

# BANet: Bilateral Aggregation Network for Mobile Stereo Matching

Gangwei Xu<sup>1</sup>, Jiaxin Liu<sup>1</sup>, Xianqi Wang<sup>1</sup>, Junda Cheng<sup>1</sup>, Yong Deng<sup>2</sup>  
Jinliang Zang<sup>2</sup>, Yurui Chen<sup>2</sup>, Xin Yang<sup>3,1†</sup>

<sup>1</sup>Huazhong University of Science and Technology <sup>2</sup>Autel Robotics <sup>3</sup>Optics Valley Laboratory

## Abstract

State-of-the-art stereo matching methods typically use costly 3D convolutions to aggregate a full cost volume, but their computational demands make mobile deployment challenging. Directly applying 2D convolutions for cost aggregation often results in edge blurring, detail loss, and mismatches in textureless regions. Some complex operations, like deformable convolutions and iterative warping, can partially alleviate this issue; however, they are not mobile-friendly, limiting their deployment on mobile devices. In this paper, we present a novel bilateral aggregation network (BANet) for mobile stereo matching that produces high-quality results with sharp edges and fine details using only 2D convolutions. Specifically, we first separate the full cost volume into detailed and smooth volumes using a spatial attention map, then perform detailed and smooth aggregations accordingly, ultimately fusing both to obtain the final disparity map. Experimental results demonstrate that our BANet-2D significantly outperforms other mobile-friendly methods, achieving 35.3% higher accuracy on the KITTI 2015 leaderboard than MobileStereoNet-2D, with faster runtime on mobile devices. Code: <https://github.com/gangweix/BANet>.

## 1. Introduction

Metric depth estimation plays a critical role in a wide range of real-world applications, such as drone navigation, smartphone photography, and robotic surgery. It can be broadly categorized into tasks include stereo matching [57], depth completion [29, 49, 69], and monocular depth estimation [19, 44, 61], and so on. Among them, stereo matching focuses on finding pixel correspondences between left and right images, enabling depth recovery via triangulation. Currently, deep learning-based methods [8, 16, 18, 20, 46–48, 59] have dominated stereo matching or depth benchmarks, consistently setting new state-of-the-art results on public leaderboards [35, 37, 38]. De-

<sup>†</sup>Corresponding author.

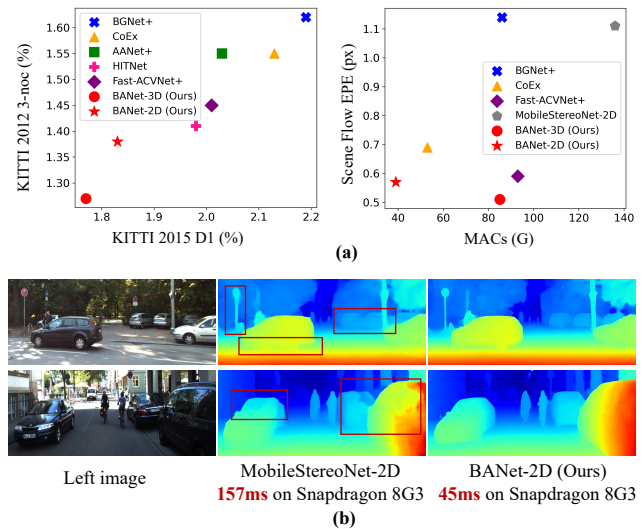


Figure 1. (a): Comparison with top-performing real-time methods (high-end GPUs) [1, 42, 50, 55, 58]. (b): Visual results and latency on Qualcomm Snapdragon 8 Gen 3 (8G3). MobileStereoNet-2D [39] blurs edges, loses details, and causes mismatches in textureless regions due to pure 2D convolutions for cost aggregation. In comparison, our bilateral aggregation effectively addresses these issues while remaining mobile-friendly, eliminating the need for complex operations [42, 58]. The latency is measured for input stereo pairs with a resolution of  $512 \times 512$ .

spite significant advancements, deep stereo matching methods still face challenges when deployed on mobile and embedded devices. These challenges include: 1) computational and memory demands exceeding mobile platforms’ capabilities; 2) difficulties in deploying complex or customized operators [42, 58]; and 3) the trade-off where simpler models tend to sacrifice accuracy. In this paper, we aim to explore a mobile-friendly stereo matching method that achieves both real-time inference speed and high accuracy using only mobile-friendly operations.

Recently, to improve stereo matching performance, some works [15, 41, 51] have designed sophisticated cost volume representations, which play a crucial role in the stereo matching pipeline. For example, GwcNet [15] designs a

group-wise correlation volume that computes correlations group by group and then concatenates them. ACVNet [51, 55] introduces an attention concatenation volume that transforms the correlation volume into attention weights, which are then used to filter the concatenation volume [3, 23]. PCWNet [41] proposes a pyramid combination and warping cost volume. These cost volume representations, in conjunction with a large number of 3D convolutions, provide a notable improvement in terms of prediction accuracy. However, the accompanying high computational and memory costs make it nearly impossible to deploy on mobile devices.

To accelerate the runtime of stereo matching, subsequent works have adopted downsampled or sparse cost volume representations [11, 24, 50, 55, 62], as well as lightweight aggregation networks [1, 27, 54, 56]. For example, StereoNet [24] and BGNet [50] construct a low-resolution cost volume (1/8 of the image resolution), while DeepPruner [11] and Fast-ACVNet [55] create sparse cost volumes by pruning the candidate disparity range. These methods can achieve real-time performance on high-end GPUs; however, they still require stacked 3D convolutions to regularize the cost volumes, which are not mobile-friendly and make it difficult for these methods to meet real-time demands on mobile devices, such as those powered by Qualcomm Snapdragon chips.

