Geometry-Aware Stereo Matching via Monocular Disparity Distribution Prior and Gradient Enhancement

Junze Zhang[1, 2], Luoxi Jing[3], Yuanyuan Wang[1], Xueqi Li[1], Guoli Yang[4], Songchang Jin[2*], Chunping Qiu[2]

[1]Academy of Military Science, Beijing 100850, China
[2]Intelligent Game and Decision Lab (IGDL), Beijing 100091, China
[3]School of Computer Science, Peking University, Beijing, 100871, China
[4]Advanced Institute of Big Data, Beijing, China

Abstract

Stereo matching recovers 3D scene information based on the correlation between corresponding pixels. Despite impressive progress, existing methods lack sufficient correlation priors in ill-posed regions such as occlusions, detailed and reflective regions. In this paper, we propose Geometry Aware Stereo Matching Network (GEAStereo) to enhance geometric structure perception and address this issue. We adaptively incorporate the Monocular Disparity Distribution Prior into the stereo cost volume, building Mono-Stereo Fusion Volume (MSFV), which effectively captures global geometric structures and rectifies the correlation information in ill-posed regions. Furthermore, we introduce rich detail information from gradient features and construct a Detail-Aware Volume (DAV) by aggregating the group-wise cost volume under the guidance of gradient spatial attention, thus enhancing the correlation modeling in detailed structures. Jointly, MSFV and DAV provide rich correlation priors for disparity iterative optimization. Experimental results show that our method achieves competitive results on the ETH3D and KITTI2015 benchmarks. Compared with the state-of-the-art methods, our method demonstrates stronger performance in zero-shot generalization.

Introduction

Stereo matching is a fundamental topic in computer vision. Its primary task is to to establish pixel-wise correspondences between the left and right image pairs (Hirschmuller 2008), and then restore the depth and geometric information of the 3D scene based on the horizontal coordinate difference of the pixel pairs, known as disparity. This technique is widely used in tasks such as 3D modeling, autonomous driving, virtual reality, and robot navigation.

In recent years, stereo matching methods based on deep learning (Kendall et al. 2017; Lipson, Teed, and Deng 2021; Xu et al. 2023) have notably enhanced the accuracy of disparity estimation on standard benchmarks. These methods can be categorized into cost aggregation-based methods and iterative optimization-based methods. Cost aggregation-based methods (Mayer et al. 2016; Pang et al. 2017; Kendall et al. 2017) construct a 3D or 4D cost volume by concatenating or performing dot products on CNN features of stereo images. Subsequently, they utilize aggregation modules based on 2D or 3D convolutions (Wang et al. 2024b) to directly aggregate the cost volume to obtain disparity. In contrast, iterative optimization-based methods (Lipson, Teed, and Deng 2021; Xu et al. 2023; Wang et al. 2024c) construct the all-pairs cost volume, which takes the current disparity as a prior. The lookup operator is employed to sample the cost volume and obtain local cost volume features. Subsequently, an iterative optimizer based on ConvGRU is utilized to iteratively optimize the disparity. These methods avoid the high computational cost of 3D convolution and update the disparity through lightweight GRU units.

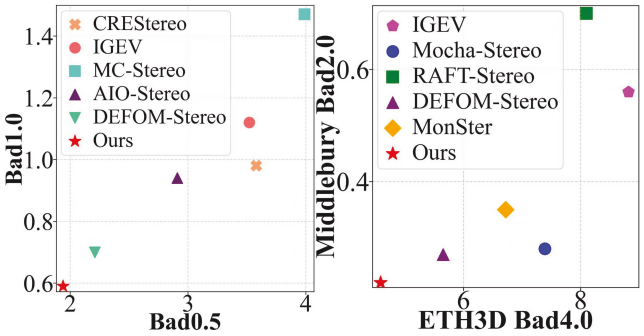

Figure 1: **Left**: Comparison with state-of-the-art stereo methods on ETH3D. **Right**: Zero-Shot Performance comparison on ETH3D and Middlebury.**bad0.6**

Both of these methods rely on the cost volume as matching priors to capture the correspondences between matching pixels. In ill-posed regions such as occlusions, textureless and reflective/detailed regions, the cost volume often fails to capture reliable matching costs, leading to mismatching and presenting a formidable challenge to stereo matching methods. Some works (Wang et al. 2024c; Zhao et al. 2023) alleviate mismatching by enhancing the processing of high-frequency disparity information. Monster (Cheng et al. 2025) and DEFOM-Stereo (Jiang et al. 2025) leverage powerful monocular depth from DepthAnyThingV2 (Yang et al. 2024) as priors to generate the initial disparity. Introducing monocular depth provides complementary structural information for ill-posed regions. However, these methods overlook the fact that matching costs of the cost volume remain

*Corresponding author's e-mail: jsc04@tsinghua.org.cn.

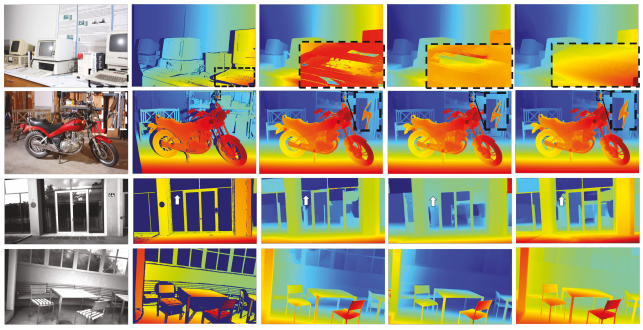

Figure 2: Zero-Shot Qualitative Comparison on two datasets. **Row 1, 2**: Middlebury. **Row 3, 4**: ETH3D.

missing or inaccurate in ill-posed regions, introducing noisy priors for disparity optimization. Meanwhile, the scale alignment between monocular depth and disparity varies across scenes, leading to limited zero-shot generalization.

