Global Occlusion-Aware Transformer for Robust Stereo Matching

Zihua Liu[1], Yizhou Li[2], and Masatoshi Okutomi[3]
Tokyo Institute of Technology, Japan
{zliu[1],yli[2]}@ok.sc.e.titech.ac.jp,mxo@ctrl.titech.ac.jp[3]

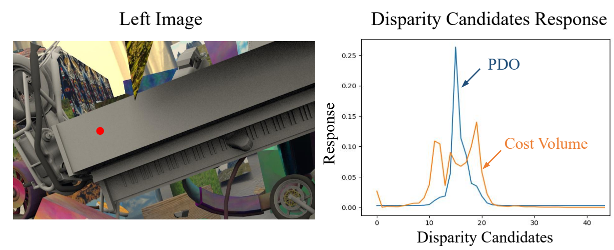
Abstract

Despite the remarkable progress facilitated by learning-based stereo-matching algorithms, the performance in the ill-conditioned regions, such as the occluded regions, remains a bottleneck. Due to the limited receptive field, existing CNN-based methods struggle to handle these ill-conditioned regions effectively. To address this issue, this paper introduces a novel attention-based stereo-matching network called Global Occlusion-Aware Transformer (GOAT) to exploit long-range dependency and occlusion-awareness global context for disparity estimation. In the GOAT architecture, a parallel disparity and occlusion estimation module (PDO) is proposed to estimate the initial disparity map and the occlusion mask using a parallel attention mechanism. To further enhance the disparity estimates in the occluded regions, an occlusion-aware global aggregation module (OGA) is proposed. This module aims to refine the disparity in the occluded regions by leveraging restricted global correlation within the focus scope of the occluded areas. Extensive experiments were conducted on several public benchmark datasets including SceneFlow [15], KITTI 2015 [16], and Middlebury [19]. The results show that proposed GOAT demonstrates outstanding performance among all benchmarks, particularly in the occluded regions.

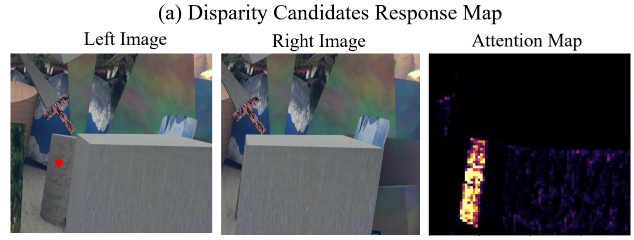
1. Introduction

Stereo-matching is one of the most fundamental tasks in computer vision. It is to infer depth from a given pair of stereo images taken by a binocular camera, which is closely related to applications like robotic navigation [17], autonomous driving [41], augmented reality [25], and so on.

Recently, the rapid development of convolutional neural networks (CNNs) has improved the performance of stereo-matching algorithms [9, 10, 15, 33, 44] significantly. Typical CNN-based methods commonly rely on a cost volume, which is constructed with a predetermined search range to evaluate the matching similarity. Existing cost volume-based stereo matching can be categorized as the



(a) Disparity Candidates Response Map



(b) Global Attention at the Occlusion Region

Figure 1. **(a)** Visualization of estimated response for disparity candidates using proposed *PDO*. Compared with a cost volume method (orange), the *PDO* (blue) can alleviate matching ambiguity in texture-less regions and show a single peak waveform. **(b)** Visualization of global attention map in the occluded regions using the proposed *OGA*.

3D correlation-volume-based methods [14, 32, 44] and the 4D concatenation-volume-based methods [1, 7, 9, 37, 43]. However, these methods perform poorly when applied in ill-conditioned regions like occluded regions, and texture-less regions.

The challenges associated with stereo matching in ill-conditioned regions can be simply summarized as follows: (1) Texture-less or repetitive regions show homogeneity in the RGB domain, which is difficult for CNN-based methods to extract distinguishable local matching features. (2) Occluded regions, which naturally lack matching correspondences and cannot be estimated by matching directly. Most methods [3, 18, 34] use CNN-based spatial propagation to refine the disparity in the occluded regions using the contextual features as a guide. However, these CNN-based networks reliant on local windows exhibit a tendency to utilize the limited receptive field information from the surrounding area for disparity refinement, which leads to limited improvement in large and irregular occluded regions. Other methods in optical flow tasks like GMA [8] use global attention instead of local correlations for the ill-conditioned region's refinement, while uncontrolled global attention is inefficient and can even affect well-conditioned areas.

In order to improve the disparity performance in the ill-conditioned regions, in this paper, we propose to leverage restricted global spatial correlation as a guide to alleviate matching ambiguities in texture-less regions and refine the disparity in occluded regions. Our idea is that disparity within a bounded region (e.g. an object) should be continuous. To realize this, we propose the Global Occlusion-Aware Transformer (*GOAT*) which introduces Vision Transformer [6] and attention mechanism to establish restricted global spatial correlation for both the matching and disparity refinement phases. In *GOAT*, a parallel disparity and occlusion estimation module (*PDO*) is proposed to estimate the initial disparity and the occlusion mask respectively with an adaptive global search range utilizing stacked self-cross attention layers for feature aggregation and parallel cross-attention for occlusion and disparity estimation. The most related prior work is the STTR [11], however, STTR employs a shared cross-attention matrix for estimating both disparity and occlusion, which leads to a trade-off between disparity and occlusion prediction accuracy. In contrast, the proposed *PDO* infers occlusion and disparity independently, eliminating any possible trade-offs between the two estimates. To further enhance the disparities in the occluded regions, an iterative occlusion-aware global aggregation module (*OGA*) is proposed to refine the disparity with a restricted focus scope of the occluded regions using global spatial correlations and context guidance.