An alternative solution is to replace 3D convolutions with 2D convolutions for cost volume aggregation; however, this can lead to edge blurring at disparity discontinuities or a loss of fine details. To mitigate these issues, AANet [58] adopts deformable convolutions [68] to perform adaptive cost aggregation. HITNet [42] introduces iterative warping operations to progressively restore the edges and fine details of the disparity map. Unfortunately, these complex operations, such as deformable convolutions and iterative warping, are not mobile-friendly and are quite expensive to deploy on mobile devices. Contrary to these methods, MobileStereoNet-2D [39] utilizes pure 2D convolutions for cost aggregation; however, it exhibits a severe degradation in accuracy, as shown in Fig. 1. To this end, a motivating question arises: *How to design a mobile-friendly stereo network while maintaining high prediction accuracy?*

In this paper, we propose a novel bilateral aggregation network (BANet) that simultaneously achieves real-time performance on mobile devices and high prediction accuracy. Considering that images contain both high-frequency detailed regions and low-frequency smooth/textureless regions, which are essential for achieving accurate predictions, directly applying standard 2D convolutions for cost aggregation across all regions tends to blur edges, lose details, and produce mismatches in textureless areas. To address this, the proposed BANet first separates the full

cost volume into a detailed and smooth cost volume, then aggregates each individually, and finally fuses them. Through this bilateral aggregation, comprising a detailed aggregation branch and a smooth aggregation branch, our method achieves state-of-the-art performance while preserving clear edges and fine details in complex scenes, as shown in Fig. 1.

Furthermore, accurately identifying high-frequency detail regions and low-frequency smooth regions is also crucial for final prediction accuracy. For this purpose, we propose a new scale-aware spatial attention module to differentiate between high-frequency details and edges, and smooth regions. The features of different scales have varying perceptions and receptive fields. Fine-scale features can perceive more high-frequency detail information, while coarse-scale features capture more low-frequency smooth information. Our scale-aware spatial attention takes full advantage of features at different scales to produce an accurate spatial attention map that effectively separates detailed and smooth regions.

To demonstrate the effectiveness of our approach, we conduct extensive experiments on the Scene Flow [34] and KITTI [13, 35] datasets. Our pure 2D convolution-based BANet-2D outperforms other lightweight methods [1, 42, 50, 55] on the KITTI 2012 and 2015 leaderboards. We examine the latency on Qualcomm Snapdragon 8 Gen 3. As shown in Fig. 1, our BANet-2D is mobile-friendly and takes only 45ms for input stereo pairs with a resolution of  $512 \times 512$ , delivering high-quality results that preserve edges and details, significantly outperforming MobileStereoNet-2D [39]. We also extend BANet to a 3D version, and the resulting BANet-3D model achieves the highest prediction accuracy among all real-time methods (on high-end GPUs) on the KITTI 2012 and 2015 leaderboards.

In summary, our main contributions are as follows:

- We present a novel bilateral aggregation network for mobile stereo matching that achieves high-quality results using only 2D convolutions.
- We propose a scale-aware spatial attention module that accurately identifies high-frequency details and edges, and low-frequency smooth regions.
- Our approach can run in real-time on mobile devices with high prediction accuracy, significantly outperforming other mobile-friendly methods.
- The extended 3D version, BANet-3D, achieves the highest accuracy among all real-time methods on GPUs.

## 2. Related Work

### 2.1. Deep Stereo Methods

Recently, deep stereo methods can primarily be categorized into two types: cost volume aggregation-based approaches [3, 7, 15, 31, 32, 41, 51, 63–65] and iterative

optimization-based approaches [2, 5, 8, 9, 12, 21, 22, 25, 26, 28, 30, 45, 52, 53, 66, 67]. A representative method within the first category is PSMNet [3]. PSMNet constructs a 4D concatenation volume and employs a stacked hourglass network, composed of 3D convolutions, to aggregate this volume. Due to the simplicity and excellent performance of PSMNet, many subsequent works have attempted to improve it in terms of cost volume construction and cost aggregation. For example, GwcNet [15] proposes group-wise correlation volume, ACVNet [51] introduces attention concatenation volume, and PCWNet [41] presents pyramid combination and warping volume. These stereo matching methods enhance the representational capacity of the cost volume, leading to improved accuracy. However, they typically come with an expensive computational cost. To improve efficiency, Cascade-Stereo [14] and CFNet [40] propose cascade cost volume representations, constructing the cost volume in a coarse-to-fine manner.

For cost aggregation, GA-Net [63] introduces two guided aggregation layers to replace the widely used 3D convolutional layer, while CoAtRS [6] proposes global attention along the disparity dimension for more comprehensive aggregation. However, these methods incur high computational costs, making real-time deployment challenging even on high-end GPUs, let alone on mobile devices.

Iterative optimization-based methods [25, 30] iteratively update disparity using matching features retrieved from a correlation volume, thus avoiding the computationally expensive cost aggregation operations. However, they typically require a large number of iterations to obtain an optimal disparity. To improve optimization efficiency and accuracy, IGEV [53, 57] introduces a more comprehensive geometry encoding volume, from which matching features are iteratively indexed to update the disparity. Despite its excellent performance, it still struggles to be deployed on mobile devices due to the iterative indexing operations.

## 2.2. Real-time Stereo Methods

To speed up stereo matching inference time, many methods [4, 11, 24, 43, 50] directly construct and aggregate a deeply downsampled cost volume, such as 1/8 of the image resolution. However, these downsampled cost volumes can lead to a significant degradation in accuracy. To maintain comparable accuracy, CoEx [1] still constructs a high-resolution cost volume but uses a more lightweight aggregation network. Fast-ACVNet [55], on the other hand, introduces a high-resolution sparse attention module that only computes sparse matches at a high resolution. However, these methods achieve real-time inference only on high-end GPUs. The extensive use of 3D convolutions makes them challenging to deploy on mobile devices.

To replace costly 3D convolutions while maintaining comparable accuracy, AANet [58] leverages deformable 2D

convolutions to enable adaptive cost aggregation, thereby alleviating the well-known edge-fattening issue. Unlike aggregation-based approaches, HITNet [42] avoids constructing an explicit cost volume and instead progressively recovers a full-resolution disparity through iterative warping operations. However, complex operations such as deformable convolutions and iterative warping are generally not mobile-friendly, making them challenging to deploy on mobile devices.

MobileStereoNet-2D [39] employs 2D MobileNet blocks [36] for cost aggregation, which are mobile-friendly. However, images typically contain information at varying frequencies, and objects exhibit different disparities. As a result, 2D aggregation often leads to edge blurring, detail loss, and mismatches in textureless regions. In contrast, our bilateral aggregation adaptively separates the cost volume based on the corresponding frequency information or disparities, and then performs targeted aggregation for each.

## 3. Bilateral Aggregation Network

In this section, we introduce the detailed structure of the proposed BANet, illustrated in Fig. 2. It consists of four steps: feature extraction, correlation volume construction, bilateral aggregation, and disparity prediction. Most existing methods use 3D convolutions to aggregate cost volumes, which improves accuracy but is computationally expensive and unsuitable for mobile devices. In contrast, 2D convolutions are lightweight but often cause blurring and mismatches. Our proposed bilateral aggregation achieves high accuracy using only efficient, mobile-friendly operations.

