Adaptive Multi-Modal Cross-Entropy Loss for Stereo Matching

Peng Xu Zhiyu Xiang* Chengyu Qiao Jingyun Fu Tianyu Pu
College of Information Science and Electronic Engineering, Zhejiang University
{xxxupeng, xiangzy, 3140104437, fujingyun, 3190105835}@zju.edu.cn

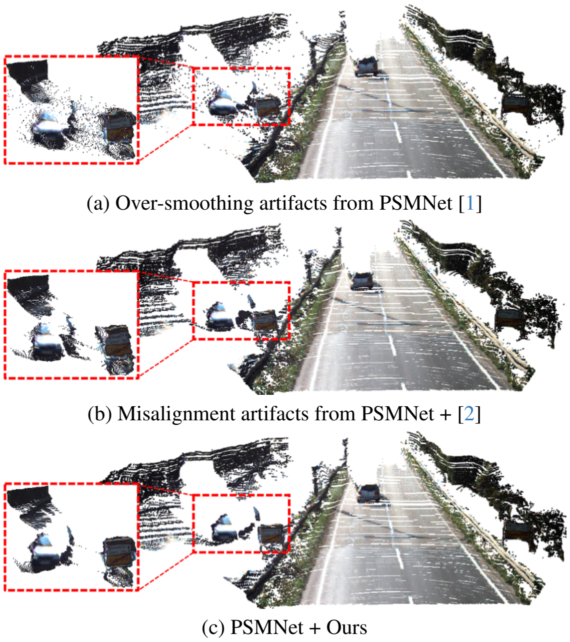
Abstract

Despite the great success of deep learning in stereo matching, recovering accurate disparity maps is still challenging. Currently, L1 and cross-entropy are the two most widely used losses for stereo network training. Compared with the former, the latter usually performs better thanks to its probability modeling and direct supervision to the cost volume. However, how to accurately model the stereo ground-truth for cross-entropy loss remains largely underexplored. Existing works simply assume that the ground-truth distributions are uni-modal, which ignores the fact that most of the edge pixels can be multi-modal. In this paper, a novel adaptive multi-modal cross-entropy loss (ADL) is proposed to guide the networks to learn different distribution patterns for each pixel. Moreover, we optimize the disparity estimator to further alleviate the bleeding or misalignment artifacts in inference. Extensive experimental results show that our method is generic and can help classic stereo networks regain state-of-the-art performance. In particular, GANet with our method ranks 1^{st} *on both the KITTI 2015 and 2012 benchmarks among the published methods. Meanwhile, excellent synthetic-to-realistic generalization performance can be achieved by simply replacing the traditional loss with ours. Code is available at https://github.com/xxxupeng/ADL.*

1. Introduction

As a long-standing and active topic in computer vision, stereo matching plays an essential role in wide applications such as autonomous driving and virtual reality. While conventional methods suffer from poor reliability in tackling illumination change and weak texture, the learning-based stereo methods show their superiority in these complex scenes.

Stereo matching is usually regarded as a regression task in deep learning [1, 3, 11, 13, 33]. In these works, L1 loss is employed for training, followed by the soft-argmax estimator [13] to predict sub-pixel disparity. The main problem of L1 loss is that it lacks direct supervision of the cost volume and is thereby prone to overfitting [36]. Moreover, soft-argmax is based on the assumption that the output distributions are uni-modal and centered on the ground-truth [13], which is not always true especially for the edge pixels with ambiguous depths. As shown in Fig. 1a, soft-argmax on edge pixels suffers from severe over-smoothing problem, causing bleeding artifacts at the edge.


(a) Over-smoothing artifacts from PSMNet [1]

(b) Misalignment artifacts from PSMNet + [2]

(c) PSMNet + Ours

Figure 1. **Comparison of the reconstructed point clouds.** Our method can alleviate the over-smoothing and misalignment artifacts, which is critical to the performance of downstream tasks.

Another line of research treats stereo matching as a classification task, where the cross-entropy loss could be used. To guide the network to output uni-modal distributions, researchers model the ground-truth disparity with discrete Laplacian or Gaussian distributions [2, 16, 28, 36]. The single-modal disparity estimator (SME) [2, 28] is further employed to extract correct modals from the predicted distributions. The cross-entropy loss can directly supervise the

[*] Corresponding author.

learning of the cost volume, thereby achieving better results than the L1 loss. However, the enforcement of uni-modal pattern seems not that effective, as evidenced by the presence of misalignment artifacts in Fig. 1b.

Our work aims to explore a better modeling for the stereo ground-truth and improve the disparity estimator. Contrary to previous works that impose the uni-modal constraints on the cost volume, we believe that the edge pixels should naturally be modeled as the multi-modal distributions. During the image capture process, edge pixels collect lights from multiple objects at different depths, implying that the depth of edge pixels inherently carries ambiguity. Enforcing the network to learn the uni-modal pattern at all areas can be confusing and misleading, causing erroneous estimation on both edge and non-edge pixels. Therefore, a better probability model encoding the true patterns of each pixel is highly desirable.

In this paper, we propose adaptive multi-modal distribution model for pixels and integrate it into cross-entropy loss for network training. We apply disparity clustering within the local window of each pixel to obtain the desired number of the modals. Laplacian distribution is then employed for modeling each cluster. We further rely on the local structural information within the window to determine the relative weight of each modal, thereby finalizing the mixture of Laplacians for the cross-entropy loss. Additionally, we propose a dominant-modal disparity estimator (DME) to better tackle the difficulties brought by the multi-modal outputs from the network. Extensive experimental results on public datasets show that our method is generic and can help classic stereo networks regain state-of-the-art performance. The comparison results in Fig. 1 exemplify the remarkable improvements of our method. Moreover, our method achieves excellent cross-domain generalization performance and exhibits higher robustness to sparser ground-truth.