Our main contributions lie in four folds:

- We explore employing restricted global spatial correlation information for stereo-matching and propose a novel stereo-matching network named *GOAT*, which enables robust disparity estimation, particularly in ill-conditioned regions.
- We propose a parallel disparity and occlusion estimation module (*PDO*) that utilizes a parallel attention mechanism to generate both disparity and occlusion masks robustly, without mutual interference.
- We also propose an occlusion-aware global aggregation module (*OGA)* that aggregates feature with a focus scope in occluded regions using self-attention, boosting disparity estimation in occluded areas.
- We conducted extensive experiments on several public datasets including SceneFlow [15], FallingThings [30], KITTI 2015 [16], and Middlebury [19]. Experimental results reveal that the proposed method achieves outstanding performance on several benchmark datasets, especially in the ill-conditioned occluded regions.

2. Related Works

Cost-Volume-based Methods. Pioneer work DispNetC [15] utilizes a correlation layer to calculate the inner product of the left and right features at each disparity level for measuring the similarity. Although correlation volume has been proven to be effective and efficient, the loss of context information during correlation limits the ultimate performance of stereo-matching. GCNet [9] firstly employs the concatenation of left and right features to construct a 4D volume that encodes abundant content information for similarity measurement. The concatenation volume following stacked 3D convolution networks for aggregation is widely used in most latest state-of-the-art works including [20, 33, 43]. In order to combine the advantages of the correlation volume and the concatenation volume, GwcNet [7] adopts a group-wise correlation method to combine the correlation volume and the concatenation volume. Later work such as PCWNet [22] follows the same architecture and exploits multi-scale volumes fusion to extract domain-invariant features, which leads to better performance.

Guidance-Incorporated Stereo Matching. Besides, depending on image similarity for stereo matching, some other methods utilize extra guidance information to improve stereo matching and achieve exceptional performance. Xiao et al propose a multi-task network called EdgeStereo [24] by applying a disparity-edge joint learning framework to leverage edge maps as the guidance for disparity refinement. Wu *et al.* [36] employ semantic guidance by introducing a designed pyramid of cost volumes for describing semantic and spatial information on multiple levels. Liu *et al.* [14] propose a normal incorporated joint learning framework to explicitly leverage the surface normal as an intuitive geometric guidance to refine the ill-conditioned regions with the surface normal affinities. Although stereo-matching with guidance information is able to introduce prior knowledge beyond RGB clues for robust stereo-matching, the implementation of these approaches requires a joint-learning framework with additional supervision, which may increase the complexity and training cost of the network.

Attention Mechanism in Stereo Matching. Recently, attention mechanisms have been introduced in the stereo-matching task to improve the quality of disparity estimation. Many works [2, 45, 46] use 2D attention block for left-right feature aggregation to adaptively calibrate weight response, improving the robustness of the feature representation. Zhang *et al.* [44] use a warped photometric error to generate a spatial attention mask for disparity

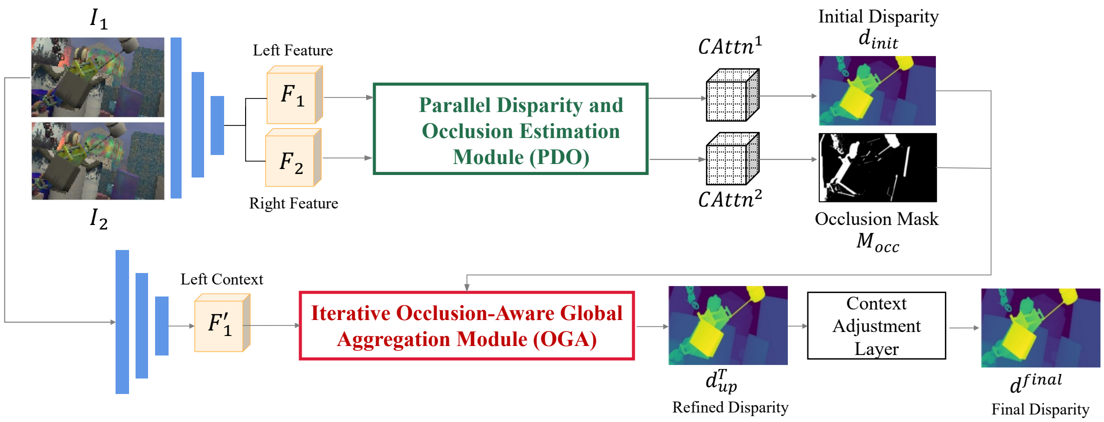


Figure 2. Overall architecture of Global Occlusion-Aware Transformer (GOAT).

residual estimation which accelerates the training process. ACVNet [37] learns an attention map from the correlation volume to suppress redundant information and enhance matching-related information in the concatenation volume. Besides, other works use an attention mechanism to replace the conventional cost volume for left-right image matching. STTR [11] takes the first attempt to use alternating self-cross attention modules to estimate the disparity and corresponding occlusion mask from an aspect of the transformer. GMStereo [40] presents a unified formulation using a cross-attention mechanism for three motion and 3D perception tasks: optical flow, rectified stereo matching, and unrectified stereo depth estimation from posed images.

3. Proposed Method

In this section, we provide a comprehensive introduction to our proposed Global Occlusion-Aware Stereo Transformer (*GOAT*). The overall architecture of the proposed work is presented in Subsection 3.1, with detailed descriptions of the proposed two specific modules provided in Subsections 3.2 and 3.3. The training mechanism and loss function are expounded upon in Subsection 3.4.

3.1. Overall Network Architecture

The overall architecture of the proposed *GOAT* is shown in Figure 2. We decouple the stereo-matching process into matching for non-occluded regions and disparity refinement for occluded regions. In the matching phase, we propose a parallel disparity and occlusion estimation module (*PDO*) which leverages both positional and global correlations between the left and right views to estimate initial disparity and the occlusion mask, respectively. In the refinement phase, we propose an iterative occlusion-aware global aggregation module (*OGA*) using restricted global correlation with occlusion guidance to optimize the disparity within the occluded regions. Finally, a context adjustment layer is employed to refine the disparity from a mono-depth aspect.