To address the limitations of previous methods, we propose a novel stereo matching method, named GEAStereo, which improves the correlation information of the cost volume in ill-posed regions through Monocular Disparity Distribution Prior Volume (MDPV) and Gradient Enhancement. Unlike previous methods that rely on monocular depth, we employ a monocular prior decoder based on CNN to generate MDPV from left features, which represents pixel-wise possibility cues belonging to different disparity levels. By doing this, we avoid scale alignment error from monocular depth across different scenes. To adaptively propagate the monocular prior into cost volume, we propose a Monocular Disparity Prior-Guided Fusion Module (MPF) to fuse the stereo cost volume and MDPV, yielding a Mono-Stereo Fusion Volume (MSFV). This enables us to capture global geometric structures and correct correlation information of cost volume in ill-posed regions. While MSFV improves global structural awareness, fine details are often lost during low-resolution cost aggregation. We then introduce RGB gradient features and design a Detail-Aware Aggregation Module (DAA) to perform aggregation on the 4D group-wise cost volume. DAA applies Gradient Spatial Attention during the aggregation process to enhance the detailed geometric information in group-wise cost volume, obtaining Detal-Aware Volume (DAV). Jointly, MSFV and DAV provide rich correlation priors for disparity iterative optimization. We conduct numerous experiments to validate the effectiveness and superiority of our method. A brief comparison with published SOTA methods is demonstrated in Figure 1.

Our primary contributions can be summarized as follows.

- We propose a Monocular Disparity Distribution Prior Volume and utilize this prior to globally refine or complete the correlation information in stereo cost volume.
- We design a Detail-Aware Volume, which introduces gradient spatial enhancement to perceive matching cost information for detailed and thin structures.
- Our proposed method outperforms existing published methods on public leaderboards such as KITTI2015, ETH3D benchmark. In zero-shot generalization experiments across diverse datasets, our method achieves best performance compared with SOTA methods.

Related Work

Learning-Based Stereo Matching

To encoder correlation between pixels, early learning-based stereo methods (Mayer et al. 2016; Pang et al. 2017) construct 2D correlation cost volume to model the correlation information between corresponding matching pixels, followed by disparity regression through a 2D CNN (Wang et al. 2025a). Subsequent works have focused on enhancing the representation ability of cost volume. GC-Net (Kendall et al. 2017) constructs a 4D concatenated cost volume by feature concatenation at disparity levels. GwcNet (Guo et al. 2019) groups channels to build group correlation volume, effectively enhancing the similarity priors of the cost volume.

RAFT-Stereo (Lipson, Teed, and Deng 2021) transfers the RAFT (Teed and Deng 2020) framework from optical flow to stereo matching tasks and proposes an iterative optimization-based stereo matching paradigm. In recent years, this paradigm has gradually become mainstream, achieving state-of-the-art results in both accuracy and efficiency. IGEV (Xu et al. 2023) performs 3D aggregation for cost volume and generates initial disparity, which introduces geometric priors for GRU iterative optimization. Selective-Stereo (Wang et al. 2024c) decouples disparity estimation by using GRUs with different kernel sizes. (Cheng et al. 2025; Jiang et al. 2025) introduce the Vision Foundation Model DepthAnythingV2 (Yang et al. 2024) into the stereo matching. By leveraging the powerful depth generalization ability of DepthAnythingV2, they generate the inverse depth map which is further aligned with GRU, providing monocular priors for ill-posed regions. AIO-Stereo (Zhou et al. 2025) leverages knowledge from multi Visual Foundation Models to stereo matching.

Gradient Enhancement in Visual Tasks

Gradient information quantifies the intensity of RGB value variations at each pixel within an image. Gradient features possess a robust capacity for representing structural details. In recent years, researchers have incorporated gradient features as supplementary structural priors to enhance the sharpness of blurred structures. TCIN(Kim et al. 2022) trains a fine-refinement network with gradient loss for image super-resolution, which aims to boost the generation of high-frequency details. DSR-EI (Qiao et al. 2023) employs the Transformer architecture to extract gradient features to recover sharp edges in depth discontinuities structures. SGNet(Wang, Yan, and Yang 2024) exploits the structural information embedded in RGB gradients to rectify the blurred structures in low-resolution depth maps.

Method

In this section, we detail our GEAStereo architecture, which consists of a multi-scale feature extractor, MSFV construction, DAV construction and a GRU-based iterative optimizer, as shown in Figure 3.