Our contributions can be summarized as follows:

- We propose an adaptive multi-modal cross-entropy loss for training stereo networks. It can effectively guide the networks to learn clear distribution patterns and suppress outliers.
- We propose a dominant-modal disparity estimator that can obtain accurate results upon the multi-modal outputs.
- Extensive experiments show that our method is general and can help the classic stereo networks regain highly competitive performance. GANet [33] with our method ranks 1^{st} on both the KITTI 2015 [19] and the KITTI 2012 [10] benchmarks among all published methods.
- Networks with our method exhibit excellent generalization performance, surpassing existing methods that specialize in cross-domain generalization.
- Our method is robust to sparser supervision, revealing great potential to save the cost of producing dense ground-truth for network training.

2. Related work

Deep stereo matching. DispNet [18], a model that constructs a correlation volume and directly regresses the disparity, is the first end-to-end deep stereo network. Later, GCNet [13] proposes constructing the cost volume with concatenated features and employing 3D convolutions for cost aggregation. PSMNet [1], the popular baseline for the following cost volume-based works [2, 27, 30, 36], adds the spatial pyramid pooling [12] to the network and stacks multiple hourglass networks to improve the accuracy. GwcNet [11] further improves the cost volume by the group-wise correlation that provides more efficient measure of feature similarity. To reduce the computational complexity, GANet [33] proposes replacing the 3D convolutions with the aggregation layers guided by semi-global and local information. Following RAFT [25], another branch of work [14, 15, 31, 37, 38] relies on iterative refinement pipeline with ConvGRU [6] to achieve high disparity precision. Recently, IGEVStereo [31] proposes combining a geometry encoding volume with the correlation feature in the iterative pipeline, achieving the state-of-the-art performance on KITTI 2015 benchmark [19].

Loss function and disparity estimator. Loss function and disparity estimator are crucial for stereo networks. The former supervises the learning process, and the latter finalizes the disparity from the distribution volume. In GCNet [13], regression-based L1 loss is adopted and the full-band weighted average operation (soft-argmax) is proposed to calculate the final disparity. Later, smooth L1 loss becomes the mainstream [1, 3, 4, 11, 30]. Different from the above works, PDSNet [28] and the following works [2, 16, 36] uses the uni-modal cross-entropy loss to impose direct supervision to the distribution volume. No matter what the loss function is, the multi-modal outputs caused by the matching ambiguity are unavoidable. Soft-argmax on these multi-modal outptus leads to over-smoothing artifacts on the edge pixels. To solve this problem, SME [2, 28] selects the modal with the maximum probability and only estimates the final disparity on it. CDN [9] determines the integer part of the disparity from the modal with maximum probability and further estimates the offsets by a small network. SMDNet [27] feeds the distribution volume to MLPs [21] to parameterize the network outputs as the mixture of two Laplacians, and chooses the modal with higher peak as the final result. Beside these post-processing methods, [32] notices the multi-modal nature of the ground-truth when supervising the coarse-level cost volume in their multi-view stereo study. Contrary to the existing works that impose uni-modal distribution for each pixel, our method models ground-truth as adaptive multi-modal distributions and encourages multi-modal outputs on edge pixels. Different from [32] that introduces multi-hot cross-entropy loss for coarse-level patch-sized pixels but still employs L1 loss for the fine-level outputs, our method directly sets up adaptive multi-modal loss for the fine-level pixels, providing more direct and effective supervision to the network. We also optimize the disparity estimator to better tackle the multi-modal outputs from the distribution volume.

Cross-domain generalization. As another important issue for deep learning-based stereo matching, capability of cross-domain generalization has been extensively studied. DSMNet [34] proposes a novel domain normalization layer combined with a learnable non-local graph-based filtering layer to reduce the domain shifts. CFNet [24] builds a cascade and fused cost volume representation to learn domain-invariant geometric scene information. ITSA [7] refers to the information bottleneck principle [26] to minimize the sensitivity of the feature representations to the domain variation. GraftNet [17] embeds a feature extractor pre-trained on large-scale datasets into the stereo matching network to extract broad-spectrum features. Without adding any additional learnable modules, we achieve outstanding generalization performance by simply changing the training loss.

3. Method

3.1. Fundamentals and problem statement

Given a calibrated stereo image pair, stereo matching aims to find the corresponding pixel in the right image for each pixel in the left image. The cost volume-based stereo networks follow the common pipeline [13]. First, features of the left and right images are extracted by a weight-sharing 2D CNN module respectively. Then a 4D cost volume is constructed upon the two obtained feature blocks. The cost aggregation module takes this 4D volume as input and outputs a distribution volume with size $D \times H \times W$, where D is the maximum range of disparity search, H and W are the height and width of the input image, respectively. Softmax operator is then applied along the disparity dimension to normalize the probability distribution $p(\cdot)$ for each pixel. Finally, the resulting disparity $\hat{d}$ is estimated by the full-band weighted average operation, which is also called soft-argmax:

$$\hat{d} = \sum_{d=0}^{D-1} d \cdot p(d) \tag{1}$$

To train the stereo network, regression-based smooth L1 loss can be employed with:

$$\mathcal{L}_{reg}(\hat{d}, d_{gt}) = \begin{cases} 0.5(\hat{d} - d_{gt})^2, & if\ |\hat{d} - d_{gt}| < 1, \\ |\hat{d} - d_{gt}| - 0.5, & otherwise. \end{cases} \tag{2}$$

where d_{gt} is the ground-truth disparity.