3.2. Parallel Disparity and Occlusion Estimation Module (PDO)

Instead of using a cost volume with a predetermined search, we proposed a global-attention-based module named *PDO* to compute the initial disparity and the occlusion mask. As illustrated in Figure 3, After obtaining the F_1 and $F_2 \in \mathbb{R}^{H\times W\times C}$ from the shared image extractor, we follow the architecture in [26] by introducing a self-cross alternating module to extract global context information and position bias, where the Swin-Transformers Blocks [13] with a window size of [h/2,w/2] are utilized for efficient feature aggregation. The self-cross attention module can be described as follows:

$$F_l = softmax(\frac{Q_l K_l^T}{\sqrt{C}})V_l, F_r = softmax(\frac{Q_r K_r^T}{\sqrt{C}})V_r,$$
$$F_l = softmax(\frac{Q_r K_l^T}{\sqrt{C}})V_l, F_r = softmax(\frac{Q_l K_r^T}{\sqrt{C}})V_r, \quad (1)$$

where the first row represents the self-attentions of the left feature and right feature, while the second row represents the cross-attentions between two views. Q, K, and V are obtained using a shared-weight linear projection layer with absolute positional encoding to indicate the position information. The alternating self-cross attention modules use the global receptive field to fully aggregate the information of the left and right views, resulting in more representative and distinguishable features. In addition, positional encoding helps to constrain the aggregation range and prevent aggregating features from distant and unrelated regions with similar textures. Once we obtained the aggregated left and right features, a parallel cross-attention module was applied to estimate the initial disparity and the occlusion mask. As illustrated in Figure 3, we conduct parallel cross-attention between the left feature and right feature and get two cross-attention matrices $CAttn^1 \in \mathbb{R}^{H\times W\times W}$ and $CAttn^2 \in \mathbb{R}^{H\times W\times W}$. Since the normalized cross-attention reflects the similarity of left and right features, the $CAttn^1$ can be regarded as a cost volume with a global

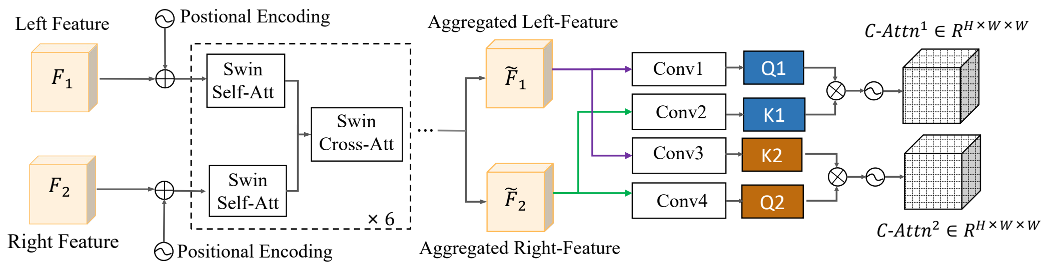


Figure 3. Parallel Disparity and Occlusion Estimation Module Architecture. (PDO)

search range. Besides, since occluded regions lack a corresponding pixel in the other image view, the summation of attention values of potential matching pixels for occlusion regions in $CAttn^2$ should yield a low response. According to the characteristics of these two cross-attentions, we compute the initial disparity and the occlusion mask:

$$\begin{aligned} \text{disp}(i,j) &= \text{Coord}x^{L(i,j)} - CAttn^1 \otimes \text{RC}, \\ \text{occlusion}_{(i,j)} &= \text{sigmoid}(f_\theta(\sum_{k=1}^{W} CAttn^2_{(i,k,j)})), \end{aligned} \tag{2}$$

where $\text{Coord}x^L(i,j) \in \mathbb{R}^{H\times W\times 1}$ is the standard coordinate of the left image in the horizontal direction, RC $= [0,1,\dots,W-1]^T$ is the range of all potential corresponding coordinates in the right image, and $\otimes$ denotes matrix multiplication. f_θ represents a small network that takes the summation of attention values of all potential matching points in $CAttn^2$ as input to regress the occlusion mask.

One related work is [31], which utilizes features extracted by a CNN for cross-attention to obtain the matching matrix for unsupervised stereo matching. However, it lacks the global context and positional encoding information introduced by alternating self-cross attention. As a result, the proposed *PDO* module is more powerful in modeling texture-less and occluded regions compared to [31].

3.3. Iterative Occlusion-Aware Global Aggregation Module (OGA)

After obtaining the initial disparity and occlusion mask at low resolution, the disparities in ill-conditioned regions, such as occluded areas, remain problematic, since they are difficult to estimate accurately via matching alone. To further enhance the disparity estimation performance, we propose an iterative refinement module based on self-attention, namely *OGA* module, which aggregates features from valid non-occluded regions into invalid occluded regions using global spatial correlation. Similar to RAFT [29], a convex upsampling layer is used to upsample the disparity to a higher resolution. The overall structure of the *OGA* module is shown in Figure 4.

