Masked representation learning for domain generalized stereo matching

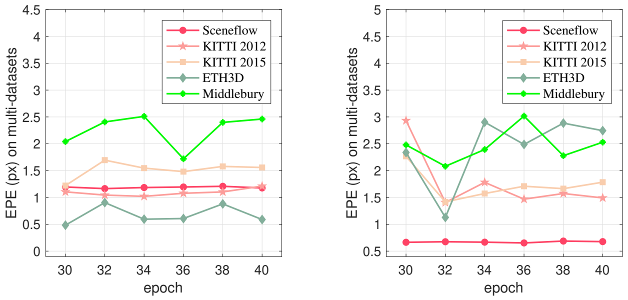
Zhibo Rao[1,2✉], Bangshu Xiong[1], Mingyi He[2], Yuchao Dai[2], Renjie He[2], Zhelun Shen[3], Xing Li[2✉]
[1]Nanchang Hangkong University, Nanchang, China
[2]Northwestern Polytechnical University, Xi'an, China
[3]Baidu Research, Beijing, China
raoxi36@foxmail.com, xiongbs@126.com, {myhe, daiyuchao, davidhrj}@nwpu.edu.cn,
1901213310@pku.edu.cn, lixing36@foxmail.com

Abstract

Recently, many deep stereo matching methods have begun to focus on cross-domain performance, achieving impressive achievements. However, these methods did not deal with the significant volatility of generalization performance among different training epochs. Inspired by masked representation learning and multi-task learning, this paper designs a simple and effective masked representation for domain generalized stereo matching. First, we feed the masked left and complete right images as input into the models. Then, we add a lightweight and simple decoder following the feature extraction module to recover the original left image. Finally, we train the models with two tasks (stereo matching and image reconstruction) as a pseudo-multi-task learning framework, promoting models to learn structure information and to improve generalization performance. We implement our method on two well-known architectures (CFNet and LacGwcNet) to demonstrate its effectiveness. Experimental results on multi-datasets show that: (1) our method can be easily plugged into the current various stereo matching models to improve generalization performance; (2) our method can reduce the significant volatility of generalization performance among different training epochs; (3) we find that the current methods prefer to choose the best results among different training epochs as generalization performance, but it is impossible to select the best performance by ground truth in practice.

1. Introduction

Stereo matching is a challenging research topic of computer vision, which aims to obtain the corresponding pixels between two rectified stereo images [3, 30, 37, 39]. It is essential in many applications, including autonomous [11], augmented reality [35], virtual reality [2], etc.



(a) CFNet (b) LacGwcNet

Figure 1. The generalization performance among different epochs on multi-datasets. When the models converge on the source domain (Sceneflow), the results are stable on the source domain. However, generalization performance has fluctuations on target datasets (KITTI 2012&2015, ETH3D, and Middlebury).

In mainstream deep stereo matching methods, the main steps include four parts: feature extraction, cost volume, feature matching, and disparity regression [10]. To improve the accuracy or speed, researchers proposed many strategies to improve the above four parts [3, 25, 36]. For example, 1) scene awareness modules are applied to the feature extraction to improve accuracy [3, 21]; 2) similarity measurements are employed to enhance physical representation in cost volume construction [23, 26, 36]; 3) various feature matching modules are proposed to solve inefficient calculations or improve accuracy [15, 25, 39]; 4) loss functions of classification or other fields are joined to disparity regression to obtain unimodal results [42]. Although the above methods have achieved significant progress in accuracy or efficiency, they failed to obtain good generalization performance in unseen domains [30, 40].