### 3.1. Bilateral Aggregation

An image typically contains both high-frequency detailed regions and low-frequency smooth or textureless regions. Therefore, using only a 2D aggregation network to aggregate the entire cost volume makes it difficult to manage both detailed and smooth regions, often resulting in edge blurring, detail loss, and mismatches in textureless regions (as shown in Fig. 1). To address these issues, we propose bilateral aggregation, which first separates the full cost volume into detailed and smooth volumes. It then uses a detailed aggregation branch for the detailed volume and a smooth aggregation branch for the smooth volume (as shown in Fig. 2).

Specifically, given a full correlation volume  $C_{cor}$  constructed through simple feature correlation, we separate it into a detailed cost volume  $C_d$  and a smooth cost volume  $C_s$ . This separation operation is based on a spatial attention map  $A$  introduced in Sec. 3.2,

$$\begin{aligned} C_d &= A \odot C_{cor}, \\ C_s &= (1 - A) \odot C_{cor}, \end{aligned} \quad (1)$$

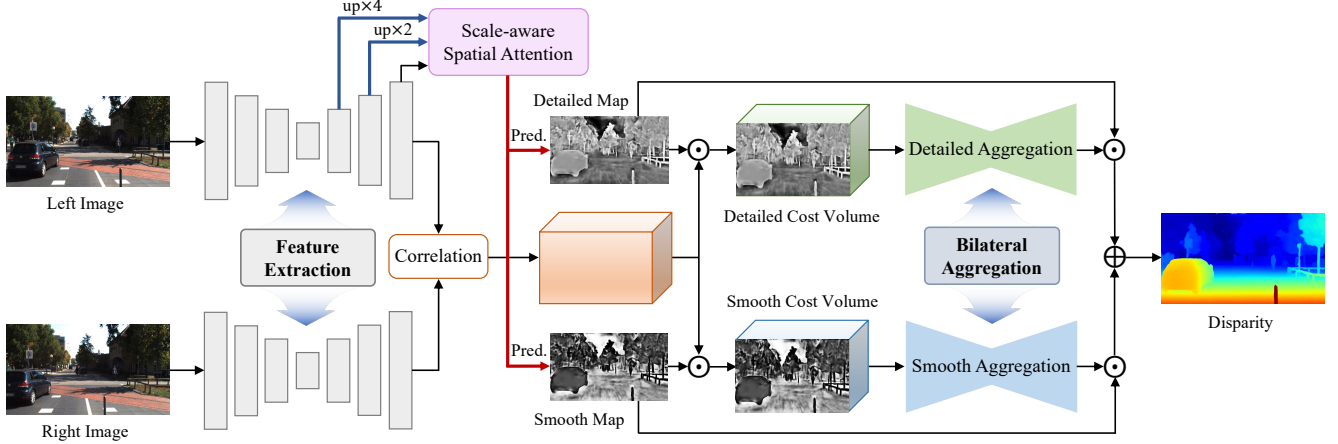


Figure 2. Overview of our proposed Bilateral Aggregation Network (BANet). To effectively handle both high-frequency detailed regions and low-frequency smooth regions, we use detailed and smooth maps to separate the full cost volume into detailed and smooth volumes. This enables targeted aggregation for each, with the detailed and smooth volumes processed independently. The detailed map highlights high-frequency detailed regions, while the smooth map highlights the opposite. We also introduce a new scale-aware spatial attention to more accurately identify detailed and smooth regions within the image.

where  $\odot$  represents the Hadamard Product. The spatial attention map  $\mathbf{A}$  highlights the high-frequency detailed regions, while the map  $(1 - \mathbf{A})$  highlights the low-frequency smooth regions.

After separating into detailed and smooth cost volumes, we accordingly employ a detailed aggregation branch  $\mathbf{G}_d$  to aggregate the detailed cost volume and a smooth aggregation branch  $\mathbf{G}_s$  for the smooth cost volume,

$$\begin{aligned} \mathbf{C}'_d &= \mathbf{G}_d(\mathbf{C}_d), \\ \mathbf{C}'_s &= \mathbf{G}_s(\mathbf{C}_s). \end{aligned} \quad (2)$$

For simplicity, we adopt the same structure for both  $\mathbf{G}_d$  and  $\mathbf{G}_s$ , but do not share their weights. Here, we only detail  $\mathbf{G}_d$ . Following previous work, the detailed aggregation network  $\mathbf{G}_d$  consists of a series of inverted residual blocks [36]: 4 blocks at 1/4 resolution, 6 blocks at 1/8 resolution, and 8 blocks at 1/16 resolution. Each inverted residual block contains point-wise, depth-wise, and another point-wise 2D convolution, with an expansion factor of 4. The channel numbers for the cost features at 1/4, 1/8, and 1/16 resolutions are 32, 64, and 128, respectively.

The detail aggregation network targets high-frequency regions, while the smooth aggregation network focuses on low-frequency, textureless regions for disparity predictions. Finally, we fuse the aggregated detailed cost volume and the smooth cost volume by:

$$\mathbf{C}_{agg} = \mathbf{A} \odot \mathbf{C}'_d + (1 - \mathbf{A}) \odot \mathbf{C}'_s. \quad (3)$$

**Extension to 3D.** Naturally, we can extend the concept of bilateral aggregation to 3D convolutions, further boosting the accuracy of our model. The 3D aggregation net-

work includes three down-sampling blocks and three up-sampling blocks. Each down-sampling block contains two  $3 \times 3 \times 3$  kernel-sized 3D convolutions, while each up-sampling block includes a  $4 \times 4 \times 4$  kernel-sized 3D transposed convolution followed by two  $3 \times 3 \times 3$  kernel-sized 3D convolutions. Our 3D extension, BANet-3D, achieves the highest accuracy among all published real-time methods on high-end GPUs.

### 3.2. Scale-aware Spatial Attention

Fine-scale image features capture more high-frequency details and edges, while coarse-scale features encompass more low-frequency, smooth, and textureless information. Therefore, to accurately separate detailed regions and smooth regions, we propose a scale-aware spatial attention module that learns the differences in multi-scale image features to generate an attention map. This map effectively distinguishes detailed regions and smooth regions.