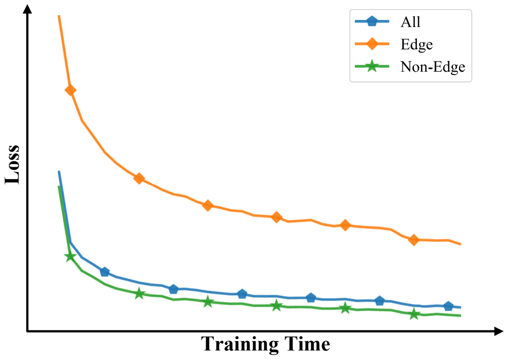


Figure 2. **Training trends** of the uni-modal cross-entropy loss on SceneFlow dataset.

In this pipeline, the distribution volume is indirectly supervised by the smooth L1 loss, which hinders the final performance [36]. By treating the stereo matching as a classification task, cross-entropy loss provides direct supervision on the distribution volume, as:

$$\mathcal{L}_{ce}(p, p_{gt}) = -\sum_{d=0}^{D-1} p_{gt}(d) \cdot \log p(d) \tag{3}$$

The new problem is that the ground-truth distribution $p_{gt}(\cdot)$ in Eq. (3) is unavailable. Existing works [2, 16, 28, 36] simply model $p_{gt}(\cdot)$ as the uni-modal Laplacian or Gaussian distribution centered on d_{gt}. However, these simple models seem unable to impose sufficient supervision for different image regions, especially the edge. As shown in Fig. 2, the training loss of edge pixels remain much larger than that of the non-edge pixels, indicating the difficulty of learning in these areas. Further statistics (details shown later in Tab. 3) on the resulting output distribution volume shows that over half of the edge and part of the non-edge pixels are actually assigned more than one modal, which conflicts with the uni-modal assumption of the pseudo-groundtruth. These undesired multi-modal outputs directly lead to the misalignment artifacts on object edges and outliers on non-edge areas, as shown in Fig. 1b. Therefore, we believe the root of the problem lies in the inappropriate uni-modal modeling of the ground-truth in all areas. In fact, edge pixels aggregate photometric information from multiple objects at different depths, implying that the intensities of edge pixels are inherently ambiguous. Imposing a uni-modal distribution pattern across the entire image will not only cause learning difficulties for edge pixels, but also confuse the learning for non-edge pixels.

3.2. Adaptive multi-modal probability modeling

Inspired by the observation in the previous section, we are dedicated to exploring a better probability modeling of ground-truth for the cross-entropy loss. We believe that

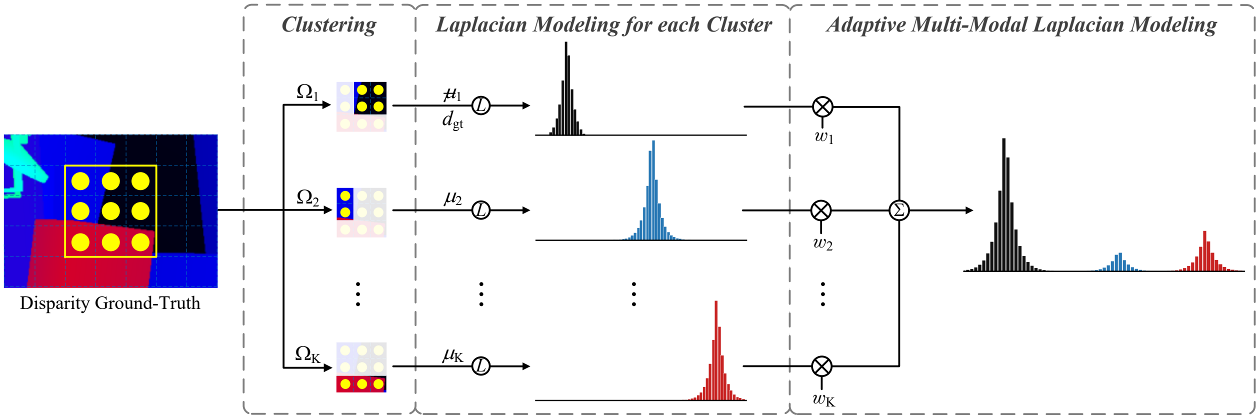

Clustering
Laplacian Modeling for each Cluster
Adaptive Multi-Modal Laplacian Modeling
Disparity Ground-Truth
Ω1
Ω2
ΩK
μ1
dgt
μ2
μK
L
w1
w2
wK
Σ

Figure 3. **Illustration of our adaptive multi-modal modeling** for cross-entropy loss. Given the pixel for modeling, the disparities within a pre-defined window are divided into K clusters $\{\Omega_1, \Omega_2, ..., \Omega_K\}$, and the mean μ_k for each cluster is calculated to form a *uni-modal* Laplacian distribution. The final adaptive *multi-modal* distribution is generated by the weighted summation of the Laplacian distributions, with the weight w_k determined by $|\Omega_k|$.

the probability distributions of edge pixels should be composed of multiple modals, with each corresponding to a specific depth/disparity. To this end, an adaptive multi-modal ground-truth modeling method is proposed. Our idea is to generate a separate Laplacian distribution for each potential depth on the edge pixels and then fuse them together to construct a mixture of Laplacians. We refer to the neighborhood of each pixel to accomplish the task, as illustrated in Fig. 3.