Therefore, many approaches learn domain-invariant representation features to achieve better generalization capability [40]. There are three solutions to accomplish this objective: the unsupervised matching method [24,32,33], domain

[1] ✉ is the corresponding author.

adaptation techniques [12, 13, 17, 31], and domain generalization approaches [16, 30, 41]. The above solutions reveal that feature presentation is crucial for improving generalization capability [40, 41]. However, all domain generalized matching methods did not mention that the generalization performance varies significantly among different training epochs, as shown in Fig. 1. Namely, the results were unstable. These methods preferred to choose the best results to represent the generalization performance by testing the models of different training epochs. However, we can not employ ground truth (can not obtain) to select the best model among different training epochs in practice. Thus, it is crucial to keep a stable generalization performance.

On the other hand, many researchers introduced multi-task learning into stereo matching [18, 27]. For instance, Rao et al. proposed a bidirectional guided attention network to cope with semantic segmentation and stereo matching simultaneously [27]. Liu et al. employed the task-shared and task-specific manner to obtain a generalizable representation and used the loss weights to balance all tasks [18]. The multi-task learning proved that feature sharing and weight sharing are essential in different tasks, promoting the learning process and obtaining a better feature representation. Inspired by these multi-task learning methods, we use masked left images as input to reconstruct the original left image as another task. By building pseudo-multi-task learning, we obtain better structural information features, urging stereo matching models to perform better generalization.

This paper addresses the domain generalization for stereo matching methods by masked representation learning. Our solution considerably increases generalization accuracy while reducing the volatility of generalization performance among different training epochs. First, we randomly mask (or remove) the part of the left image with a fixed ratio following a uniform distribution. Second, we add a simple decoder to the feature extraction module's tail to recover the original left image. Finally, we train the model with two tasks (stereo matching and image reconstruction) as a pseudo-multi-task learning framework and test generalization performance on the multi-datasets. In this task, our strategy helps existing methods to achieve significant gains in generalization performance. Furthermore, we exhibit the generalization performance over different training epochs to demonstrate the superiority of our approach. We hope these observations will help other tasks explore better generalization results. The main contributions are as follows:

- We combine masked image modeling and stereo matching as a pseudo-multi-task learning framework to increase generalization accuracy in stereo matching.
- Our approach is employed for various stereo matching networks, significantly improving cross-domain accuracy and reducing the volatility of generalization performance among different training epochs.
- We find that the accuracy of the existing stereo matching domain generalization methods varies significantly among different training epochs. Thus, we advise that stability should be evaluated in cross-domain methods.

2. Related Work

Deep learning based stereo matching. Since Zbontar *et al*. used convolution neural networks (CNNs) to replace hand-crafted features and compute matching costs, many researchers started to apply this novel technology to update the traditional pipelines [38]. Mayer *et al*. built a cost volume in a correlation manner and created a sizeable synthetic dataset to meet the data requirements in stereo matching [19]. Kendall *et al*. first proposed 4D cost volume by stacking two view features and used 3D convolution to regularize costs, producing a profound impact in stereo matching [10]. Chang *et al*. introduced the scene awareness module to extract features and applied a 3D hourglass module to aggregate cost volume [3]. Zhang *et al*. utilized the traditional semi-global matching method to reform the feature matching module (3D convolution), capturing the local or global cost dependencies [39]. Shen *et al*. replaced the 3D hourglass module to cascade structure as a coarse to fine process, achieving a state-of-the-art generalization performance [30]. Li *et al*. employed a hierarchical network with recurrent refinement to modify a stacked cascaded architecture and proposed an adaptive group correlation layer to reduce the impact of erroneous rectification [11].

Domain adaptation stereo matching. To improve generalization performance, many researchers converted the original domain (synthetic scenes) to the target domain (real scenes) [31]. Liu *et al*. employed an end-to-end training framework with domain translation to tackle the problem of pixel distortion and stereo mismatch after translation [17]. Song *et al*. designed a bottom-up domain adaptation method based on color transfer and cost regularization to improve generalization without any learnable parameters [31]. Li *et al*. proposed Fourier-based amplitude transform (FAT), mapping the source image to the target style without altering semantic content [12].