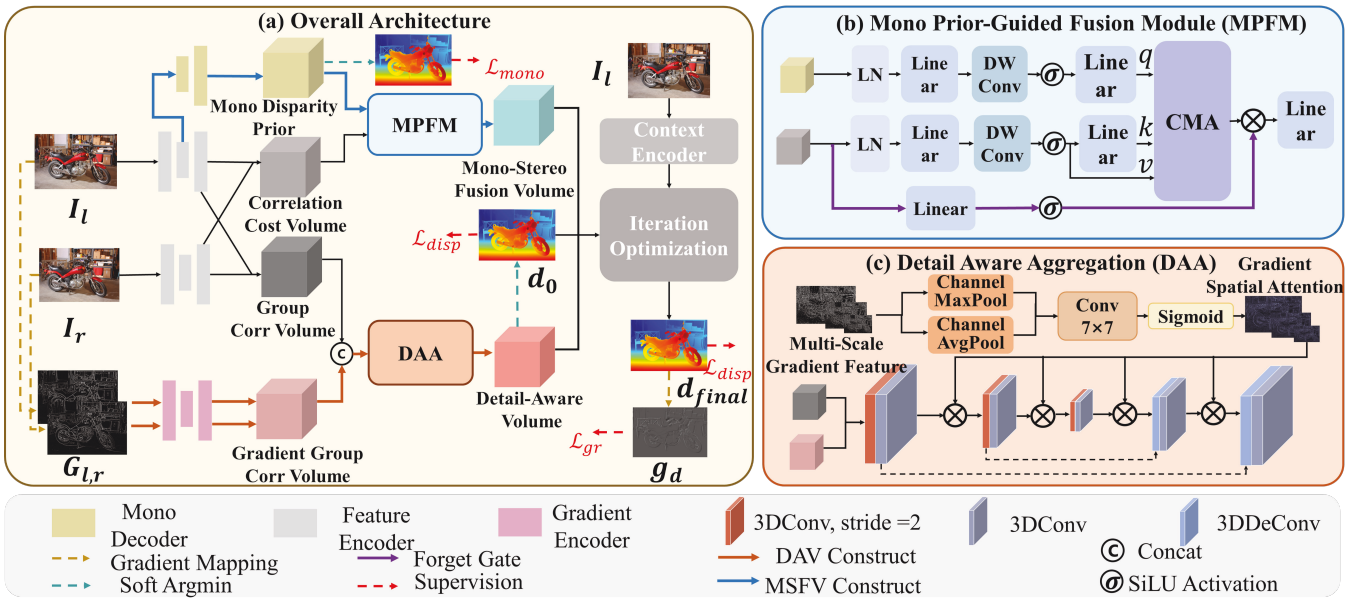

Figure 3: Left: The overall framework of GEAStereo. GEAStereo builds a Mono-Stereo-Fusion Volume (MSFV) and a Detal-Aware Volume (DAV). Then we regress an initial disparity from DAV and iteratively optimize it with ConvGRUs. Right: The structure of our proposed Mono-Stereo Fusion module and Detail Aware Aggregation module.

Feature Extractor

We separate the feature extractor into three parts: context encoder, feature encoder, and gradient encoder. For the left and right images$\{\mathbf{I}_l, \mathbf{I}_r \in \mathbb{R}^{3\times H\times W}\}$, we follow (Xu et al. 2023) and employ feature encoder to generate multi-scale features$\{\mathbf{f}_l^i, \mathbf{f}_r^i \in \mathbb{R}^{C_i\times H/2^{i+1}\times W/2^{i+1}}\}(i = 1,2,3,4)$, which are used to construct the cost volume and encode the pixel correlations between the stereo images.

Following (Xu et al. 2023), the context encoder is processed to generate multi-scale context features$\{\mathbf{f}_c^i \in \mathbb{R}^{C\times H/2^{j+1}\times W/2^{j+1}}\}(j = 1,2,3)$ from left images. These features provide the initial hidden state for the GRU and inject the context information of the left image into iterative optimization.

Meanwhile, to introduce the rich geometric detail information in gradient map, we use a gradient mapping to transform the stereo images into the gradient maps $\mathbf{G}_l, \mathbf{G}_r$. The calculation formula is as follows:

$$\mathbf{G}_{l/r}(x,y) = \left\| \begin{pmatrix} \mathbf{I}_{l/r}(x+1,y) - \mathbf{I}_{l/r}(x-1,y) \\ \mathbf{I}_{l/r}(x,y+1) - \mathbf{I}_{l/r}(x,y-1) \end{pmatrix} \right\|_2 \quad (1)$$

To process the mapped gradient maps, we employ a light UNet-like network to extract multi-scale gradient features$\{\mathbf{g}_l^k, \mathbf{g}_r^k \in \mathbb{R}^{C_k\times H/2^{k+1}\times W/2^{k+1}}\}(k = 1,2,3)$.

Mono-Stereo Fusion Volume Based on Monocular Disparity Distribution Prior

Pixels lacking correspondences cannot provide valid matching costs. This introduces noisy correlation information into the GRU iterative optimization, causing mismatches in ill-posed regions.

We first construct a monocular texture-guided prior decoder that further processes the left features $\mathbf{f}_l$. This process infers monocular disparity distribution prior information from the left image texture, providing geometric and matching information guidance for the ill-posed regions. Starting from the feature at 1/32 resolution, we employ the Trap Block (Ning and Gan 2023) for decoding. Through multiple stages of Trap Block and Trap interpolation (Ning and Gan 2023) for upsampling, we obtain multi-scale prior features. Then we upsample all prior feature maps to $H/4 \times W/4$ resolution using bilinear interpolation and concatenate them together. After processing them through a monocular prior head, we obtain the Monocular Disparity Prior Volume (MDPV) $\mathbf{P}_{mono}$ with dimensions $H/4\times W/4\times D_{max}$, representing the distribution possibility at different disparity levels.

By applying $soft\ argmin$ operation to MDPV, we obtain the "monocular disparity" $\mathbf{d}_{mono}$ guided by monocular feature, which is used for intermediate supervision and guiding the generation of accurate prior information of MDPV.

$$\mathbf{d}_{mono} = \sum_{d=0}^{D_{max}-1} d \times \mathrm{Softmax}(\mathbf{P}_{mono}(d)) \quad (2)$$

Meanwhile, we construct the stereo correlation cost volume $\mathbf{C}_{stereo}$ by taking the vector inner product of the stereo feature pairs, calculating the pixel correlation for the corresponding disparity level. As previously mentioned, the stereo cost volume often fails to capture effective correlation relationship in ill-posed regions.

$$\mathbf{C}_{stereo}(d,x,y) = \langle \mathbf{f}_l^1(x,y), \mathbf{f}_r^1(x-d,y) \rangle \quad (3)$$

To integrate the geometric prior information and disparity distribution information from MDPV into the stereo cost

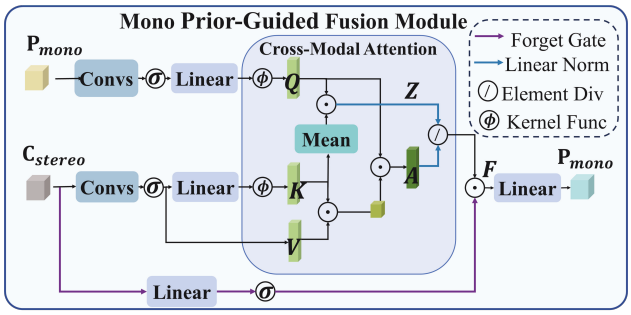

Figure 4: Monocular Disparity Prior-Guided Fusion Module (MPF).

volume, and to guide the alignment of matching costs between ill-posed and normal regions in the correlation cost volume, we employ a Monocular Disparity Prior-Guided Fusion Module (MPF) to fuse the stereo cost volume and the disparity distribution prior volume.