The input of the *OGA* module is the disparity d^{t-1} of stage $t-1$ as well as the left context F'_1 extracted from a CNN. We also construct a local cross-attention that measures the similarity between the left and right features around the d^{t-1} with a search range of r by sampling from the cross-attention matrix $CAttn^1$ in the *PDO* module. The current disparity d^{t-1} and its corresponding local cross-attention are then passed to a disparity encoder to obtain the matching feature $F^t_{matching}$. Meanwhile, the left context F'_1 is further concatenated with $F^t_{matching}$ to supplement local feature F^t_{local} from a mono-depth aspect. Such information is sufficient for disparity optimization in the non-occluded regions. As for occluded regions, we calculate the global spatial correlation of the left image through the self-attention module and obtain a self-attention matrix $A \in \mathbb{R}^{H\times W\times H\times W}$. For arbitrary specific point (i,j), we obtain its correlation with all other pixels in the left view by consulting the attention map $A_{i,j} \in \mathbb{R}^{H\times W}$. Then, we perform feature aggregation to derive global feature F^t_{global}. With local feature F^t_{local} and global feature F^t_{global} obtained, we then adopt an occlusion-aware global aggregation mechanism as shown in Figure 4. We reserve the local feature at the non-occluded region and keep the global feature at the occluded region to generate an adaptive feature F^t_{ada} for overall disparity refinement. On the one hand, local features are sufficient for non-occluded regions to perform disparity refinement. On the other hand, we can prevent the local features of occluded regions, which are less confident because of the matching ambiguity, from propagating to non-occluded regions through the attention map like [8]. This can effectively reduce the degradation of features. Therefore, the proposed *OGA* module can make good use of the global spatial correlations at the ill-conditioned regions as well as avoid harmful propagation. The whole process can be described as follows:

$$\begin{aligned} F^t_{ada} &= A \otimes F^t_{local} \odot M_{occ} + F^t_{global} \odot (I - M_{occ}), \\ F^t_{local} &= concat(F^t_{matching}, F^{'}_1), \end{aligned} \tag{3}$$

where M_{occ} indicates the occlusion mask, I is an identity matrix, and $\odot$ denotes element-wise multiplication. After feature aggregation, we employ a GRU [5] unit to regress the disparity residual and an upsample mask, where we compute the disparity d^t at the current iteration and use the

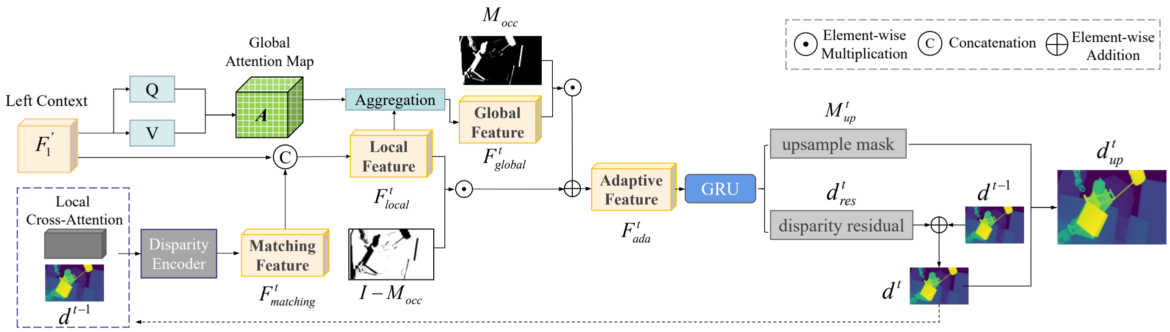


Figure 4. Iterative Occlusion-Aware Global Aggregation Module (OGA).

upsample mask to increase the resolution:

$$\begin{aligned} d_{res}^t, M_{up}^t &= GRU(F_{ada}^t), \\ d^t &= max(0, d_{res}^t + d^{t-1}), \\ d_{up}^t &= d^t * M_{up}^t, \end{aligned} \tag{4}$$

where the $M_{up}^t \in \mathbb{R}^{H \times W \times S \times S}$, S is the upsample scale, and $*$ denotes convolution. The upsampled disparity d_{up}^T of the last iteration T is further passed to a context adjustment layer [11] to derive final disparity d^{final}, which recovers fine-grained disparity details from a mono-depth aspect. This layer utilizes the left image and the current disparity map to regress the disparity residual.

3.4. Occlusion and Disparity Supervision

We supervised the network with groundtruth disparity and occlusion mask. Since the *GOAT* is an iterative network, we follow the sequence loss proposed in [29] to supervise the disparity at different iterations, which is the l_1 distance between the ground truth disparity and the estimated disparity at each iteration with exponentially increasing weights. The loss can be defined as follows:

$$L_{disp} = \sum_{i=0}^{T} \gamma^{T-t} \left\| d^{gt} - d_{up}^t \right\| + \left\| d^{gt} - d^{final} \right\|, \tag{5}$$

where the T is the iteration number which in our case equals 12 and set the increasing weight γ to 0.95. For occlusion supervision, the cross-entropy loss is deployed for effective training:

$$L_{occ} = -\frac{1}{2} \sum_{i}^{2} (O_{gt} \log(O_i) + (1 - O_{gt}) \log(1 - O_i)). \tag{6}$$

The final loss is the weight summation of disparity loss and occlusion loss.

$$L_{total} = \lambda_1 \times L_{disp} + \lambda_2 \times L_{occ}. \tag{7}$$

4. Experimental Results

4.1. Datasets

We evaluate our method on multiple public benchmark datasets including SceneFlow [15], Falling Things [30] KITTI 2015 [16], and Middlebury [19]. As the proposed network requires ground-truth occlusion masks for training, which are not provided in the several datasets, we generate the ground-truth occlusion masks using left-right consistency. More details can be seen in our *supplementary material*. The SceneFlow dataset is a synthetic dataset containing 39,823 stereo image pairs with random flying objects. The Falling Things dataset is another synthetic dataset with more realistic indoor scenes. The KITTI 2015 dataset comprises real-world scenes that have sparse ground-truth disparity captured using LiDAR. For the Middlebury dataset, the evaluation is conducted using the standard Middlebury Stereo Evaluation-Version 3.