Domain generalization stereo matching. Generalization performance is an important indicator that reveals the ability of networks to cope with the unseen domain. Tonioni *et al*. joint unsupervised and continuous online learning to preserve the model's accuracy in multiple environments [33]. Zhang *et al*. proposed a domain-invariant stereo matching network by regularizing the distribution of learned representations and extracting robust structural representations [40]. Cai *et al*. introduced matching functions and confidence measures to replace the learning-based feature extraction module, leading to superior generalization to unseen environments [1]. Li *et al*. introduced a transformer to the stereo matching task, which showed strong

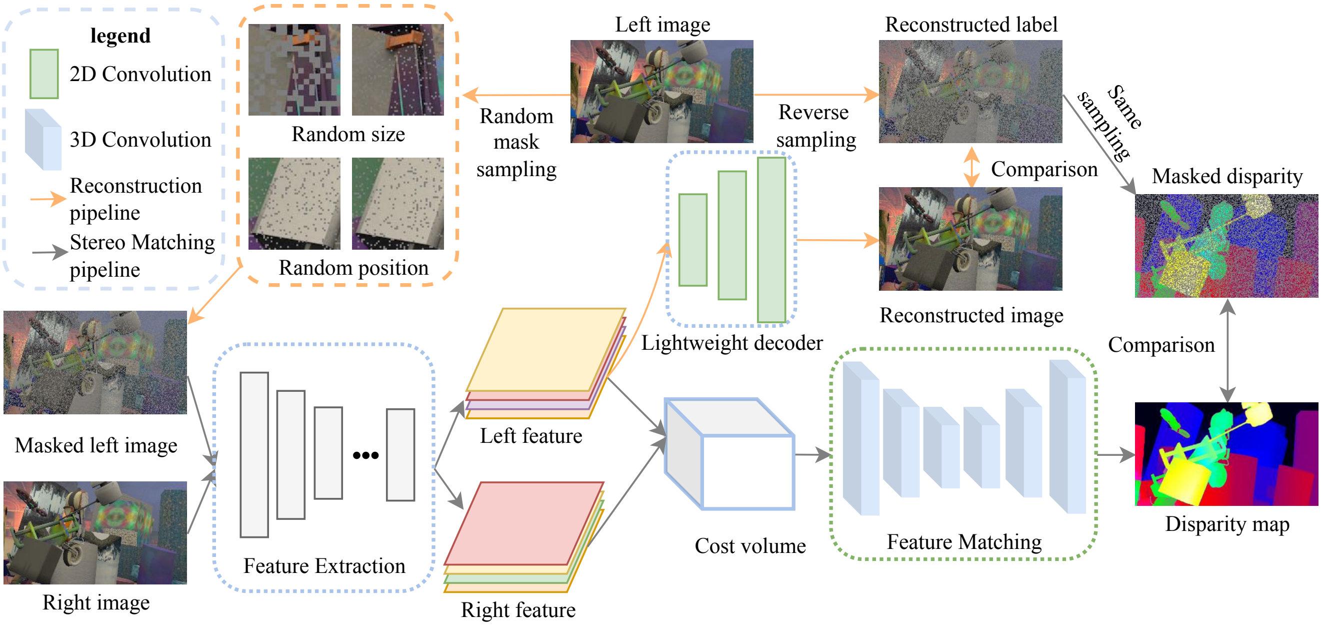


Figure 2. The overall pipeline of our method. Inspired by masked representation learning and multi-task learning, we build a pseudo-multi-task learning framework based on image reconstruction and stereo matching. The feature extraction module can learn better structure information by feature fusion in pseudo-multi-task learning, improving generalization performance and stability.

robustness [14]. Liu *et al*. leveraged the feature of a model trained on the large-scale dataset to deal with the domain shift and used a cosine similarity-based cost volume to graft task-oriented features [16]. Zhang *et al*. argued maintaining feature consistency between matching pixels is a vital factor for the generalization capability and employed a simple pixel-wise contrastive learning across the viewpoints [41].