Ill-posed regions, like occlusions, often manifest as aggregated regions and CNN networks have limited receptive fields, they are unable to effectively capture global pixel relationships. Inspired by (Han et al. 2024), we combine the forget-gating mechanism (Gu and Dao 2024) with a Cross-Modal Attention (CMA) to construct MPF. MPF aims to capture long-range dependencies and extract global associations between prior and cost, obtaining Mono-Stereo Fusion Volume (MSFV). The detailed architecture of MPF is shown in Figure 4.

For given Volume $\mathbf{P}_{mono}, \mathbf{C}_{stereo} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times D_{max}}$, we first process them through shallow convolutional layers (Wang et al. 2024a) to obtain the corresponding volume features $\mathbf{F}_{mono}, \mathbf{F}_{stereo}$.

Cross-Modal Attention. We then employ CMA to fuse the volume features. First, we utilize the convolution operation $\theta(\cdot, W)$ along with the kernel function $\phi(\cdot) = elu(\cdot)+1$ to transform $\mathbf{F}_{mono}$ into the query feature $\mathbf{Q}_{mono}$. Additionally, we apply the similar operation to transform $\mathbf{F}_{stereo}$ into the key feature $\mathbf{K}_{stereo}$, while directly using $\mathbf{F}_{stereo}$ as the value feature $\mathbf{V}_{stereo}$. Specifically, the kernel function $\phi(\cdot)$ aims to perform feature mapping on $\mathbf{Q}_{mono}$ and $\mathbf{K}_{stereo}$, ensuring that each element in the features maintains non-negativity. As a result, a linear normalization (Katharopoulos et al. 2020) operation can be adopted to replace the non-linear Softmax normalization, thus reducing the computational complexity by exchanging the order of attention calculation.

After adjusting two feature shapes to $\left(\frac{H\times W}{16}, D_{\max}\right)$, the attention $\mathbf{A}_{fusion}$ is computed by multiplying $\mathbf{Q}_{mono}$, $\mathbf{K}_{stereo}$, and $\mathbf{V}_{stereo}$, extracting the global pixel correspondence between the monocular and stereo features. This process improves the global geometry and correlation information of the stereo cost volume in ill-posed regions, refining the matching costs of the unmatched regions. Subsequently, we compute the normalization factor $\mathbf{Z}$ and apply linear normalization to $\mathbf{A}_{fusion}$, obtaining the fused feature $\mathbf{F}_{fusion}$, which is computed as:

$$\begin{aligned}\mathbf{F}_{fusion} &= (\mathbf{Q}_{mono} \otimes (\mathbf{K}^T_{stereo} \otimes \mathbf{V}_{stereo}))/\mathbf{Z}, \\ \mathbf{Z} &= \mathbf{Q}_{mono} \otimes \text{mean}(\mathbf{K}^T_{stereo}),\end{aligned} \tag{4}$$

where $\text{mean}(\cdot)$ represents the averaging operation along the spatial dimension.

Forget Gate Filtering (FGF). In order to preserve the precise information from the stereo cost volume during the fusion process and suppress the interference of erroneous priors or low-confidence regions, we introduce the forget-gating mechanism (Gu and Dao 2024). By computing the forget gate FGF for the stereo cost volume and multiplying it with $\mathbf{F}_{fusion}$, we achieve information filtering in the fused information and retain the accurate matching costs in normal regions of the stereo cost volume, eventually obtaining MSFV, which is represented by $\mathbf{C}_{fusion}$.

$$\begin{aligned}\text{FGF} &= \text{SiLU}(\text{Linear}(\mathbf{C}_{stereo})), \\ \mathbf{C}_{fusion} &= \text{Linear}(\mathbf{F}_{fusion} \otimes \text{FGF})\end{aligned} \tag{5}$$

Through MPF, the stereo cost volume utilizes monocular disparity distribution prior to refine correlation information, providing global geometric structure and accurate correlation for iterative optimization.

Detail-Aware Volume Based on Gradient

While MSFV improves global structural awareness, feaAbure channels exhibit varying degrees of geometric structure loss during the feature extraction process (Chen et al. 2024). Additionally, construction of the cost volume at low resolutions leads to loss of geometric details. Therefore, we introduce the rich geometric details from the gradient map to enhance the cost computation in detailed regions.

Given binocular features $\mathbf{f}^g_{l,1}, \mathbf{f}^g_{r,1}$ at 1/4 resolution, we follow (Guo et al. 2019) to construct the group correlation volume. First, we divide the feature map channels into N_g groups and compute the correlation within each group:

$$\mathbf{C}_{gp}(g,d,x,y) = \frac{1}{N_c/N_g}\langle \mathbf{f}^{1,g}_l(x,y), \mathbf{f}^{1,g}_r(x-d,y)\rangle, \tag{6}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, N_c is the number of channels in the feature map, d is the disparity level. However, simply performing group correlation fails to effectively model the geometric structural correspondence. Therefore, we construct a gradient correlation volume $\mathbf{C}_{gd}$:

$$\mathbf{C}_{gd}(g,d,x,y) = \frac{1}{N_c/N_g}\langle \mathbf{g}^{1,g}_l(x,y), \mathbf{g}^{1,g}_r(x-d,y)\rangle \tag{7}$$