4.2. Implementation Details

We implemented our *GOAT* network by PyTorch trained with 4 NVIDIA 3090 GPUs. For the SceneFlow dataset, we trained the networks for 80 epochs using a batch size of 8 with an initial learning rate of 4e-4 following a step learning rate decay strategy. For the Falling Things dataset, we trained for 10 epochs with a constant learning rate of 4e-4. Compared to SceneFlow dataset, the Falling Things dataset [30] has enhanced scene realism and better semantics in occluded region, therefore we use it for more comprehensive ablation studies. For both above dataset, we randomly cropped the input images to 320×640. For the KITTI 2015 dataset, we fine-tune our networks with the SceneFlow pre-trained model. Mixed datasets of KITTI 2012 and KITTI 2015, totaling 400 image pairs, were used for the initial 400 epochs with a random crop size of 320×1088. The model with the best validation performance was chosen, followed by another 200 epochs of fine-tuning on the KITTI 2015 training set to obtain the final model. For the Middlebury dataset with only 23 images, we first evaluated generalization on the Middlebury training set using the SceneFlow pre-trained model, then fine-tuned it at half-resolution for benchmark assessment. Please refer to the *supplementary material* for more training details.

Table 1. Ablation study of our proposed *GOAT* network on the SceneFlow dataset. We conduct ablation studies on the proposed *PDO* and *OGA* modules. As well as compared with other attention-based disparity estimation and refinement modules like *STTR* [11] and *GMA* [8]. The '*' represents a higher resolution. We calculated the EPE and P1(outliers) both in the overall and the occluded regions separately.

Method	Disparity Estimation			Update Module			**CA Layer**	EPE		P1(%)		Occ mIOU
	Cost Volume	STTR	**PDO**	RAFT	GMA	**OGA**		All	Occ	All	Occ	
Baseline	✓			✓				0.79	2.27	9.2%	25.6%	-
STTR		✓		✓				0.78	2.31	10.0%	28.4%	0.81
PDO			✓	✓				0.65	1.96	7.2%	22.2%	0.83
PDO + GMA			✓		✓			0.62	1.86	7.0%	21.9%	0.83
PDO + OGA			✓			✓		0.57	1.78	6.7%	20.9%	0.83
PDO + OGA + CA(Full)			✓			✓	✓	0.55	1.72	6.6%	19.9%	0.83
PDO + OGA + CA*(Full)			✓			✓	✓	0.47	1.53	5.6%	18.6%	0.94

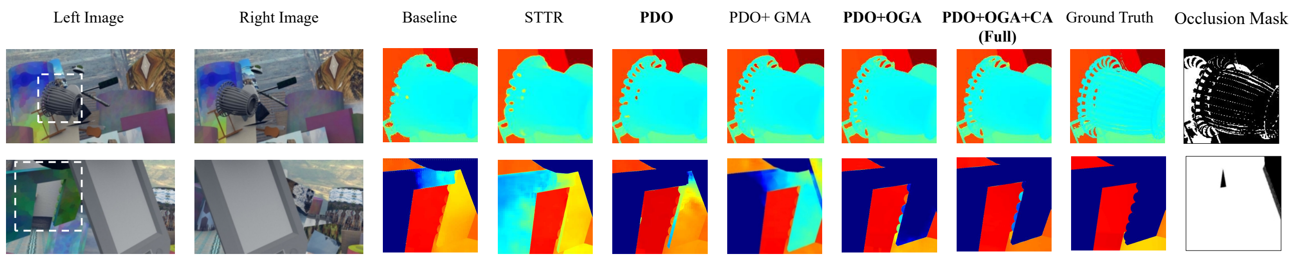


Figure 5. Visualizations of ablation study on SceneFlow dataset. We cropped and enlarged the selected part of the disparity map for easier viewing.

Table 2. Quantitative comparison of *GOAT* and other methods on the SceneFlow. We adopt the EPE-All results from the original papers. Due to incomplete disparity evaluation of the occluded regions in some works, we calculate EPE-Occ using the corresponding official pre-trained models. Proposed *GOAT* ranks top for overall and occluded regions. **Red Bold:Best**. **Bold:Second**.

Model	PSMNet [1]	AANet++ [39]	RaftStereo [12]	PCW-Net [22]	STTR-light [11]	ACVNet [37]	IGEVStereo [38]	**GOAT (Ours)**
EPE-All	1.09	0.72	0.69	0.86	4.14	**0.48**	**0.47**	**0.47**
EPE-Occ	3.14	2.44	2.14	2.54	23.9	1.65	**1.61**	**1.53**

4.3. Ablation Studies

We conducted ablation studies on the SceneFlow and Falling Things datasets. We report the standard end-point error (EPE) and P1-value (outliers) for overall regions (All) and occluded regions (Occ), respectively. For occlusion mask evaluation, we compute the mean Intersection over Union (mIoU) between the ground truth and the predicted occlusion mask. The relevant results of the Sceneflow dataset are shown in Table 1, where we use a simplified version of [12] as the Baseline. For more ablation studies in the FallingThings Datset, please refer to our *supplemenatry aterials*

Parallel Disparity and Occlusion Estimation Module (PDO): As depicted by Table 1, compared with the Baseline integrating the *PDO* module (designated as PDO) exhibits a remarkable improvement in terms of EPE for both overall and occluded regions. We also compared our proposed *PDO* modules with another transformer-based method by replacing the PDO module with a disparity estimation module proposed in STTR [11]. As demonstrated in Table 1, our *PDO* shows better disparity estimation performance with smaller errors, especially in the occluded regions, where the STTR-based method reveals even bigger EPE errors than the baseline. Further insight into the efficacy of the *PDO* module can be gained from the 1st row of Figure 5, which demonstrates that the PDO derives a more accurate structural representation of the object compared with Baseline and *STTR*, as *PDO* module reduces the matching ambiguity when dealing with the texture-less and occluded regions.