For each pixel labeled with ground-truth disparity, we consider a $m \times n$ local window centered on it. The entire set of disparity values within the window is then divided into $K (K \geq 1)$ disjoint subsets $\{\Omega_1, \Omega_2, ..., \Omega_K\}$ by the DBScan clustering algorithm [8], with each cluster corresponding to a different potential depth. In DBScan, the distance threshold ϵ and density threshold $minPts$ are set manually to adjust the resulting number of clusters. This clustering method offers the following advantages: (1) there is no need to pre-define the number of clusters; (2) $K = 1$ can be regarded as an indicator of non-edges; (3) it is robust for slanted planes with continuous but varying depths.

The ground-truth distribution of each pixel can then be modeled as the mixture of Laplacians:

$$
\begin{aligned}
p_{gt}(d) &= \sum_{k=1}^{K} w_k \cdot \text{Laplacian}_{\mu_k, b_k}(d) \\
&= \sum_{k=1}^{K} w_k \cdot \frac{e^{\frac{-|d-\mu_k|}{b_k}}}{\sum_{d_i=0}^{D-1} e^{\frac{-|d_i-\mu_k|}{b_k}}}
\end{aligned}
\tag{4}
$$

where the Laplacians are discretized and normalized over the disparity candidates $d \in \{0, 1, ..., D-1\}$, and μ_k, b_k, and w_k are the mean, scale, and weight parameters for the k^{th} Laplacian distribution, respectively. μ_k is set to the mean value of the disparities within the cluster Ω_k. Defining that Ω_1 contains the central pixel to be modeled, μ_1 is replaced by the central pixel's ground-truth to ensure the accuracy of the supervision. The weight w_k is designed to adjust the relative proportions of the obtained multiple modals, and can be assigned based on the local structure within the window. We take the cardinality of Ω_k as an indicator of the local structure, *e.g.*, a smaller $|\Omega_k|$ corresponds to a thinner structure, which should have smaller weight accordingly. Finally, w_k is defined as:

$$
w_k = \begin{cases} \alpha + (|\Omega_k| - 1) \cdot \frac{1-\alpha}{mn-1}, & k = 1 \\ |\Omega_k| \cdot \frac{1-\alpha}{mn-1}, & k \neq 1 \end{cases}
\tag{5}
$$

where α is a fixed weight for the central pixel. We set $\alpha \geq 0.5$ to ensure the dominance of the ground-truth modal. The rest $(1 - \alpha)$ weights are equally distributed to the rest $(mn - 1)$ neighboring pixels. For datasets with sparse ground-truth like KITTI [10, 19], only valid disparities within the local window are counted and mn in Eq. (5) is replaced with $\sum_{k=1}^{K} |\Omega_k|$. For non-edge pixels which have only one cluster within the window, w_1 is equal to one, and Eq. (4) degenerates into a uni-modal Laplacian distribution.

3.3. Dominant-modal disparity estimator

Stereo networks trained by cross-entropy loss often yield more multi-modal outputs than the L1 loss, thereby a better disparity estimator is highly desired. SME [2] alleviates the over-smoothing problem by aggregating disparities only within the "most likely" modal, but still suffers from the misalignment artifacts caused by the erroneous modal selection. In contrast to the uni-modal loss, our new loss encourages the network to generate more multi-modal patterns at the edge. Consequently, an enhanced disparity estimator becomes imperative to address the complexities introduced by our method.

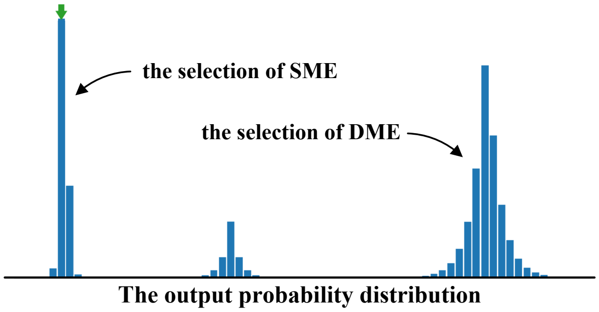


Figure 4. **Illustration of modal selection strategy** during inference. SME [2] prefers the modal with maximum probability candidate (aimed by the green arrow). Our proposed DME prefers the one with maximum cumulative probability.

SME first locates the disparity candidate with the maximum probability density, and then traverses left and right respectively until the probability stops decreasing, thereby determining the range of the dominant modal $[d_l, d_r]$ for disparity estimation. However, this pixel-level winner-take-all strategy is sensitive to noises that can produce sharp and narrow modals in the output distribution.

We propose our DME to solve this problem, as illustrated in Fig. 4. Specifically, we split each modal from the multi-modal output and calculate their cumulative probability separately. Each modal corresponds to a potential matching object with a specific depth, and the cumulative probability of the modal reflects the matching possibility of that object. Therefore, we adopt an object-level winner-take-all strategy, *i.e.*, selecting the modal with the maximum cumulative probability as the dominant modal. The selected modal is then normalized as:

$$\overline{p}(d) = \begin{cases} \frac{p(d)}{\sum_{d_i=d_l}^{d_r} p(d_i)}, & if\ d_l \leq d \leq d_r, \\ 0, & otherwise. \end{cases} \tag{6}$$

Finally, Eq. (1) is used to estimate the disparity by substituting $\overline{p}(d)$ for $p(d)$.