By leveraging the powerful structural representation capability of gradient features, the gradient cost volume can effectively model the correlation in edge and detailed regions. To further aggregate geometric structural perception within the cost volume, we concatenate $\mathbf{C}_{gp}$ and $\mathbf{C}_{gd}$ and design a Detail-Aware Aggregation (DAA) module to aggregate the concatenated cost volume, obtaining Detail-Aware Volume (DAV). The aggregation module is shown in the Figure 3(c).

$$\mathbf{C}_{detail} = DAA(Concat\{\mathbf{C}_{gp}, \mathbf{C}_{gd}\}, \mathbf{g}_l) \tag{8}$$

Method	RAFT-Stereo (2021)	ACVNet (2022)	IGEV (2023)	Selective-IGEV (2024c)	DEFOM-Stereo (2025)	Ours
EPE(px)↓	0.60	0.48	0.47	0.44	<u>0.42</u>	**0.40**
Params(M)	11.23	6.22	12.60	13.14	382.62	14.84

Table 1: Quantitative Evaluation on Scene Flow test set.The **best** result is bolded, and the <u>second-best</u> result is underlined.

Method	ETH3D						KITTI2015			
	Non-Occ				All		Non-Occ		All	
	Bad0.5	Bad1.0	Bad2.0	Avgerr	Bad0.5	Bad1.0	D1-bg	D1-all	D1-bg	D1-all
ACVNet (2022)	10.36	2.58	0.57	0.23	10.83	2.86	1.26	1.52	1.37	1.65
RAFT-Stereo (2021)	7.04	2.44	0.44	0.18	7.33	2.62	-	-	-	-
CREStereo (2022)	3.58	0.98	0.22	<u>0.13</u>	3.75	<u>1.09</u>	1.33	1.54	1.45	1.69
IGEV (2023)	3.52	1.12	<u>0.21</u>	0.14	3.97	1.51	1.27	1.49	1.38	1.59
MC-Stereo (2024)	3.99	1.47	0.24	0.14	4.47	1.87	1.24	1.46	1.36	<u>1.55</u>
Selective-IGEV (2024c)	3.06	1.23	0.22	**0.12**	3.46	1.56	1.22	1.44	1.33	<u>1.55</u>
IGEV++ (2025)	2.98	1.14	0.36	<u>0.13</u>	3.48	1.58	<u>1.20</u>	<u>1.42</u>	<u>1.31</u>	**1.51**
AIO-Stereo (2025)	<u>2.91</u>	<u>0.94</u>	<u>0.21</u>	<u>0.13</u>	<u>3.32</u>	1.30	1.22	1.43	1.34	<u>1.55</u>
GEAStereo (Ours)	**1.94**	**0.59**	**0.16**	**0.12**	**2.49**	**1.04**	**1.16**	**1.41**	**1.27**	**1.51**

Table 2: Quantitative Evaluation on ETH3D and KITTI2015. "All" denotes evaluation over all pixels, whereas "Non-Occ" denotes evaluation with a non-occlusion mask.

Following (Xu et al. 2023; Bangunharcana et al. 2021), DAA is based on a lightweight 3D UNet, consisting of three downsampling blocks and three upsampling blocks. Since excessive convolution layers (Wang et al. 2025b) and low resolution result in poor aggregation of details and edges, we employ multi-scale left gradient features $\mathbf{g}_l$ to enhance details in the cost volume during the 3D aggregation process. First, we apply max pooling and average pooling operations on the gradient features along the channel dimension. Then, we concatenate pooled maps and employ a 7×7 convolution layer and a sigmoid function σ to generate gradient spatial attention weights $\boldsymbol{w}_s$ with values between 0 and 1. Through these operations, the gradient space attention map will have higher weights in edge and detailed regions.

For the intermediate cost volumes $\mathbf{C}_i$ during aggregation, we apply Gradient Space Attention at the same scale to spatially enhance the cost volume, guiding the cost volume to focus on the detailed geometric information. The spatial enhancement process is expressed as,

$$\begin{aligned} \boldsymbol{w}_s^i &= \sigma\left(Conv7 \times 7(ChannelPooling(\mathbf{g}_l^i))\right) \\ \mathbf{C}_i^{detail} &= \boldsymbol{w}_s^i \odot \mathbf{C}_i \end{aligned} \tag{9}$$

GRU-based Iterative Optimization

First, we apply the $soft\ argmin$ operation to process DAV, obtaining the initial disparity enhanced with geometry details:

$$\mathbf{d}_0 = \sum_{d=0}^{D_{max}-1} d \times \mathrm{Softmax}(\mathbf{C}_{detail}(d)) \tag{10}$$

Due to the incorporation of rich geometric detail information from gradient maps, the initial disparity provides precise prior of geometric detail for subsequent GRU iterations.

Following (Wang et al. 2024c), the iterative optimization part consists of dual-kernel GRUs and obtain the disparity residual $\Delta\mathbf{d}_k$ during each iteration. We update the iterative disparity map as $\mathbf{d}_{k+1} = \mathbf{d}_k + \Delta\mathbf{d}_k$.

Loss function

Our loss function consists of three components: disparity loss $\mathcal{L}_{disp}$, disparity gradient loss $\mathcal{L}_{grad}$, and monocular disparity loss $\mathcal{L}_{mono}$. For the initial disparity generated by the DAV and the disparity obtained at each iteration, we use Smooth L1 loss (Chang and Chen 2018) and L1 loss with weights exponentially increasing (Lipson, Teed, and Deng 2021) for supervision, respectively.

$$\mathcal{L}_{disp} = |\mathbf{d}_0 - \mathbf{d}_{gt}|_{\text{smooth}} + \sum_{i=1}^{N-1} \gamma^{N-i} \|\mathbf{d}_i - \mathbf{d}_{gt}\|_1, \tag{11}$$

where $\mathbf{d}_{gt}$ is the ground truth disparity, N is the number of disparity iterations, and γ = 0.9. Additionally, we employ Smooth L1 loss to supervise the monocular disparity:

$$\mathcal{L}_{mono} = \mathrm{Smooth}_{L_1}(\mathbf{d}_{mono} - \mathbf{d}_{gt}) \tag{12}$$

We also apply the similar loss as $\mathcal{L}_{disp}$ to supervise the gradients of the initial and final disparities, enhancing the disparity in edge and detailed regions.

$$\begin{aligned} \mathcal{L}_{gr} &= |\mathbf{G}_0 - \mathbf{G}_{gt}|_{\text{smooth}} + \sum_{i=1}^{N} \gamma^{N-i} \|\mathbf{G}_i - \mathbf{G}_{gt}\|_1, \\ \mathbf{G}_{gt} &= \nabla_{u,v}\mathbf{d}_{gt},\ \mathbf{G}_i = \nabla_{u,v}\mathbf{d}_i, \end{aligned} \tag{13}$$

where $\nabla_{u,v}$ denotes the gradient of the variable in the horizontal and vertical directions.