As shown in Fig. 2, the multi-scale left image features  $\mathbf{F}_{l,16}$ ,  $\mathbf{F}_{l,8}$ , and  $\mathbf{F}_{l,4}$  are scaled to 1/4 resolution before being input into our spatial attention module. The scaled features are represented as  $\mathbf{F}_{l,16}^{up}$ ,  $\mathbf{F}_{l,8}^{up}$ , and  $\mathbf{F}_{l,4}$ . First, we apply a convolutional layer to each scaled feature to obtain intermediate features with the same number of channels, and then we concatenate them. Second, we use another convolutional layer followed by a *sigmoid* activation function to predict the spatial attention map. In this way, the attention map  $\mathbf{A}$  is obtained by:

$$\begin{aligned} \mathbf{S} &= \text{Concat}([\text{Conv}(\mathbf{F}_{l,16}^{up}), \text{Conv}(\mathbf{F}_{l,8}^{up}), \text{Conv}(\mathbf{F}_{l,4})]), \\ \mathbf{A} &= \sigma(\text{Conv}(\mathbf{S})), \end{aligned} \quad (4)$$

where *Concat* indicates the concatenation operator, *Conv* represents the 2D convolutional operator, and  $\sigma$  denotes the *sigmoid* function.

As shown in Fig. 2 and Fig. 3, the spatial attention map  $\mathbf{A}$  effectively highlights high-frequency details and edges, as these regions typically exhibit high feature values and distinct variations within the scale-aware perception. By applying a reverse operation, we obtain an inverse spatial attention map  $(1 - \mathbf{A})$  that highlights low-frequency smooth and textureless regions.

### 3.3. Network Architecture

**Feature Extraction.** Given the left image  $\mathbf{I}_l \in \mathbb{R}^{3 \times H \times W}$  and the right image  $\mathbf{I}_r \in \mathbb{R}^{3 \times H \times W}$ , we employ a pre-trained MobileNetV2 on ImageNet [10] as our backbone, extracting multi-scale feature maps at 1/4, 1/8, 1/16, and 1/32 of the original resolution, respectively. Starting from the 1/32 resolution image features, we iteratively apply up-sampling blocks until reaching a 1/4 resolution. In more detail, each up-sampling block employs a transpose convolution with a  $4 \times 4$  kernel and a stride of 2, followed by a  $3 \times 3$  kernel-sized convolution. Finally, we obtain multi-scale left image features:  $\mathbf{F}_{l,4}$  at 1/4 resolution,  $\mathbf{F}_{l,8}$  at 1/8 resolution, and  $\mathbf{F}_{l,16}$  at 1/16 resolution, which are then used for scale-aware spatial attention generation, while  $\mathbf{F}_{l,4}$  and  $\mathbf{F}_{r,4}$  are used for correlation volume construction, as shown in Fig. 2.

**Correlation Volume Construction.** Given the left feature map  $\mathbf{F}_{l,4}$  and right feature map  $\mathbf{F}_{r,4}$ , we construct the correlation volume by,

$$\mathbf{C}_{cor}(d, x, y) = \frac{1}{N_c} \langle \mathbf{F}_{l,4}(x, y), \mathbf{F}_{r,4}(x - d, y) \rangle, \quad (5)$$

where  $\langle \cdot, \cdot \rangle$  is the inner product of two feature vectors,  $N_c$  denotes the number of channels, and  $d$  represents the disparity level.

**Disparity Prediction.** We use the proposed bilateral aggregation to aggregate the correlation/cost volume, which is detailed in Sec. 3.1. After obtaining the aggregated cost volume, we apply the softmax operation to it to regress the disparity map  $\mathbf{d}_0$ :

$$\mathbf{d}_0 = \sum_{d=0}^{D_{max}/4-1} d \times \text{Softmax}(\mathbf{C}_{agg}(d)), \quad (6)$$

where  $D_{max}$  denotes the predefined maximum disparity value, and  $d$  represents the predefined disparity range from 0 to  $D_{max}/4 - 1$ . The disparity map  $\mathbf{d}_0$  has a size of  $H/4 \times W/4$ . We use superpixel weights [60] around each pixel in the left image for a weighted combination of local neighboring points in  $\mathbf{d}_0$ , resulting in a full-resolution disparity map  $\mathbf{d}_1 \in \mathbb{R}^{H \times W}$ .

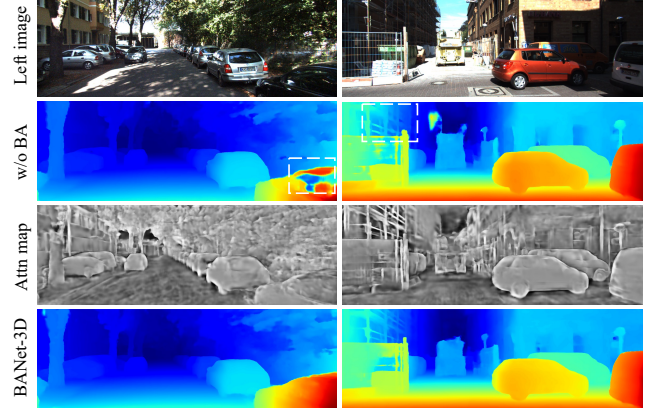


Figure 3. Our bilateral aggregation (BA) significantly enhances performance in both detailed and textureless regions. The Attention map highlights the high-frequency detailed regions. Significant improvements are highlighted by white dashed boxes.

### 3.4. Loss Function

The entire network is trained in a supervised, end-to-end manner, with the final loss function defined as follows:

$$\mathcal{L} = \lambda_0 \text{Smooth}_{L_1}(\mathbf{d}_0 - \mathbf{d}_{gt}) + \lambda_1 \text{Smooth}_{L_1}(\mathbf{d}_1 - \mathbf{d}_{gt}) \quad (7)$$

where  $\mathbf{d}_{gt}$  denotes the ground-truth disparity map, and  $\text{Smooth}_{L_1}$  represents Smooth L1 loss.

## 4. Experiments

### 4.1. Datasets and Evaluation Metrics

**Scene Flow** [34] is a synthetic dataset, consisting of Flyingthings3D, Driving, and Monkaa. The dataset provides 35,454 training pairs and 4,370 testing pairs of size  $960 \times 540$  with dense disparity maps. The Scene Flow dataset provides two versions: Cleanpass and Finalpass. We use the Finalpass of Scene Flow for training and testing since it is more like real-world images than the Cleanpass, which contains more motion blur and defocus. The end-point error (EPE) and disparity outlier rate Bad 3.0 are used as the evaluation metrics. Bad 3.0 is defined as the percentage of pixels with disparity error (EPE) greater than 3 pixels.

