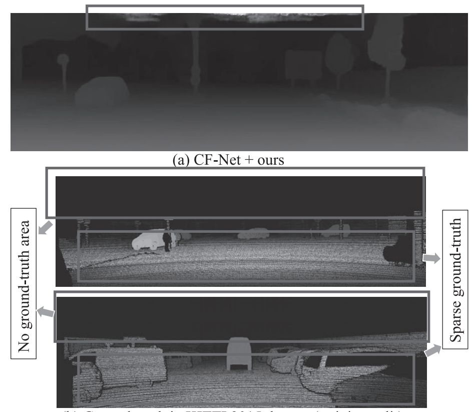
(Complexity and Efficiency Analysis of Different Cost Aggregation Strategies (The Baseline Model Is HSM-Net). Due to Hardware Limitations, We Do Not Run CARAFE++ at 1/8 to 1 Resolution. The Best Results for Each Evaluation Metric Are Shown in Bold

Upsampling	Resolution (RES) $\frac{1}{16}$ to $\frac{1}{8}$	$\frac{1}{8}$ to 1	Memory (MB)	Extra Parameter (KB)	Times (s)
HSM + CARAFE++[52]	✓		5153.81	467	0.15
HSM + 3D Deconv	✓		1574.71	216	0.08
HSM + Ours	✓		**1094.86**	**113**	**0.06**
HSM + CARAFE++[52]		-	-	-	-
HSM + 3D Deconv		✓	10936.72	216	0.20
HSM + Ours		✓	**7988.72**	**113**	**0.09**

H. Limitation

1) Lack of Dense Outdoor Data: The performance gains for outdoor scenes are smaller compared to those in virtual and indoor datasets. Additionally, in the CF-Net baseline, our method still fails to completely correct the erroneous depth estimation for the sky, as shown in Fig. 10 (a). We believe there are two main reasons for this: 1) Poor ground truth quality. Outdoor datasets like KITTI use LiDAR scanning, resulting in sparse depth maps. Ground truth is missing in areas beyond the LiDAR scan range, as shown in Fig. 10 (b). This sparsity affects model training. 2) Lack of fine-grained regions. Our method focuses on fine-grained areas, but the coarse nature of LiDAR scans in outdoor datasets means many details are overlooked. For this scenario, we believe that employing some advanced depth completion methods to refine sparse areas in the ground truth could be a reasonable approach.

2) Future Work: In future work, we aim to delve into super-resolution techniques to augment the detail information within extensive textureless areas, which will significantly bolster the performance in outdoor environments. Furthermore, the present study has adopted a distinct spatial domain modeling strategy to address the issue of detail loss. Yet, the utilization of high-frequency components in the frequency domain for such fine-grained information presents itself as an inherently viable alternative. Moving forward, we intend to experiment with frequency domain analysis techniques, including wavelet transformations, to facilitate the restoration of fine-grained regional information.



(a) CF-Net + ours

(b) Ground-truth in KITTI 2015 dataset (training split)

Fig. 10. Failure case and ground-truth in the outdoor scenarios.

VI. Conclusion

We have presented an inter-scale similarity guided cost aggregation method designed to adaptively recover details in fine-grained areas. By leveraging both low-resolution and high-resolution information, our approach effectively exploits detail while generating inter-scale similarity measurements. Additionally, our stereo-content-aware cost aggregation method employs a decomposition strategy that divides the 3D disparity-spatial space into 1D disparity space and 2D spatial space, significantly reducing computational costs associated with 3D cost volumes. Experimental results across three benchmarks demonstrate the effectiveness of our method with various models.

References

[1] R. Chabra, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, "StereoDRNet: Dilated residual StereoNet," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11778–11787.

[2] Q. Chen et al., "Virtual blood vessels in complex background using stereo X-ray images," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 99–106.

[3] T. Pire, T. Fischer, J. Civera, P. De Cristóforis, and J. J. Berlles, "Stereo parallel tracking and mapping for robot localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 1373–1378.