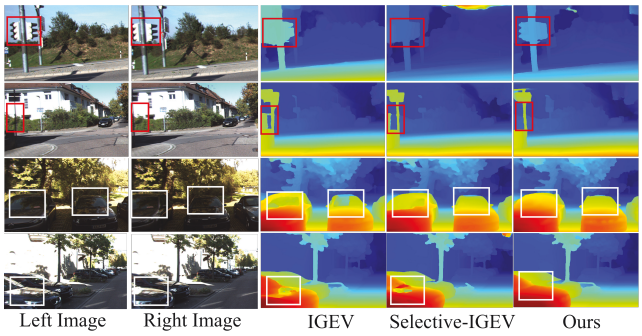

Figure 5: Qualitative Comparison on KITTI. **Row 1, 2**: Occlusions. **Row 3, 4**: Reflective Regions.

Method	Reflective Regions			
	Out-3 Noc	Out-3 All	Out-4 All	Out-5 All
CREStereo (2022)	6.27	7.27	5.55	4.59
IGEV (2023)	4.35	5.00	3.57	2.86
Selective-IGEV (2024c)	3.79	4.38	3.05	2.31
Mocha-Stereo (2024)	3.83	4.50	3.08	2.28
IGEV++ (2025)	3.71	4.35	3.00	2.22
GEAStereo (Ours)	**3.42**	**3.94**	**2.76**	**2.11**

Table 3: Results of the reflective regions on KITTI2012.

Method	Edge		Non-Edge	
	EPE	>1px	EPE	>1px
RAFT-Stereo (2021)	3.21	29.16	0.53	6.53
IGEV (2023)	2.23	20.42	0.41	4.58
Selective-IGEV (2024c)	2.18	20.01	0.38	4.35
DEFOM-Stereo (2025)	1.82	17.80	0.38	4.55
GEAStereo (Ours)	**1.50**	**15.54**	**0.33**	**4.00**

Table 4: Comparison on Scene Flow test set in different regions

Experiment

We implement our model using the PyTorch framework and adopt the AdamW (Loshchilov and Hutter 2019) optimizer with gradient clipping in the range of [-1, 1]. We use a one-cycle learning rate schedule with an initial learning rate of 2e-4. Our model is first pre-trained for 200k iterations on the Scene Flow (Mayer et al. 2016) dataset with batch size of 8.

Benchmark Performance

To demonstrate the effectiveness and superiority of our method, we evaluate it on widely used datasets including Scene Flow (Mayer et al. 2016), KITTI2015 (Menze and Geiger 2015), and ETH3D (Schöps et al. 2017) benchmarks.

Scene Flow. As shown in Table 1, our method achieves an EPE of 0.40 on the Scene Flow dataset, surpassing Selective-Stereo (Wang et al. 2024c) by 9.1%. With only 4% of the parameters of DEFOM-Stereo (Jiang et al. 2025), our EPE metric acheieves an improvement of 4.8%.

ETH3D. As shown in Table 2, our method achieves the best performance on most metrics among all published methods. Compared with the SOTA method AIO-Stereo (Zhou et al. 2025), our error rates for Bad 0.5 and Bad 1.0 in non-occluded regions are reduced by 33.3% and 37.2%, respectively. Moreover, our model does not incorporate the substantial extra parameters introduced by Vision Foundation Model. Compared with MC-Stereo (Feng et al. 2024), our method achieves an improvement of 44.3% on the Bad 0.5 metric for all pixels.

KITTI2015. Our method achieves competitive results on the KITTI2015 (Menze and Geiger 2015). We reduces the background error rate D1 in non-occluded regions by 4.9% compared with AIO-Stereo (Zhou et al. 2025).

Performance in ill-posed Regions

To demonstrate that our method effectively improves the disparity prediction performance in ill-posed regions, we conducted comparisons in representative ill-posed regions, such as reflective areas, edges and non-edge regions.

Reflective Regions. We evaluate our method on reflective regions of the KITTI2012 (Geiger, Lenz, and Urtasun 2012) test set. Our method achieves the best performance on almost all metrics as shown in Table 3. On the Out-4(All) metric, our method surpasses Selective-IGEV (Wang et al. 2024c) and IGEV++ (Xu et al. 2025) by 9.5% and 8.0%, respectively. Compared to the SOTA method Mocha-Stereo (Chen et al. 2024), our approach improves Out-3(Noc) and Out-5(All) metrics by 10.7% and 7.5%. The binocular geometric constraints are no longer applicable in reflective regions. However, we introduce monocular disparity priors to stereo matching, which can recognize reflective areas through texture, thus eliminating mismatches. As illustrated in Figure 5, GEAStereo demonstrates superior visualization performance, particularly in occlusions and reflective regions.

Edge/Non-Edge Region. To validate our method's performance in edge and low-texture areas, we use the Canny operator to split the Scene Flow (Mayer et al. 2016) test set into edge and non-edge regions and evaluate our method separately. As shown in the Table 4, compared with the baseline method Selective-IGEV (Wang et al. 2024c), our method achieves a 31.2% and 13.2% improvement in EPE on edge and non-edge regions, respectively.