**KITTI 2012** [13] and **KITTI 2015** [35] are datasets for real-world driving scenes. KITTI 2012 contains 194 training pairs and 195 testing pairs, and KITTI 2015 contains 200 training pairs and 200 testing pairs. Both datasets provide sparse ground-truth disparities obtained with LiDAR. We submit our predicted disparity maps to the KITTI website to obtain quantitative evaluation results. For KITTI 2012, we report the percentage of pixels with errors larger than  $x$  disparities in both non-occluded ( $x$ -noc) and all regions ( $x$ -all), as well as the overall EPE in both non-occluded (EPE-noc) and all the pixels (EPE-all). For KITTI 2015, we report the percentage of pixels with EPE larger

Agg. Type	Model	Bilateral Aggregation	Scale-aware Spatial Attn	EPE (px)	Bad 3.0 (%)	MACs (G)
2D Conv	Baseline	✗	✗	0.63	2.75	29
	BA	✓	✗	0.59	2.57	38
	BA+SSA (BANet-2D)	✓	✓	<b>0.57</b>	<b>2.49</b>	39
3D Conv	Baseline	✗	✗	0.56	2.43	57
	BA	✓	✗	0.53	2.27	80
	BA+SSA (BANet-3D)	✓	✓	<b>0.51</b>	<b>2.21</b>	85

Table 1. Ablation study on the Scene Flow test set. We integrate bilateral aggregation into two types of aggregation networks: 2D aggregation networks and 3D aggregation networks. Results demonstrate that our bilateral aggregation can significantly improve the accuracy of existing aggregation networks.

Method	KITTI 2012		KITTI 2015		
	3-noc	3-all	D1-bg	D1-fg	D1-all
w/o BA	1.77	2.21	1.85	3.67	2.15
BANet-2D	<b>1.38</b>	<b>1.79</b>	<b>1.59</b>	<b>3.03 (17%↑)</b>	<b>1.83</b>
w/o BA	1.54	2.01	1.66	3.87	2.03
BANet-3D	<b>1.27</b>	<b>1.72</b>	<b>1.52</b>	<b>3.02 (22%↑)</b>	<b>1.77</b>

Table 2. Ablation study on the test sets of KITTI 2012 and 2015. Our bilateral aggregation (BA) effectively enhances the prediction accuracy for both background (D1-bg) and foreground (D1-fg) regions, particularly in the foreground, which typically contains more high-frequency details and edges.

Method	EPE (px)	MACs (G)
DeepPruner-Fast [11]	0.97	219
StereoNet [24]	1.10	104
BGNet+ [50]	1.14	86
MobileStereoNet-2D [39]	1.11	136
MobileStereoNet-3D [39]	0.80	615
CoEx [1]	0.67	53
LightStereo-L [17]	0.59	92
Fast-ACVNet [55]	0.64	79
Fast-ACVNet+ [55]	0.59	93
BANet-2D (Ours)	<b>0.57</b>	<b>39</b>
BANet-3D (Ours)	<b>0.51</b>	85

Table 3. Quantitative evaluation on the Scene Flow test set. The MACs are measured for an input size of  $960 \times 540$ .

than  $\max(3\text{px}, 0.05d_{gt})$  in background regions (D1-bg), foreground regions (D1-fg), and all (D1-all).

## 4.2. Implementation Details

We implement our approaches with PyTorch and perform our experiments using NVIDIA RTX 3090 GPUs. We first train our approaches on the Scene Flow dataset for 200k steps with a batch size of 16, and then fine-tune the pre-

trained Scene Flow model on a mixed dataset of KITTI 2012 and 2015 training sets for 50k steps. During training, images are randomly cropped to a size of  $256 \times 512$ . For all experiments, we use the AdamW [33] optimizer with a one-cycle learning rate schedule, where the maximum learning rate is set to  $8e-4$ . In our experiments, we set  $\lambda_0$  and  $\lambda_1$  to 0.3 and 1.0, respectively.  $D_{max}$  is set to 192.

## 4.3. Ablation Study

We conduct extensive ablation studies on the Scene Flow [34] and KITTI [13, 35] datasets to validate the effectiveness of the proposed approaches. The proposed bilateral aggregation is versatile and applicable to various aggregation networks. As shown in Tab. 1, we apply it to 2D aggregation networks composed of 2D convolutions (BANet-2D) and 3D aggregation networks composed of 3D convolutions (BANet-3D), respectively. Compared to a single-branch aggregation network (Baseline), our bilateral aggregation (BA) can adaptively divide the image into high-frequency detail regions and low-frequency smooth regions, and then aggregate them accordingly. Without SSA, the attention map  $\mathbf{A}$  is generated from the  $1/4$  scale feature  $\mathbf{F}_{l,4}$ . As a result, BA provides significant improvements with minimal computational cost.

To more accurately identify high-frequency and low-frequency regions in the image, we propose a scale-aware spatial attention module that learns the differences in multi-scale image features to generate a spatial attention map for effectively separating high-frequency details and edges, and low-frequency smooth regions. Tab. 1 shows that scale-aware spatial attention can further boost performance.

We also present the ablation results on the test sets of KITTI 2012 [13] and 2015 [35], as shown in Tab. 2. Compared to the improvements on the synthetic Scene Flow test set, our bilateral aggregation achieves more significant gains on the challenging real-world test sets of KITTI 2012 and 2015. We highlight the improvements in the foreground regions (D1-fg) on KITTI test sets, as these regions usually contain more high-frequency details and edges. Specif-