Iterative Occlusion-Aware Global Aggregation Module (*OGA*): Table 1 illustrates the effectiveness of the *OGA* module. Model with the *OGA* module, which is named as *PDO+OGA* can reduce the EPE in the occluded regions from 1.96 to 1.78 in the SceneFlow dataset with an improvement of 10.1%, which is more effective compared with naive global-attention-based *GMA* [8] module with an improvement of 5.1%. Moreover, The *OGA* module is also able to maintain the disparity at the non-occluded regions due to the restricted global attention mechanism. As depicted in Figure 5, the *PDO+OGA* shows less error and enhanced robustness in the occluded regions (marked by white boxes) compared to the *PDO* only and *PDO+GMA*. Besides, it also shows better disparity estimation at the non-

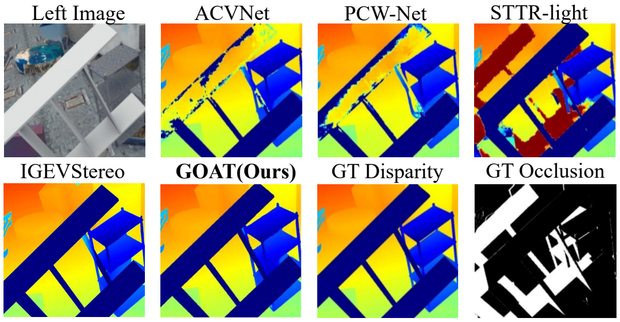


Figure 6. Qualitative comparison on SceneFlow dataset with other superior works.

occluded regions while the *PDO+GMA* fails to estimate well. Moreover, incorporating the context adjustment module into the whole, designated as *PDO+OGA+CA*, results in further improved performance.

Resolution: Like [12], we employed the *PDO* and *OGA* modules at both 1/8 and 1/4 resolutions. As shown in Table 1, increasing the resolution yields better performance, while consuming much bigger GPU memory for the self-attention computation.

4.4. Performance Evaluation

In this subsection, we compare our method with other top-performing methods using multiple datasets.

SceneFlow. For quantitative evaluation demonstrated in Table 2, our proposed method ranks at the top for occluded regions, surpassing all competing methods and even very recent state-of-the-art methods such as IGEVStereo [38] and PCW-Net [22]. Note that while IGEVStereo [38] requires 32 iterations for disparity refinement, our proposed *GOAT* achieves equivalent disparity performance in overall regions with only 12 iterations, and surpasses IGEVStereo [38] in occluded regions by a large margin. This further illustrates the advantages of our proposed GOAT in optimizing disparity in the occluded regions. For qualitative evaluation shown in Figure 6, proposed *GOAT* generates disparity maps with more detailed and precise structures in textureless areas. In contrast, other methods exhibit less satisfactory performance, with missing details and artifacts.

KITTI 2015. For KITTI dataset evaluation, we follow the standard protocol to submit our fine-tuned results to KITTI leaderboard [16]. Table 3 demonstrates the evaluation performance on the KITTI 2015 test set. In our assessment of overall (All) regions, including occluded areas, our method distinctly excels in its performance on foreground (fg) objects with key items like cars and pedestrians, achieving a D1-Error of 2.51. The results surpass very recent methods including PCWNet [22] and IGEVStereo [38]. Importantly, in the context of real-world autonomous driving applications, foreground regions like pedestrians and cars are of great importance, where our method proves notably proficient. As evidenced in Figure 8, for out-of-view regions marked by the red box which lack the corresponding pixels, our proposed *GOAT* still succeeds in estimating the disparity by showing better depth consistency and clearer structures. At the same time, other methods fail to generate sat-

Table 3. Benchmark results on KITTI 2015 test set. The **"Noc"** and **"All"** indicate the non-occluded and overall regions, respectively. The **"fg"** and **"all"** indicate the foreground and overall regions, respectively. The results report the percentage of outliers over the available ground truth disparities.

Method	Noc (%)		All (%)		Time (s)
	fg	all	fg	all	
GANet [43]	3.37	1.73	3.82	1.93	0.36
PSMNet [1]	4.31	2.14	4.62	2.32	0.41
GwcNet [7]	3.49	1.92	3.93	2.11	0.32
AANet [39]	4.93	2.32	5.39	2.55	0.075
DispNetC [15]	3.72	4.05	4.41	4.34	0.06
FADNet [32]	3.07	2.59	3.50	2.82	0.05
IGEVStereo [38]	**2.62**	1.49	**2.67**	**1.59**	0.83
HITNet [28]	2.72	1.74	3.20	1.98	0.02
LEAStereo [4]	2.65	1.51	2.91	1.65	0.30
RAFTStereo [12]	2.94	**1.45**	2.94	1.82	0.38
GMStereo [40]	2.97	1.61	3.14	1.77	0.38
CFNet [21]	3.25	1.73	3.56	1.88	0.38
ACVNet [37]	2.84	1.52	3.07	**1.65**	0.20
PCW-Net [23]	2.93	**1.26**	3.16	1.67	0.44
GOAT(Ours)	**2.43**	1.71	**2.51**	1.84	0.29