Zero-shot Performance

Generating accurate disparity maps for real images is a major challenge in stereo matching. Therefore, the generalization ability of the model is crucial. We follow (Cheng et al. 2025) and employ DepthAnythingV2 (Yang et al. 2024) as feature encoder, denoted as GEAStereo-DFM. Since we do not need to generate depth maps, depth decoder is discarded, reducing the network parameters compared with SOTA methods(Cheng et al. 2025; Jiang et al. 2025).

We directly employ the pre-trained model on the Scene Flow dataset without any fine-tuning and conduct tests on the MiddleBury (Scharstein et al. 2014) and ETH3D

Method	Middlebury half >2px	Middlebury quarter >2px	ETH3D >4px
RAFT-Stereo (2021)	12.59	8.10	0.70
DLNR (2023)	9.82	7.82	21.35
IGEV (2023)	13.36	8.82	0.56
Selective-IGEV (2024c)	13.27	9.82	0.66
Mocha-Stereo (2024)	11.49	7.39	0.28
DEFOM-Stereo (2025)	5.91	<u>5.65</u>	<u>0.27</u>
MonSter (2025)	<u>4.06</u>	6.72	0.35
GEAStereo-DFM (Ours)	**3.92**	**4.58**	**0.22**

Table 5: Zero-shot evaluation on Middlebury and ETH3D.

Model	EPE(px)	>3px(%)	Params(M)
Baseline	0.441	2.42	13.12
+**D**AV	0.423	2.32	13.41
+**M**SFV	0.411	2.27	14.68
Full(+**D**+**M**)	**0.398**	**2.22**	14.84

Table 6: Ablation Study of overall framework.

(Schöps et al. 2017) datasets.

By learning monocular disparity distribution rather than monocular depth, our approach avoids the scale mapping error between disparity and inverse depth across different datasets. Our method achieves best performance compared with SOTA methods across all metrics.

As shown in the Table 5, our method reduces the Middlebury error rate by 18.9% at quarter resolution and the ETH3D error rate by 18.5% compared to the previous best method DEFOM-Stereo (Jiang et al. 2025). Compared to baseline method (Wang et al. 2024c), our method reduces the error rate by up to 70% on the Middlebury dataset at half resolution. The visualization results in Figure 2 show that our method delivers better performance in detailed and reflective regions, demonstrating strong generalization even in unseen scenes.

Ablation Study

To demonstrate the effectiveness of our proposed method, we take Selective-IGEV (Wang et al. 2024c) as the baseline and conduct ablation studies on the Scene Flow dataset, as shown in Table 6.

Mono-Stereo Fusion Volume (MSFV). Compared with the baseline, our MSFV achieves a 6.8% improvement in EPE, as MSFV leveraging monocular disparity distribution prior to correct the matching cost in ill-posed regions.

Further ablation of MSFV is shown in Table 7. To demonstrate the effectiveness of DPV, we first remove the intermediate supervision of MDPV (w/o Mono Loss), the network fails to establish the relationship between the monocular texture information and the disparity distribution, thus resulting in a performance decline compared to Full (MPF). Subsequently, we remove MDPV (w/o MDPV) and fuse the two stereo cost volumes, due to the absence of the monocular prior, there is still no matching cost in ill-posed regions, leading to poor performance.

Module	EPE(px)	>3px(%)	Params(M)
Ablation on Mono Disparity Prior Volume			
w/o Mono Loss	0.446	2.45	14.68
w/o MDPV	0.440	2.43	13.22
Ablation on Fusion			
Add	0.455	2.47	14.57
Concat&3D-CNN	0.442	2.44	14.60
MPF w/o FGF	0.420	2.30	14.64
Full (MPF)	**0.411**	**2.27**	14.68

Table 7: Ablation Study of MSFV.

Meanwhile, we evaluate different strategies for fusing MDPV and stereo cost volume. Firstly, we simply add MDPV and cost volume for fusion. However, due to the lack of alignment between the prior information and matching costs, it would destroy the matching information in the stereo cost volume, leading to a decline in accuracy. Then we directly concatenate MDPV and the stereo cost volume, employing a 3D CNN block for fusion (Concat & 3DCNN). Nevertheless, the limited receptive field of CNNs constrains the model's ability to capture global relationships across different regions, while also introducing additional noise. Subsequently, we remove the forget gate filtering from the MPF (MPF w/o FGF). In this case, the fusion module is able to utilize the monocular disparity distribution prior to recover matching costs in occluded regions, but it still introduces erroneous priors in regions with accurate matching costs.

Finally, after adding the forget gate mechanism to MPF, we effectively preserve the precise matching costs from stereo cost volume in normal regions, thereby achieving the best performance.

Detail-Aware Volume (DAV). As shown in the Tab.6, by introducing gradient information during the aggregation process of the group correlation volume, we reduce EPE (from 0.441 to 0.423), a 4.1% improvement. To further verify the effectiveness of the DAV, we replace DAA with 3D Aggregation (Xu et al. 2023). Quantitative results show that DAV can restore more detailed structures in disparity map.

Conclusion

In this paper, we propose GEAStereo, a novel stereo matching method. GEAStereo aims to address the problem of mismatching in ill-posed regions such as occlusions and reflections. By integrating Monocular Disparity Distribution Prior, MSFV effectively improves the correlation of the stereo cost volume and provides global structural guidance for ill-posed regions. Meanwhile, DAV utilizes the rich detail information in gradient features, enabling us to recover more refined disparity structures in detailed regions. Our method achieves the best performance on multiple benchmarks including ETH3D, KITTI, and Scene Flow, especially outperforming existing methods in zero-shot generalization.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China under Grant 42201513, and in part by the China Postdoctoral Science Foundation under Grant 2022M723902 and Grant 2023T160789.