4. Experiments

4.1. Datasets and evaluation metrics

We evaluate our method on five popular stereo datasets. SceneFlow [18] is a large synthetic dataset containing 35454 image pairs for training and 4370 for testing. KITTI 2012 [10] and KITTI 2015 [19] are the two real outdoor datasets, each containing hundreds of images collected from driving scenes. Middlebury [22] and ETH3D [23] are also real-world datasets, with a few dozen of image pairs acquired in indoor or outdoor scenes. We only use the training sets of Middlebury and ETH3D to additionally validate the cross-domain generalization performance of our method.

As usual, EPE (End-Point-Error) and $k$px (the percentage of outliers with an absolute error greater than k pixels) are employed to evaluate the networks' performance. For KITTI 2015, D1 metric (the percentage of disparity outliers) is reported.

4.2. Implementation details

We separately apply our method to three classic cost volume-based stereo networks, namely, PSMNet [1], GwcNet [11], and GANet [33]. We implement all networks in PyTorch and use Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as the optimizer. We train the networks from scratch using two NVIDIA 3090 GPUs. When training on SceneFlow, the learning rate is set to 1×10^{-3} for the first 30 epochs and then reduced to 1×10^{-4} for the rest 15 epochs. On KITTI 2012 and 2015, we fine-tune the SceneFlow pre-trained networks for 600 epochs with the learning rate of 1×10^{-3}. b_k in Eq. (4) and α in Eq. (5) are both set to 0.8 after parameter tuning. The distance threshold ϵ and the density threshold $minPts$ in DBScan [8] are set to 3 and 1, respectively.

4.3. Ablation study

We perform ablations on the SceneFlow dataset. The original PSMNet [1] trained with smooth L1 loss is taken as the baseline for comparison. To ensure fairness, all models are trained from scratch for 15 epochs on the training set and then validated on the test set.

Loss function. As shown in Tab. 1, compared with the baseline, PSMNet with our multi-modal cross-entropy loss boosts performance drastically, with EPE by 19.59%, 1px error by 40.06%, and 3px error by 32.75%. Our method also outperforms the uni-modal cross-entropy loss [2] on all metrics, demonstrating its superiority in effectively guiding the network to learn the explicit distribution patterns.

Disparity estimator. We validate our disparity estimator by comparing with SME [2]. As shown in Tab. 1, our DME consistently outperforms SME for the networks trained with either uni- or multi-modal losses, which proves its effectiveness in selecting the correct modals from multi-modal outputs.

Window shape and size. The disparities within the window are used to construct the ground-truth distributions. We ablate to determine the optimal window shape and size. Tab. 2 shows that local windows with horizontal 1D shape usually perform better than others and the 1×9 window achieves the best. This can be attributed to the nature of stereo matching, *i.e.*, a 1D matching task along the horizontal direction.

4.4. Analysis of the output distribution patterns

To have a deeper understanding of the change in the distribution volume brought by our new loss, we count the proportions of pixels with different number of output

Method	Cross-Entropy Loss		Disparity Estimator		EPE	>1px	>3px
	Uni-Modal (UM)	Multi-Modal (MM)	SME	DME			
PSMNet [1]					0.97	10.51	4.03
PSMNet + UM + SME [2]	✓		✓		0.84	6.65	2.85
PSMNet + UM + DME	✓			✓	0.82	6.61	2.82
PSMNet + MM + SME		✓	✓		0.80	6.31	2.72
PSMNet + MM + DME (Ours)		✓		✓	**0.78**	**6.30**	**2.71**

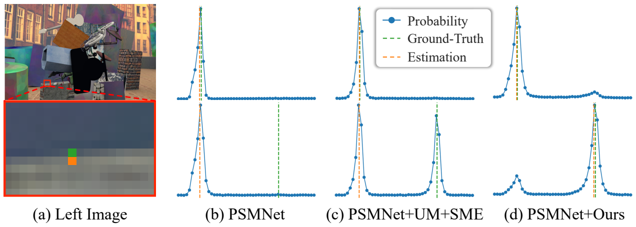
Table 1. **Ablation study** of the loss function in training and the disparity estimator in inference on SceneFlow.

Local Window		EPE	>1px	>3px
Shape	3×3	0.83	6.67	2.86
	1×9	**0.78**	**6.30**	**2.71**
	9×1	0.84	6.65	2.87
Size	1×3	0.82	6.76	2.85
	1×5	0.81	6.58	2.79
	1×7	0.81	6.62	2.83
	1×9	**0.78**	**6.30**	**2.71**
	1×11	0.84	6.72	2.91

Table 2. **Ablation study** of window shape and size on SceneFlow.

Method	Region	The number of modals			Outliers
		1	2	≥3	
PSMNet [1]	All	98.50	1.00	0.50	4.03
	Edge	87.79	10.30	1.91	25.59
	Non-Edge	98.92	0.64	0.44	3.17
+UM +SME [2]	All	94.67	4.60	0.73	2.85
	Edge	40.16	52.70	7.14	19.39
	Non-Edge	96.78	2.77	0.45	2.20
+MM +DME (Ours)	All	94.92	4.35	0.73	2.71
	Edge	35.35	57.28	7.37	18.97
	Non-Edge	97.23	2.33	0.44	2.07

Table 3. **Statistics of pixels w.r.t. the number of modals and their corresponding outliers** (>3px) for PSMNet variants on SceneFlow. Modals with the peak density lower than 1% are not included in the statistics.