[4] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2722–2730.

[5] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

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

[7] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5510–5519.

[8] 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.

[9] 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.

[10] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12981–12990.

[11] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A context integrated residual pyramid network for stereo matching," in *Proc. Asian Conf. Comput. Vis.* Springer, 2018, pp. 20–35.

[12] 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.

[13] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3268–3277.

[14] V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "HITNet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14362–14372.

[15] H. Hirschmüller, P. R. Innocent, and J. Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 229–246, 2002.

[16] K.-J. Yoon and I. So Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.

[17] C. Xu, C. Wu, D. Qu, F. Xu, H. Sun, and J. Song, "Accurate and efficient stereo matching by log-angle and pyramid-tree," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4007–4019, Oct. 2021.

[18] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, vol. 2, Jul. 2001, pp. 508–515.

[19] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[20] 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.

[21] X. Zhang, H. Dai, H. Sun, and N. Zheng, "Algorithm and VLSI architecture co-design on efficient semi-global stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4390–4403, Nov. 2020.

[22] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A deep visual correspondence embedding model for stereo matching costs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 972–980.

[23] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proc. BMVC*, 2016, vol. 2, no. 3, p. 4.

[24] J. L. Schonberger, S. N. Sinha, and M. Pollefeys, "Learning to fuse proposals from multiple scanline optimizations in semi-global matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 739–755.

[25] J. Žbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1592–1599.

[26] A. Seki and M. Pollefeys, "SGM-Nets: Semi-global matching with neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 231–240.

[27] K. Batsos, C. Cai, and P. Mordohai, "CBMV: A coalesced bidirectional matching volume for disparity estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2060–2069.

[28] Y. Zhang et al., "Adaptive unimodal cost volume filtering for deep stereo matching," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12926–12934.

[29] 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.

[30] 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.

[31] 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.

[32] H. Dai, X. Zhang, Y. Zhao, H. Sun, and N. Zheng, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3099–3110, May 2022.

[33] T. Song, S. Kim, and K. Sohn, "Unsupervised deep asymmetric stereo matching with spatially-adaptive self-similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13672–13680.

[34] J. Jing et al., "Uncertainty guided adaptive warping for robust and efficient stereo matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, vol. 29, Oct. 2023, pp. 3295–3304.

[35] J. Zeng, C. Yao, L. Yu, Y. Wu, and Y. Jia, "Parameterized cost volume for stereo matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18347–18357.

[36] C. Yao and L. Yu, "FoggyStereo: Stereo matching with fog volume representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13033–13042.

[37] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Depth-aware multi-grid deep homography estimation with contextual correlation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4460–4472, Jul. 2022.

[38] J. Cheng, X. Yang, Y. Pu, and P. Guo, "Region separable stereo matching," *IEEE Trans. Multimedia (TMM)*, vol. 26, pp. 4880–4893, 2022.

[39] 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.

[40] X. Ye, J. Zhang, Y. Yuan, R. Xu, Z. Wang, and H. Li, "Underwater depth estimation via stereo adaptation networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 5089–5101, Sep. 2023.

[41] 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.

[42] Q. Su and S. Ji, "ChiTransformer: Towards reliable stereo from cues," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1929–1939.

[43] H. Zhao, H. Zhou, Y. Zhang, J. Chen, Y. Yang, and Y. Zhao, "High-frequency stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1327–1336.

[44] G. Xu, Y. Wang, J. Cheng, J. Tang, and X. Yang, "Accurate and efficient stereo matching via attention concatenation volume," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2461–2474, Apr. 2024.

[45] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 185–194.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] O. Rukundo and H. Cao, "Nearest neighbor value interpolation," 2012, *arXiv:1211.1768*.

[51] M. Mastyło, "Bilinear interpolation theorems and applications," *J. Funct. Anal.*, vol. 265, no. 2, pp. 185–207, Jul. 2013.