References

Bangunharcana, A.; Cho, J. W.; Lee, S.; Kweon, I. S.; Kim, K.-S.; and Kim, S. 2021. Correlate-and-Excite: Real-Time Stereo Matching via Guided Cost Volume Excitation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3542–3548.

Chang, J.-R.; and Chen, Y.-S. 2018. Pyramid Stereo Matching Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5410–5418.

Chen, Z.; Long, W.; Yao, H.; Zhang, Y.; Wang, B.; Qin, Y.; and Wu, J. 2024. MoCha-Stereo: Motif Channel Attention Network for Stereo Matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27768–27777.

Cheng, J.; Liu, L.; Xu, G.; Wang, X.; Zhang, Z.; Deng, Y.; Zang, J.; Chen, Y.; Cai, Z.; and Yang, X. 2025. MonSter: Marry Monodepth to Stereo Unleashes Power. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 6273–6282.

Feng, M.; Cheng, J.; Jia, H.; Liu, L.; Xu, G.; and Yang, X. 2024. MC-Stereo: Multi-Peak Lookup and Cascade Search Range for Stereo Matching. In *2024 International Conference on 3D Vision (3DV)*, 344–353.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3354–3361.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

Guo, X.; Yang, K.; Yang, W.; Wang, X.; and Li, H. 2019. Group-Wise Correlation Stereo Network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3268–3277.

Han, D.; Wang, Z.; Xia, Z.; Han, Y.; Pu, Y.; Ge, C.; Song, J.; Song, S.; Zheng, B.; and Huang, G. 2024. Demystify Mamba in Vision: A Linear Attention Perspective. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 127181–127203. Curran Associates, Inc.

Hirschmuller, H. 2008. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2): 328–341.

Jiang, H.; Lou, Z.; Ding, L.; Xu, R.; Tan, M.; Jiang, W.; and Huang, R. 2025. DEFOM-Stereo: Depth Foundation Model Based Stereo Matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 21857–21867.

Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 5156–5165. PMLR.

Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; and Bry, A. 2017. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 66–75.

Kim, S. Y.; Aberman, K.; Kanazawa, N.; Garg, R.; Wadhwa, N.; Chang, H.; Karnad, N.; Kim, M.; and Liba, O. 2022. Zoom-to-Inpaint: Image Inpainting With High-Frequency Details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 477–487.

Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; and Liu, S. 2022. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16242–16251.

Lipson, L.; Teed, Z.; and Deng, J. 2021. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. In *2021 International Conference on 3D Vision (3DV)*, 218–227.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4040–4048.

Menze, M.; and Geiger, A. 2015. Object scene flow for autonomous vehicles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3061–3070.

Ning, C.; and Gan, H. 2023. Trap Attention: Monocular Depth Estimation With Manual Traps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5033–5043.

Pang, J.; Sun, W.; Ren, J. S. J.; Yang, C.; and Yan, Q. a. 2017. Cascade Residual Learning: A Two-stage Convolutional Neural Network for Stereo Matching. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 878–886.

Qiao, X.; Ge, C.; Zhang, Y.; Zhou, Y.; Tosi, F.; Poggi, M.; and Mattoccia, S. 2023. Depth super-resolution from explicit and implicit high-frequency features. *Computer Vision and Image Understanding*, 237: 103841.

Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; and Westling, P. 2014. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In Jiang, X.; Hornegger, J.; and Koch, R., eds., *Pattern Recognition*, 31–42. Cham: Springer International Publishing.

Schöps, T.; Schönberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2538–2547.

Teed, Z.; and Deng, J. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. *Computer Vision – ECCV 2020*, 402–419.

Wang, M.; Liu, J.; Luo, G.; Wang, S.; Wang, W.; Lan, L.; Wang, Y.; and Nie, F. 2025a. Smooth-Guided Implicit Data Augmentation for Domain Generalization. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3): 4984–4995.

Wang, M.; Liu, Y.; Yuan, J.; Wang, S.; Wang, Z.; and Wang, W. 2024a. Inter-Class and Inter-Domain Semantic Augmentation for Domain Generalization. *IEEE Transactions on Image Processing*, 33: 1338–1347.

Wang, M.; Su, H.; Wang, S.; Wang, S.; Yin, N.; Shen, L.; Lan, L.; Yang, L.; and Cao, X. 2025b. Graph Convolutional Mixture-of-Experts Learner Network for Long-Tailed Domain Generalization. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7): 6936–6947.

Wang, M.; Wang, S.; Yang, X.; Yuan, J.; and Zhang, W. 2024b. Equity in Unsupervised Domain Adaptation by Nuclear Norm Maximization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 5533–5545.

Wang, X.; Xu, G.; Jia, H.; and Yang, X. 2024c. Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19701–19710.

Wang, Z.; Yan, Z.; and Yang, J. 2024. SGNet: Structure Guided Network via Gradient-Frequency Awareness for Depth Map Super-resolution. In *38th AAAI Conference on Artificial Intelligence, AAAI 2024*, 5823–5831.

Xu, G.; Cheng, J.; Guo, P.; and Yang, X. 2022. Attention Concatenation Volume for Accurate and Efficient Stereo Matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12971–12980.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023. Iterative Geometry Encoding Volume for Stereo Matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21919–21928.

Xu, G.; Wang, X.; Zhang, Z.; Cheng, J.; Liao, C.; and Yang, X. 2025. IGEV++: Iterative Multi-Range Geometry Encoding Volumes for Stereo Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8): 7108–7122.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything V2. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 21875–21911. Curran Associates, Inc.

Zhao, H.; Zhou, H.; Zhang, Y.; Chen, J.; Yang, Y.; and Zhao, Y. 2023. High-Frequency Stereo Matching Network. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1327–1336.

Zhou, J.; Zhang, H.; Yuan, J.; Ye, P.; Chen, T.; Jiang, H.; Chen, M.; and Zhang, Y. 2025. All-in-One: Transferring Vision Foundation Models into Stereo Matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10): 10797–10805.