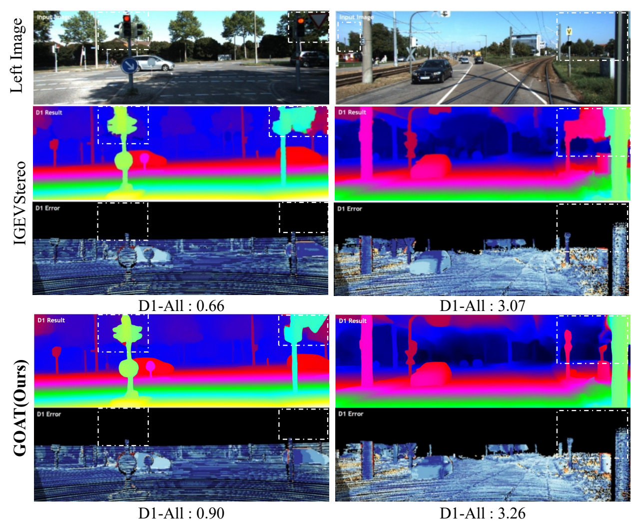


Figure 7. Visualization comparison on KITTI 2015 test set between IGEVStereo [38] and our *GOAT*. The 2^{nd} and 4^{th} line show estimated disparity maps, and the 3^{rd} and 5^{th} line display the corresponding errors. The error map indicates that colored regions have LiDAR annotation while black regions lack annotation, which means the **D1-All cannot fully represents the disparity estimation performance on the whole scene**. Although our model has a higher D1-All error, it exhibits improved structures and fewer artifacts in regions where the ground-truth disparity is missing.

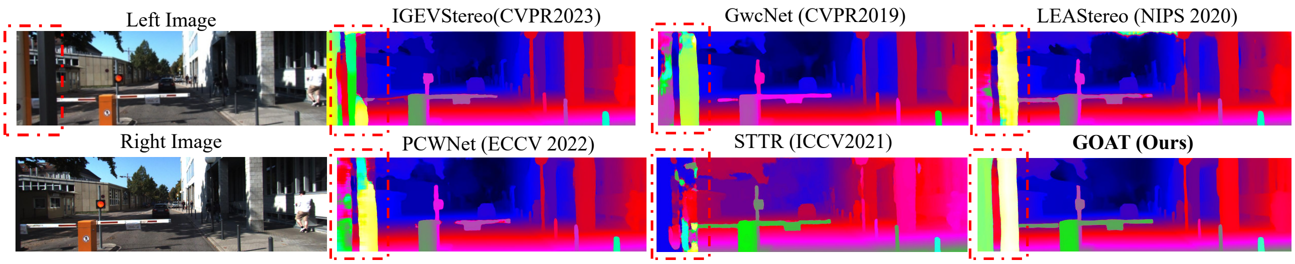


Figure 8. Performance on KITTI 2015 test set. Our method obviously exhibits better results in the severely occluded regions.

Table 4. Quantitative generalization evaluation on the Middlebury training dataset. **"Occ"** represents occluded regions, and **"Non"** represents non-occluded regions. Note the **Red Bold** means the best and the **Bold** means the second-best.

Method	AvgErr		RMSE		Bad 4.0		Bad 2.0	
	Occ	Non	Occ	Non	Occ	Non	Occ	Non
AANet [39]	9.9	5.5	15.3	10.8	39.5	28.2	56.4	28.3
PSMNet [1]	17.7	10.7	29.9	22.1	47.4	23.3	62.1	32.3
GwcNet [7]	10.3	6.3	17.6	13.7	34.1	15.1	47.9	21.9
ACVNet [32]	9.4	6.3	16.4	14.2	30.2	13.9	43.4	19.0
PCW-Net [22]	**7.7**	3.9	**14.9**	9.3	**26.5**	9.7	**39.1**	14.9
STTR-light [11]	35.2	**3.0**	47.7	10.2	74.7	**8.3**	82.0	**13.3**
RAFTStereo [12]	10.0	3.6	15.9	**9.1**	34.4	9.5	46.5	**14.4**
GOAT(Ours)	**5.7**	**2.0**	**9.4**	**5.3**	**28.0**	**9.2**	**43.3**	15.7

isfactory results.

It is noteworthy that the KITTI dataset lacks LiDAR ground truth for the upper portions of the images as shown in Figure 7, s.t. these parts of results are not evaluated in D1-All error. This lack of annotation may introduce bias into the final D1-All error, preventing a complete revelation of the network's effectiveness. Figure 7 illustrates this challenge by comparing the proposed *GOAT* with the most advanced IGEVStereo [38]. Although *GOAT* produces a larger D1-All error, the visualization results exhibit clearly better structures and fewer artifacts in regions where the ground-truth disparity is missing.

Middlebury. As the Middlebury dataset only includes 23 images for training, we first evaluate the generalization of the pre-trained SceneFlow model on the Middlebury training set with half resolution. As depicted in Table 4, the proposed *GOAT* generates the best-performing disparity map with the lowest AvgErr and RMSE compared to other methods. Especially in occluded regions, proposed *GOAT* outperforms the latest PCW-Net [22] by 26% in terms of AvgErr and 33.6% in terms of RMSE. Figure 9 shows the visual comparison. Besides, we also fine-tune our model on the Middlebury dataset with half resolutions (H) because of memory issues. As depicted in Table 5, compared with other competing methods submitted at the same resolution, our *GOAT* demonstrates state-of-the-art performance by showing the smallest AvgErr and RMSE. Please refer to the *supplementary material* for more results on datasets.

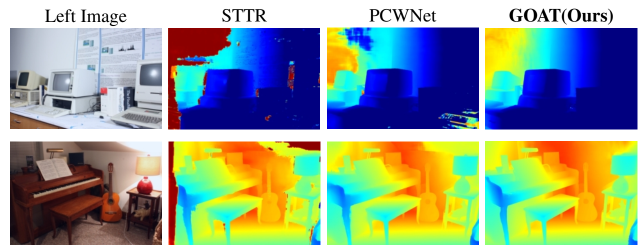


Figure 9. Generalization evaluation on Middlebury Dataset.

Table 5. Fine-Tuned Results on Middlebury Benchmarks with half resolution in **'all'** regions. **Red Bold: Best**, **Bold: Second**.