(a) Left Image (b) PSMNet (c) PSMNet+UM+SME (d) PSMNet+Ours

Figure 5. Visualization of output distributions at the edge. Top row: background pixel, bottom row: foreground pixel.

patterns and list them in Tab. 3. PSMNet [1] yields the most uni-modal distributions for all of the pixels, namely, 98.5%. However, a part of these uni-modal distributions are not correct, as they may be centered on the wrong disparities which lead to large outliers. When uni-modal cross-entropy loss [2] is employed, the proportion of multi-modal distributions in the edge regions rises from 12.21% (10.30%+1.91%) to 59.84% (52.70%+7.14%), indicating the failure of the uni-modal constraints on the distribution volume. Compared with the uni-modal loss, our adaptive multi-modal loss yields about 5% more multi-modals at the edge while resulting in lower outliers. This demonstrates the superiority of our loss in supervising the network to produce easily distinguishable modals for disparity estimation. Interestingly, our method achieves better non-edge performance than [2], reducing outliers from 2.20% to 2.07%. It can also be observed that more uni-modal distributions than [2] are produced for the non-edge areas, which means that the supervision of clear patterns is beneficial for learning on not only edge but also non-edge pixels.

Method	EPE	>1px	>3px
PSMNet [1]	1.09	12.1	4.56
AcfNet [36]	0.87	–	4.31
GANet [33]	0.78	8.70	–
GwcNet [11]	0.77	8.00	3.30
PSMNet + [2]	0.77	–	2.21
IGEVStereo [31]	0.47	–	2.47
ACVNet [30]	**0.46**	4.89	1.98
PSMNet + Ours	0.64	5.14	2.19
GwcNet + Ours	0.62	5.07	2.16
GANet + Ours	0.50	**4.25**	**1.81**

Table 4. **Quantitative results on SceneFlow test set.**

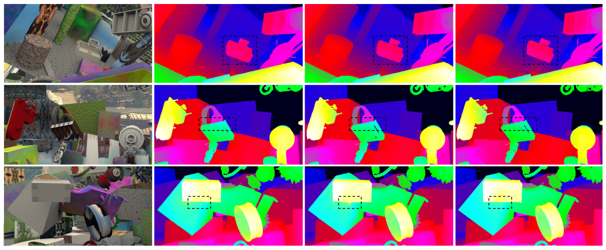
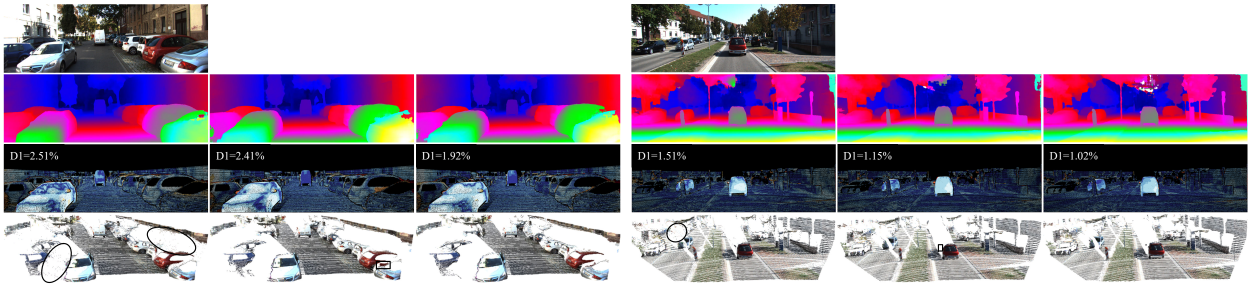


Figure 6. **Qualitative comparison on SceneFlow.** From left to right: input images, disparity results from PSMNet, PSMNet+[2], and PSMNet+Ours. Our method can recover more accurate object structure and reduce undesired defects at the edge.

Fig. 5 shows the output distributions of the two exampling pixels. We can observe that: 1) PSMNet outputs little multi-modal distributions but large disparity error; 2) SME incurs misalignment artifacts from ambiguous distribution;

Method	KITTI 2015						KITTI 2012			
	All			Noc			>2px		>3px	
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	Out-Noc	Out-All	Out-Noc	Out-All
PDSNet [28]	2.29	4.05	2.58	2.09	3.68	2.36	3.82	4.65	1.92	2.53
PSMNet [1]	1.86	4.62	2.32	1.71	4.31	2.14	2.44	3.01	1.49	1.89
PSMNet + [2]	1.54	4.33	2.14	1.70	3.90	1.93	2.17	2.81	1.35	1.81
GwcNet [11]	1.74	3.93	2.11	1.61	3.49	1.92	2.16	2.71	1.32	1.70
PSMNet + SMDNet [27]	1.69	4.01	2.08	1.54	3.70	1.89	–	–	–	–
CDN [9]	1.66	3.20	1.92	1.50	2.79	1.72	–	–	–	–
AcfNet [36]	1.51	3.80	1.89	1.43	3.25	1.73	1.83	2.35	1.17	1.54
PSMNet + [32] *	1.56	3.49	1.88	1.42	3.29	1.73	–	–	–	–
GANet [33]	1.48	3.46	1.81	1.34	3.11	1.63	1.89	2.50	1.19	1.60
GANet + LaC [16]	1.44	2.83	1.67	1.26	2.64	1.49	1.72	2.26	1.05	1.42
ACVNet [30]	**1.37**	3.07	1.65	1.26	2.84	1.52	1.83	2.34	1.13	1.47
LEAStereo [5]	1.40	2.91	1.65	1.29	2.65	1.51	1.90	2.39	1.13	1.45
IGEVStereo [31]	1.38	2.67	1.59	1.27	2.62	1.49	1.71	2.17	1.12	1.44
CroCoStereo [29]	1.38	2.65	1.59	1.30	2.56	1.51	–	–	–	–
PSMNet + Ours	1.44	3.25	1.74	1.30	3.04	1.59	1.80	2.32	1.14	1.50
GwcNet + Ours	1.42	3.01	1.68	1.30	2.76	1.54	1.65	2.17	1.05	1.42
GANet + Ours	1.38	**2.38**	**1.55**	**1.24**	**2.18**	**1.40**	**1.52**	**2.01**	**0.98**	**1.29**

Table 5. **Quantitative results on KITTI 2015 and 2012 Benchmarks.** * retrained network.