Method	KITTI 2012 [13]						KITTI 2015 [35]			MACs (G)
	3-noc	3-all	4-noc	4-all	EPE noc	EPE all	D1-bg	D1-fg	D1-all	
DispNetC [34]	4.11	4.65	2.77	3.20	0.9	1.0	4.32	4.41	4.34	-
AANet+ [58]	1.55	2.04	1.20	1.58	0.4	0.5	1.65	3.96	2.03	-
DecNet [62]	-	-	-	-	-	-	2.07	3.87	2.37	-
BGNet+ [50]	1.62	2.03	1.16	1.48	0.5	0.6	1.81	4.09	2.19	76
CoEx [1]	1.55	1.93	1.15	1.42	0.5	0.5	1.79	3.82	2.13	49
DeepPruner-Fast[11]	-	-	-	-	-	-	2.32	3.91	2.59	194
HITNet [42]	1.41	1.89	1.14	1.53	0.4	0.5	1.74	3.20	1.98	47
Fast-ACVNet+ [55]	1.45	1.85	1.06	1.36	0.5	0.5	1.70	3.53	2.01	85
Fast-ACVNet [55]	1.68	2.13	1.23	1.56	0.5	0.6	1.82	3.93	2.17	72
MobileStereoNet-2D [39]	-	-	-	-	-	-	2.49	4.53	2.83	127
MobileStereoNet-3D [39]	-	-	-	-	-	-	2.75	3.87	2.10	564
BANet-2D (Ours)	<u>1.38</u>	<u>1.79</u>	<u>1.01</u>	<u>1.32</u>	0.5	0.5	<u>1.59</u>	<u>3.03</u>	<u>1.83</u>	<b>36</b>
BANet-3D (Ours)	<b>1.27</b>	<b>1.72</b>	<b>0.95</b>	<b>1.27</b>	0.5	0.5	<b>1.52</b>	<b>3.02</b>	<b>1.77</b>	78

Table 4. Quantitative evaluation on the test sets of KITTI 2012 [13] and KITTI 2015 [35]. Previous methods reported their runtime on their own GPUs; however, the runtime can vary across different GPU models. For a fair comparison, we measure MACs, which are consistent across GPU models. The MACs are measured for an input size of  $1242 \times 375$ . **Bold**: Best, Underline: Second best.

Method	EPE (px)	Bad 3.0 (%)
PSMNet [3]	1.09	4.68
BA+PSMNet	<b>0.77</b>	<b>3.28</b>
GwcNet [15]	0.76	3.30
BA+GwcNet	<b>0.67</b>	<b>2.89</b>
Fast-ACVNet+ [55]	0.59	2.70
BA+Fast-ACVNet+	<b>0.53</b>	<b>2.25</b>

Table 5. Performance of Bilateral Aggregation (BA). Our BA can be seamlessly integrated into cost-volume aggregation methods, significantly enhancing their performance.

ically, for the D1-fg metric, our bilateral aggregation improves accuracy by 17% for 2D convolution-based aggregation and 22% for 3D convolution-based aggregation.

Qualitative results are shown in Fig. 3. The attention map distinguishes high-frequency details and edges, and our bilateral aggregation preserves fine structures while ensuring accurate matching in textureless regions.

#### 4.4. Comparisons with State-of-the-Art Methods

**Quantitative Comparisons.** Tab. 3 and Tab. 4 present quantitative comparison results on the Scene Flow, KITTI 2012, and KITTI 2015 test sets. Our BANet-3D achieves the highest accuracy among the published real-time methods [1, 39, 42, 50, 55] on high-end GPUs for almost all metrics. However, due to the use of 3D convolutions, BANet-3D is challenging to deploy on mobile devices. Without the use of 3D convolutions, our BANet-2D is more

mobile-friendly and easy to deploy on mobile platforms. Although it is slightly inferior to BANet-3D, it surpasses all other lightweight methods. Specifically, on the KITTI 2015 test set, BANet-2D surpasses MobileStereoNet-2D [39] by 35.3%, and BANet-3D outperforms FastACVNet+ [55] by 11.9% for the D1-all metric.

Previous methods [1, 42, 50, 55, 58] reported their runtime on their respective GPUs; however, the runtime can vary across different GPUs. To ensure a fair comparison of the computational complexity across methods, we uniformly measure MACs (Multiply-Accumulate Operations), which remain consistent regardless of GPU model. Our BANet-2D achieves the lowest MACs among all methods.

**Qualitative Comparisons.** We compared the visual results of our methods with the mobile-friendly 2D convolution-based MobileStereoNet-2D [39] and the state-of-the-art 3D convolution-based Fast-ACVNet+ [55]. As shown in Fig. 4, a single-branch aggregation network often struggles to effectively handle both high-frequency edges and details, as well as large textureless regions, resulting in edge blurring, detail loss, and mismatches in textureless regions. In contrast, by employing this divide-and-conquer idea, our proposed bilateral aggregation produces clear edges and preserves intricate detail structures, resulting in accurate matching in large areas of textureless regions. In particular, our performance gains are more pronounced when integrated into simpler 2D convolution-based methods.

**Latency on Mobile Device.** We compare the latency with the latest mobile-friendly method, MobileStereoNet-2D [39], on the Qualcomm Snapdragon 8 Gen 3, as shown