Method	AvgErr	RMSE	Bad 4.0	Bad 2.0
CFNet [21]	5.07	18.20	11.30	16.10
LEAStereo [4]	**2.89**	**13.70**	**6.33**	12.10
AANet++ [39]	9.77	24.90	16.40	22.00
NOSS_ROB [35]	4.80	19.80	8.37	**11.20**
LocalExp [27]	5.13	21.10	8.83	**11.30**
FADNet_RVC [27]	21.00	48.30	24.20	33.30
MC-CNN-acrt [42]	17.90	55.00	15.80	19.10
HITNet [28]	3.29	14.50	8.66	12.80
ACVNet [37]	12.10	38.60	12.60	19.50
GOAT (Ours)	**2.71**	**11.20**	**8.18**	13.80

5. Conclusions

In this paper, we have proposed a novel attention-based stereo-matching network called *GOAT* that exploits long-range dependency and global context for disparity estimation in ill-conditioned regions. The parallel disparity and occlusion estimation module *(PDO)* is proposed to estimate the initial disparity and the occlusion with a parallel attention mechanism, which improves the disparity estimation performance as well as provides the occlusion mask for further disparity refinement. The iterative occlusion-aware global aggregation module (*OGA*) uses a restricted global correlation with a focus scope marked by the occlusion mask to refine the disparity in the occluded regions. Extensive experiments on various datasets have demonstrated the effectiveness and generalization ability of the proposed method. By the time we finish this paper, our method outperforms recent state-of-the-art methods on the SceneFlow dataset and also ranks 1^{st} on the KITTI 2015 leaderboard for foreground objects.

References

[1] J. Chang and Y. Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 1, 6, 7, 8

[2] Jia-Ren Chang, Pei-Chun Chang, and Yong-Sheng Chen. Attention-aware feature aggregation for real-time stereo matching on edge devices. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2

[3] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 10615–10622, 2020. 1

[4] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22158–22169. Curran Associates, Inc., 2020. 7, 8

[5] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1597–1600. IEEE, 2017. 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[7] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3268–3277, 2019. 1, 2, 7, 8

[8] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021. 2, 4, 6

[9] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 66–75, 2017. 1, 2

[10] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 573–590, 2018. 1

[11] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6197–6206, 2021. 2, 3, 5, 6, 8

[12] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 6, 7, 8

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 3

[14] Zihua Liu, Songyan Zhang, Zhicheng Wang, and Masatoshi Okutomi. Digging into normal incorporated stereo matching. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 6050–6060, 2022. 1, 2

[15] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 1, 2, 5, 7

[16] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 1, 2, 5, 7

[17] Don Murray and James J Little. Using real-time stereo vision for mobile robot navigation. *Autonomous Robots*, 8:161–171, 2000. 1

[18] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 120–136. Springer, 2020. 1

[19] D. Scharstein, H. Hirschmuller, Y. Kitajima, G. Krathwohl, N. Nesic, and P. Westling X Wang. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR)*, 2014. 1, 2, 5

[20] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, 2021. 2

[21] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, 2021. 7, 8

[22] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 280–297. Springer, 2022. 2, 6, 7, 8

[23] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 280–297. Springer, 2022. 7

[24] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *Proceedings of the Asian Conference

on Computer Vision (ACCV)*, pages 20–35. Springer, 2018. 2

[25] Hideyuki Suenaga, Huy Hoang Tran, Hongen Liao, Ken Masamune, Takeyoshi Dohi, Kazuto Hoshi, and Tsuyoshi Takato. Vision-based markerless registration using stereo vision and an augmented reality surgical navigation system: a pilot study. *BMC Medical Imaging*, 15(1):1–11, 2015. 1

[26] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8922–8931, 2021. 3

[27] Tatsunori Taniai, Yasuyuki Matsushita, Yoichi Sato, and Takeshi Naemura. Continuous 3d label stereo matching using local expansion moves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2725–2739, 2017. 8

[28] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14362–14372, 2021. 7, 8

[29] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419. Springer, 2020. 4, 5

[30] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2038–2041, 2018. 2, 5

[31] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12250–12259, 2019. 4

[32] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. Fadnet: A fast and accurate network for disparity estimation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 101–107. IEEE, 2020. 1, 7, 8

[33] Tianyuan Wang, Can Ma, Haoshan Su, and Weiping Wang. Cspn: Multi-scale cascade spatial pyramid network for object detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1490–1494. IEEE, 2021. 1, 2

[34] Tianyuan Wang, Can Ma, Haoshan Su, and Weiping Wang. Cspn: Multi-scale cascade spatial pyramid network for object detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1490–1494, 2021. 1

[35] Wenhuan Wu, Hong Zhu, Shunyuan Yu, and Jing Shi. Stereo matching with fusing adaptive support weights. *IEEE Access*, 7:61960–61974, 2019. 8

[36] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Semantic stereo matching with pyramid cost volumes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7484–7493, 2019. 2

[37] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Acvnet: Attention concatenation volume for accurate and efficient stereo matching. *arXiv preprint arXiv:2203.02146*, 2022. 1, 3, 6, 7, 8

[38] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 6, 7, 8

[39] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020. 6, 7, 8

[40] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *arXiv preprint arXiv:2211.05783*, 2022. 3, 7

[41] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 899–908, 2019. 1

[42] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016. 8

[43] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr. Ganet: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 7

[44] Songyan Zhang, Zhicheng Wang, Qiang Wang, Jinshuo Zhang, Gang Wei, and Xiaowen Chu. Ednet: Efficient disparity estimation with cost volume combination and attention-based spatial residual. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5433–5442, 2021. 1, 2

[45] Yaru Zhang, Yaqian Li, Yating Kong, and Bin Liu. Attention aggregation encoder-decoder network framework for stereo matching. *IEEE Signal Processing Letters*, 27:760–764, 2020. 2

[46] Yaru Zhang, Yaqian Li, Chao Wu, and Bin Liu. Attention-guided aggregation stereo matching network. *Image and Vision Computing*, 106:104088, 2021. 2