PSMNet **PSMNet+UM+SME** **PSMNet+Ours** **PSMNet** **PSMNet+UM+SME** **PSMNet+Ours**

Figure 7. **Qualitative comparison on KITTI 2015.** From top to bottom: left images, disparity maps, error maps, and reconstructed point clouds. The elliptical and rectangular boxes show partial over-smoothing and misalignment artifacts, respectively.

3) Our loss outputs more easily distinguishable multi-modal distributions than the uni-modal one.

4.5. Performance evaluation

We integrate our method into several baseline networks and compare them with other methods.

SceneFlow. As shown in Tab. 4, our method significantly improves the performance of all of the baselines by simply changing the loss function and disparity estimator. In particular, the EPE metrics are improved by 41.28%, 19.48%, and 35.90% for the baselines PSMNet, GwcNet, and GANet, respectively. GANet with our method achieves the state-of-the-art results on 1px and 3px metrics. Additionally, our multi-modal trained PSMNet also performs much better than the uni-modal trained one [2]. Qualitative results shown in Fig. 6 also validate the improvements.

KITTI 2015 & KITTI 2012 benchmarks. As the ground-truth of KITTI is sparse [10, 19], leveraging the adjacent rows for ground-truth modeling would be beneficial. Therefore, the size of the local window for generating the ground-truth distributions is enlarged to 3 × 9 when fine-tuning the SceneFlow pre-trained network on KITTI datasets. As the results shown in Tab. 5, all of the three baselines are lifted to a highly competitive level by our method. In particular, GANet with our method achieves new state-of-the-art results on both KITTI 2015 and KITTI 2012 benchmarks. Furthermore, we outperform those methods [2, 16, 28, 36], whose loss function contains a uni-modal cross-entropy term, by a large margin.

Since [32] doesn't have results on KITTI, we retrain the PSMNet with this method for the purpose of comparison. As shown in Tab. 5, our method also preforms better than those involving multi-modal modeling [27, 32].

To show our improvements more clearly, we convert the resulting disparity maps to point clouds. As shown in Fig. 7, our method dramatically improves the over-smoothing artifacts, and can obtain the point clouds with precise edge structures. More accurate point clouds can be very beneficial for downstream tasks, such as pseudo-LiDAR-based 3D object detection [20].

Method	KT 15 >3px	KT 12 >3px	MB >2px	ETH3D >1px
PSMNet [1]	16.3	15.1	25.1	23.8
GwcNet [11]	12.8	11.7	18.1	9.0
GANet [33]	11.7	10.1	20.3	14.1
DSMNet [34]	6.5	6.2	13.8	6.2
CFNet [24]	5.8	4.7	13.5	5.8
FC-GANet [35]	5.3	4.6	10.2	5.8
Graft-GANet [17]	4.9	4.2	9.8	6.2
ITSA-CFNet [7]	4.7	4.2	10.4	5.1
IGEVStereo [31]	–	–	**7.1**	3.6
PSMNet + Ours	4.78	4.23	8.85	3.44
GwcNet + Ours	**4.52**	4.19	9.11	3.79
GANet + Ours	4.84	**3.93**	8.72	**2.31**

Table 6. **Cross-domain generalization evaluation.**

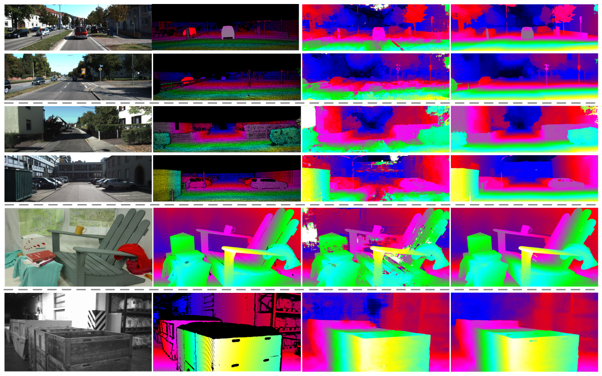


Figure 8. **Qualitative comparison of cross-domain generalization.** From top to bottom: KITTI 2015 [19], KITTI 2012 [10], Middlebury [22], and ETH3D [23].

4.6. Cross-domain generalization performance

Besides the fine-tuning performance, generalization is also crucial for deploying networks in the real world. In this section, we compare our generalization performance with baselines, as well as other methods that are specially designed for cross-domain generalization. All methods are only trained on SceneFlow and then tested on four real-world datasets [10, 19, 22, 23].

As shown in Tab. 6, the generalization performance of the baselines are greatly enhanced by our method. Meanwhile, by guiding the networks to learn explicit multi-modal patterns, our method shows superior performance than existing generalization-focused works on all four datasets. This proves that multi-modal distribution is more in line with the nature of stereo matching. Fig. 8 shows the qualitative comparison between the original PSMNet [1] and ours.