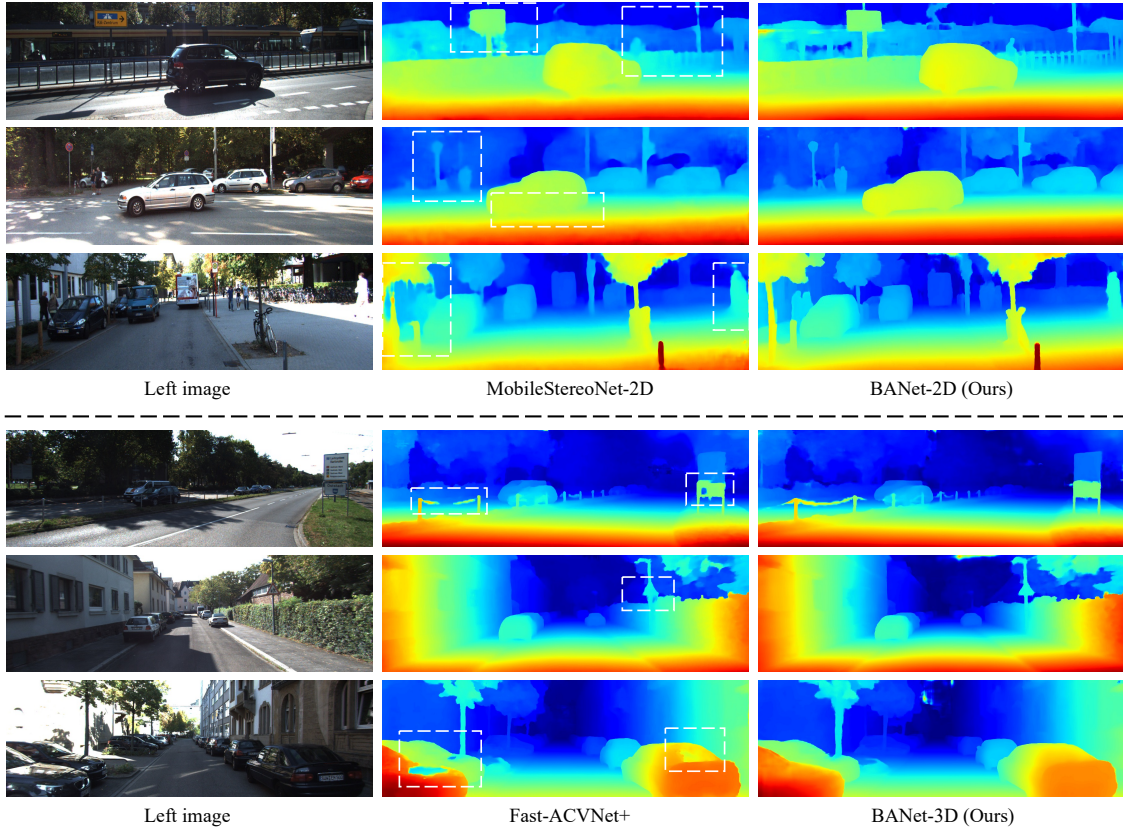


Figure 4. Qualitative comparisons on the test sets of KITTI 2012 [13] and 2015 [35]. By employing this divide-and-conquer approach, our bilateral aggregation produces clear edges and preserves intricate details, enabling accurate matching even in large textureless regions.

in Fig. 1. Benefiting from the proposed bilateral aggregation, our BANet-2D achieves sharp edges and accurate matching in textureless regions, while requiring only 45ms, which is less than one-third of MobileStereoNet-2D’s latency. Furthermore, we present a detailed breakdown of the latency: 16ms for feature extraction, 6.5ms for correlation volume construction, and 22.5ms for bilateral aggregation.

#### 4.5. Universality and Superiority of BA

To demonstrate the universality and superiority of the proposed bilateral aggregation, we integrate it into three representative methods, namely PSMNet [3], GwcNet [15], and Fast-ACVNet+ [55], and compare the performance of the original methods with their counterparts enhanced by bilateral aggregation. The comparison results are presented in Tab. 5. Our bilateral aggregation significantly enhances those cost-volume aggregation-based methods, such as improving Fast-ACVNet+ by 10.2% in the EPE metric.

## 5. Conclusion

This paper presents a novel bilateral aggregation network for mobile stereo matching that achieves high-quality re-

sults using only 2D convolutions. To effectively handle both detailed and smooth regions, we propose bilateral aggregation, which separates the full cost volume into detailed and smooth cost volumes, and then performs detailed and smooth aggregations accordingly. To more accurately distinguish between detailed and smooth regions, we propose a new scale-aware spatial attention module. Experimental results demonstrate that our method can run in real-time on mobile devices with high prediction accuracy, significantly outperforming existing methods.

An exciting future direction could involve extending our approach to other aggregation-based tasks, *e.g.* multi-view stereo and optical flow estimation. Additionally, we believe that our mobile-friendly design could offer significant advantages for practical applications, such as drone navigation and intelligent photography.

**Acknowledgements.** This research is supported by the National Natural Science Foundation of China (623B2036, 62472184), the National Key R&D Program of China(2024YFE0217700), the Fundamental Research Funds for the Central Universities, and the Innovation Project of Optics Valley Laboratory (Grant No. OVL2025YZ005).

## References

- [1] Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *IROS*, pages 3542–3548. IEEE, 2021. 1, 2, 3, 6, 7
- [2] Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In *CVPR*, pages 1013–1027, 2025. 3
- [3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 2, 3, 7, 8
- [4] Jia-Ren Chang, Pei-Chun Chang, and Yong-Sheng Chen. Attention-aware feature aggregation for real-time stereo matching on edge devices. In *ACCV*, 2020. 3
- [5] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu. Mocha-stereo: Motif channel attention network for stereo matching. In *CVPR*, pages 27768–27777, 2024. 3
- [6] Junda Cheng, Gangwei Xu, Peng Guo, and Xin Yang. Coatsnet: Fully exploiting convolution and attention for stereo matching by region separation. *IJCV*, 132(1):56–73, 2024. 3
- [7] Junda Cheng, Wei Yin, Kaixuan Wang, Xiaozhi Chen, Shijie Wang, and Xin Yang. Adaptive fusion of single-view and multi-view depth for autonomous driving. In *CVPR*, pages 10138–10147, 2024. 2
- [8] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. In *CVPR*, pages 6273–6282, 2025. 1, 3
- [9] Zhien Dai, Zhaohui Tang, Hu Zhang, Can Tian, Mingjun Pan, and Yongfang Xie. Eglcr: Edge structure guidance and scale adaptive attention for iterative stereo matching. In *ACMMM*, pages 4197–4206, 2024. 3
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5
- [11] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *ICCV*, pages 4384–4393, 2019. 2, 3, 6, 7
- [12] Miaojie Feng, Junda Cheng, Hao Jia, Longliang Liu, Gangwei Xu, and Xin Yang. Mc-stereo: Multi-peak lookup and cascade search range for stereo matching. In *3DV*, pages 344–353. IEEE, 2024. 3
- [13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 2, 5, 6, 7, 8
- [14] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, pages 2495–2504, 2020. 3
- [15] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, pages 3273–3282, 2019. 1, 2, 3, 7, 8
- [16] Xianda Guo, Chenming Zhang, Juntao Lu, Yiqi Wang, Yiqun Duan, Tian Yang, Zheng Zhu, and Long Chen. Openstereo: A comprehensive benchmark for stereo matching and strong baseline. *arXiv preprint arXiv:2312.00343*, 2023. 1
- [17] Xianda Guo, Chenming Zhang, Dujun Nie, Wenzhao Zheng, Youmin Zhang, and Long Chen. Lightstereo: Channel boost is all your need for efficient 2d cost aggregation. *arXiv preprint arXiv:2406.19833*, 2024. 6
- [18] Xianda Guo, Chenming Zhang, Youmin Zhang, Dujun Nie, Ruilin Wang, Wenzhao Zheng, Matteo Poggi, and Long Chen. Stereo anything: Unifying stereo matching with large-scale mixed data. *arXiv preprint arXiv:2411.14053*, 2024. 1
- [19] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE TPAMI*, 2024. 1
- [20] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. *arXiv preprint arXiv:2501.09466*, 2025. 1
- [21] Junpeng Jing, Jiankun Li, Pengfei Xiong, Jiangyu Liu, Shuaicheng Liu, Yichen Guo, Xin Deng, Mai Xu, Lai Jiang, and Leonid Sigal. Uncertainty guided adaptive warping for robust and efficient stereo matching. In *ICCV*, pages 3318–3327, 2023. 3
- [22] Junpeng Jing, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. Stereo any video: Temporally consistent stereo matching. *arXiv preprint arXiv:2503.05549*, 2025. 3
- [23] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, pages 66–75, 2017. 2
- [24] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for edge-aware depth prediction. In *ECCV*, 2018. 2, 3, 6
- [25] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *CVPR*, pages 16263–16272, 2022. 3
- [26] Kunhong Li, Longguang Wang, Ye Zhang, Kaiwen Xue, Shunbo Zhou, and Yulan Guo. Los: Local structure-guided stereo matching. In *CVPR*, pages 19746–19756, 2024. 3
- [27] Ximeng Li, Chen Zhang, Wanjuan Su, and Wenbing Tao. Iinet: Implicit intra-inter information fusion for real-time stereo matching. In *AAAI*, pages 3225–3233, 2024. 2
- [28] Zhaohuai Liang and Changhe Li. Any-stereo: Arbitrary scale disparity estimation for iterative stereo matching. In *AAAI*, pages 3333–3341, 2024. 3
- [29] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *CVPR*, pages 17070–17080, 2025. 1