[52] D. A. Rajon and W. E. Bolch, "Marching cube algorithm: Review and trilinear interpolation adaptation for image-based dosimetric models," *Computerized Med. Imag. Graph.*, vol. 27, no. 5, pp. 411–435, Sep. 2003.

[53] R. E. Carlson and F. N. Fritsch, "Monotone piecewise bicubic interpolation," *SIAM J. Numer. Anal.*, vol. 22, no. 2, pp. 386–400, Apr. 1985.

[54] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[55] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[56] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 269–284.

[57] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, 2016, pp. 1–9.

[58] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3007–3016.

[59] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE++: Unified content-aware reassembly of features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4674–4687, Sep. 2022.

[60] D. Mazzini, "Guided upsampling network for real-time semantic segmentation," 2018, *arXiv:1807.07466*.

[61] 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.

[62] D. Scharstein et al., "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.

[63] T. Schöps, T. Sattler, and M. Pollefeys, "BAD SLAM: Bundle adjusted direct RGB-D SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 134–144.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[65] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[66] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[67] Q. Zhang, X. Zhang, B. Li, Y. Chen, and A. Ming, "HDA-net: Horizontal deformable attention network for stereo matching," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 32–40.

[68] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "DeepPruner: Learning efficient stereo matching via differentiable PatchMatch," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4384–4393.

[69] Q. Wang, S. Shi, S. Zheng, K. Zhao, and X. Chu, "FADNet++: Real-time and accurate disparity estimation with configurable networks," 2021, *arXiv:2110.02582*.

[70] A. F. Kadmin, R. A. Hamzah, M. N. A. Manap, M. S. Hamid, and T. F. T. Wook, "Local stereo matching algorithm using modified dynamic cost computation," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 22, no. 3, pp. 1312–1319, Jun. 2021.

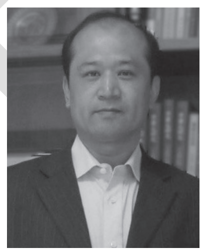
[71] M. N. Z. Azali, R. A. Hamzah, and Z. M. Noh, "Disparity map algorithm using census transform and hierarchical segment-tree from stereo image," in *Proc. Eng. Technol. Int. Conf. (ETIC)*, Sep. 2022, pp. 244–249.

[72] M. Zahari, R. A. Hamzah, N. A. Manap, and A. I. Herman, "Stereo matching algorithm for autonomous vehicle navigation using integrated matching cost and non-local aggregation," *Bull. Electr. Eng. Informat.*, vol. 12, no. 1, pp. 328–337, Feb. 2023.

[73] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, and P. Torr, "Domain-invariant stereo matching networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 420–439.

Pengxiang Li received the B.S. degree from the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China, in 2021, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and machine learning.

Chengtang Yao received the B.S. and M.S. degrees in computer science from Beijing Institute of Technology (BIT), China, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include computer vision, pattern recognition, and machine learning.



Yunde Jia (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Beijing Institute of Technology (BIT), China, in 1983, 1986, and 2000, respectively. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University (CMU), from 1995 to 1997. He is currently a Professor with the Department of Engineering, Shenzhen MSU-BIT University (SMBU). He is also the Director of Guangdong Provincial Key Laboratory of Machine Perception and Intelligent Computing. His research interests include computer vision, vision-based HCI and HRI, and intelligent robotics.

Yuwei Wu (Member, IEEE) received the Ph.D. degree in computer science from Beijing Institute of Technology (BIT), Beijing, China, in 2014. From August 2014 to August 2016, he was a Post-Doctoral Research Fellow with the Rapid-Rich Object Search (ROSE) Laboratory, School of Electrical and Electronic Engineering (EEE), Nanyang Technological University (NTU), Singapore. He is currently a tenured Associate Professor with the School of Computer Science, BIT. He has strong research interests include computer vision and machine learning.