4.7. Influence of sparser ground-truth

Acquiring dense and accurate disparity ground-truth are difficult and expensive, especially for outdoor scenes. KITTI 2012 registers consecutive LiDAR point clouds with ICP to increase the ground-truth density [10], and KITTI 2015 further leverages detailed 3D CAD models to recover points on dynamic objects [19]. Despite these efforts, the valid ground-truth density on KITTI 2015 is only about 30%. In addition, the error of point cloud registration also needs to be considered, which affects the ground-truth quality. Therefore, a network that can be trained with sparser LiDAR ground-truth will be applauded.

Density	PSMNet [1]	PSMNet+Ours
100%	1.90	1.58
80%	2.15 (-13.16%)	1.61 (-1.90%)
60%	2.24 (-17.89%)	1.68 (-6.33%)
40%	2.30 (-21.05%)	1.71 (-8.23%)
20%	2.35 (-23.68%)	1.74 (-10.13%)

Table 7. **Influence of the ground-truth density.** D1 metric is reported on KITTI 2015 validation set.

In this experiment, we simulate different densities by randomly down-sampling the original ground-truth on KITTI 2015. Tab. 7 lists the influence of the ground-truth density to the final performance. The performance of both the original PSMNet [1] and ours degrades with less supervision signal. However, our method is much less affected than its counterpart, with just 10.13% degradation when trained with only 20% of the original ground-truth density. Even with this worst result, it is still much better than the best result of PSMNet trained with 100% original density. This clearly indicates the large potential of our method in saving the cost of collecting dense ground-truth for real applications.

5. Conclusion

In this work, we propose an adaptive multi-modal cross-entropy loss for stereo matching networks. Contrary to the previous works that impose uni-modal constraints on the distribution volume, our method encourages multi-modal outputs for edge pixels to avoid confusion in network learning. The number of modals and their corresponding weights in the distribution are determined by clustering and statistics within a local window. We also optimize the disparity estimator to robustly locate the dominant modal from the multi-modal outputs. Our method is general and can be easily implemented to enhance the performance of most of the existing stereo networks. GANet with our method achieves the new state-of-the-art on the KITTI 2012 and 2015 benchmarks. Our method is also robust to sparser ground-truth and exhibits excellent cross-domain generalization performance.

Acknowledgement. We thank all the reviewers for their valuable comments. The work is supported by the Key Research & Development Plan of Zhejiang Province (2024C01017).

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 1, 2, 5, 6, 7, 8

[2] Chuangrong Chen, Xiaozhi Chen, and Hui Cheng. On the over-smoothing problem of cnn based disparity estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8997–9005, 2019. 1, 2, 3, 4, 5, 6, 7

[3] Shuya Chen, Zhiyu Xiang, Chengyu Qiao, Yiman Chen, and Tingming Bai. Sgnet: Semantics guided deep stereo matching. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 2

[4] Shuya Chen, Zhiyu Xiang, Chengyu Qiao, Yiman Chen, and Tingming Bai. Pgnet: Panoptic parsing guided deep stereo matching. *Neurocomputing*, 463:609–622, 2021. 2

[5] 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. *Advances in Neural Information Processing Systems*, 33:22158–22169, 2020. 7

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[7] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13022–13032, 2022. 3, 8

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 4, 5

[9] Divyansh Garg, Yan Wang, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Wasserstein distances for stereo disparity estimation. *Advances in Neural Information Processing Systems*, 33:22517–22529, 2020. 2, 7

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 4, 5, 7, 8

[11] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 1, 2, 5, 6, 7, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 2

[13] 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 *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 1, 2, 3

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 2

[15] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2

[16] Biyang Liu, Huimin Yu, and Yangqi Long. Local similarity pattern and cost self-reassembling for deep stereo matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1647–1655, 2022. 1, 2, 3, 7

[17] Biyang Liu, Huimin Yu, and Guodong Qi. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13012–13021, 2022. 3, 8

[18] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2, 5

[19] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 2, 4, 5, 7, 8

[20] Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. End-to-end pseudo-lidar for image-based 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5881–5890, 2020. 7

[21] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2

[22] 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 *German conference on pattern recognition*, pages 31–42. Springer, 2014. 5, 8

[23] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017. 5, 8

[24] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13906–13915, 2021. 3, 8

[25] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[26] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015. 3

[27] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8942–8952, 2021. 2, 7

[28] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. *Advances in neural information processing systems*, 31, 2018. 1, 2, 3, 7

[29] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pretraining for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 7

[30] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022. 2, 6, 7

[31] 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. 2, 6, 7, 8

[32] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Non-parametric depth distribution modelling based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8626–8634, 2022. 2, 7

[33] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 1, 2, 5, 6, 7, 8

[34] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision*, pages 420–439. Springer, 2020. 3, 8

[35] Jiawei Zhang, Xiang Wang, Xiao Bai, Chen Wang, Lei Huang, Yimin Chen, Lin Gu, Jun Zhou, Tatsuya Harada, and Edwin R Hancock. Revisiting domain generalized stereo matching networks from a feature consistency perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13001–13011, 2022. 8

[36] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12926–12934, 2020. 1, 2, 3, 6, 7

[37] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Yong Zhao, Yitong Yang, and Ting Ouyang. Eai-stereo: Error aware iterative network for stereo matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 315–332, 2022. 2

[38] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1327–1336, 2023. 2