- [30] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, pages 218–227. IEEE, 2021. 3
- [31] Biyang Liu, Huimin Yu, and Yangqi Long. Local similarity pattern and cost self-reassembling for deep stereo matching networks. In *AAAI*, pages 1647–1655, 2022. 2
- [32] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, Yao Yao, and Luc Van Gool. Stereo risk: A continuous modeling approach to stereo matching. *arXiv preprint arXiv:2407.03152*, 2024. 2
- [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [34] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 2, 5, 6, 7
- [35] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015. 1, 2, 5, 6, 7, 8
- [36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 3, 4
- [37] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, pages 31–42. Springer, 2014. 1
- [38] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 1
- [39] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *WACV*, pages 2417–2426, 2022. 1, 2, 3, 6, 7
- [40] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnets: Cascade and fused cost volume for robust stereo matching. In *CVPR*, pages 13906–13915, 2021. 3
- [41] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *ECCV*, pages 280–297. Springer, 2022. 1, 2, 3
- [42] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *CVPR*, pages 14362–14372, 2021. 1, 2, 3, 7
- [43] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. Fadnet: A fast and accurate network for disparity estimation. In *ICRA*, pages 101–107. IEEE, 2020. 3
- [44] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, pages 5261–5271, 2025. 1
- [45] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *CVPR*, pages 19701–19710, 2024. 3
- [46] Xianqi Wang, Hao Yang, Gangwei Xu, Junda Cheng, Min Lin, Yong Deng, Jinliang Zang, Yurui Chen, and Xin Yang. Zerostereo: Zero-shot stereo matching from single images. *arXiv preprint arXiv:2501.08654*, 2025. 1
- [47] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Johann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *ICCV*, pages 17969–17980, 2023.
- [48] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *arXiv preprint arXiv:2501.09898*, 2025. 1
- [49] Jijun Xiang, Xuan Zhu, Xianqi Wang, Yu Wang, Hong Zhang, Fei Guo, and Xin Yang. Depthor: Depth enhancement from a practical light-weight dtof sensor and rgb image. *arXiv preprint arXiv:2504.01596*, 2025. 1
- [50] Bin Xu, Yuhua Xu, Xiaoli Yang, Wei Jia, and Yulan Guo. Bilateral grid learning for stereo matching networks. In *CVPR*, pages 12497–12506, 2021. 1, 2, 3, 6, 7
- [51] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *CVPR*, pages 12981–12990, 2022. 1, 2, 3
- [52] Gangwei Xu, Shujun Chen, Hao Jia, Miaojie Feng, and Xin Yang. Memory-efficient optical flow via radius-distribution orthogonal cost volume. *arXiv preprint arXiv:2312.03790*, 2023. 3
- [53] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, pages 21919–21928, 2023. 3
- [54] Gangwei Xu, Huan Zhou, and Xin Yang. Cgi-stereo: Accurate and real-time stereo matching via context and geometry interaction. *arXiv preprint arXiv:2301.02789*, 2023. 2
- [55] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE TPAMI*, 46(4):2461–2474, 2024. 1, 2, 3, 6, 7, 8
- [56] Gangwei Xu, Yujin Wang, Jinwei Gu, Tianfan Xue, and Xin Yang. Hdrflow: Real-time hdr video reconstruction with large motions. In *CVPR*, pages 24851–24860, 2024. 2
- [57] Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. Igev++: Iterative multi-range geometry encoding volumes for stereo matching. *IEEE TPAMI*, 2025. 1, 3
- [58] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, pages 1959–1968, 2020. 1, 2, 3, 7
- [59] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatofghi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE TPAMI*, 2023. 1
- [60] Fengting Yang, Qian Sun, Hailin Jin, and Zihan Zhou. Superpixel segmentation with fully convolutional networks. In *CVPR*, pages 13964–13973, 2020. 5

- [61] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NIPS*, 37:21875–21911, 2024. [1](#)
- [62] Chengtang Yao, Yunde Jia, Huijun Di, Pengxiang Li, and Yuwei Wu. A decomposition model for stereo matching. In *CVPR*, pages 6091–6100, 2021. [2](#), [7](#)
- [63] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2019. [2](#), [3](#)
- [64] Jiawei Zhang, Lei Huang, Xiao Bai, Jin Zheng, Lin Gu, and Edwin Hancock. Exploring the usage of pre-trained features for stereo matching. *IJCV*, pages 1–22, 2024.
- [65] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. In *AAAI*, pages 12926–12934, 2020. [2](#)
- [66] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *CVPR*, pages 1327–1336, 2023. [3](#)
- [67] Yang Zhao, Gangwei Xu, and Gang Wu. Hybrid cost volume for memory-efficient optical flow. In *ACM MM*, pages 8740–8749, 2024. [3](#)
- [68] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. [2](#)
- [69] Xuan Zhu, Jijun Xiang, Xianqi Wang, Longliang Liu, Yu Wang, Hong Zhang, Fei Guo, and Xin Yang. Svdc: Consistent direct time-of-flight video depth completion with frequency selective fusion. In *CVPR*, pages 16619–16628, 2025. [1](#)