Masked representation learning. Masked modeling is a highly successful way for pre-training in natural language processing (NLP) and computer vision (CV) [5, 9]. Vincent *et al*. proposed denoising autoencoders (DAE), which corrupted an input signal and learned to reconstruct the original, uncorrupted signal [34]. Pathak *et al*. presented an unsupervised visual feature learning algorithm driven by context-based pixel prediction, proving the effectiveness of CNN pre-training on classification, detection, and segmentation tasks [22]. Devlin *et al*. held out a portion of the input sequence and trained the transformer to predict the missing content [5]. Chen *et al*. inspired by progress in unsupervised representation learning in NLP and trained a sequence transformer to predict pixels without incorporating knowledge of structure information [4]. He *et al.* attempted to demonstrate that masked autoencoders (MAE) are scalable self-supervised learners for computer vision, which achieved the best accuracy among approaches that only used ImageNet-1K data [9]. Feichtenhofer *et al*. studied a simple extension of Masked Autoencoders (MAE) to spatiotemporal representation learning from videos [6].

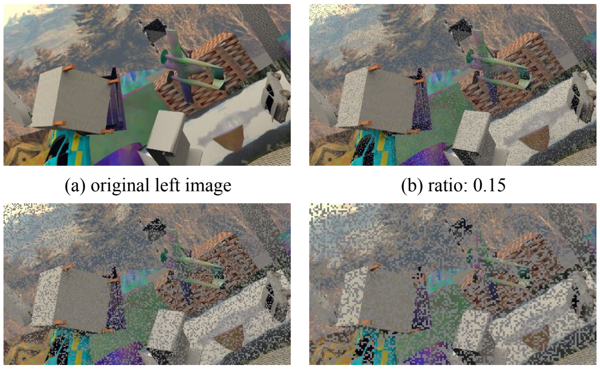
3. Method

3.1. Overall Pipeline

In a typical stereo matching method based on deep learning [3, 10, 11, 30, 39], the rectified left and right images are fed to the feature extraction module. Then, these methods construct the cost volume by concatenating the left and traveled right features. Finally, they use 3D convolutions to aggregate the cost volume and regress to the disparity map. Therefore, the poor generalization ability of stereo matching must be caused by two learnable modules (feature extraction module and feature matching module).

In previous studies [16, 40], many achievements have demonstrated that feature representation in the feature extraction module is a vital factor in the model's generalization ability. Meanwhile, we have noticed that multi-task learning helps the models obtain a better matching accuracy, especially when another task has a recognition attribute (e.g., semantic segmentation, object detection, classification, etc.) [18, 27]. Why can recognition algorithms help models to get better performance in multi-task learning? Because feature-sharing in two or more tasks can help the models learn better structural or scene information [16, 18]. Inspired by the above research, we elegantly employ masked representation learning to build a pseudo-multi-task learning framework, promoting models to learn better structural information in the feature extraction module to improve generalization ability, as shown in Fig. 2.

Masked Input image. Following MAE [6, 9], we follow a uniform distribution to randomly mask pixels as gray in the left image with the fixed size and ratio (e.g., $0.15, 0.3, 0.45$), as shown in Fig. 3. Unlike MAE, we use a low ratio and small mask size to mask pixels and do not employ random positions of patches. Because 1) we hope the models extrapolate the missing pixel from visible neighboring patches to perceive structural information; 2) lacking too many pixels is not conducive to learning correlation in the stereo matching task, mainly feature matching module; 3) stereo matching is sensitive to location information, and the goal is to build the relationship of correlation, so we do not use random positions. This design helps the models to learn structural information in the feature extraction module and does not impact the training process of other learnable modules (e.g., the feature matching module.).



(a) original left image

(b) ratio: 0.15

(c) ratio: 0.3

(d) ratio: 0.4

Figure 3. Illustration of mask sampling strategy. It determines the difficulty of the reconstruction task and affects the reconstruction quality. More masked pixels mean more difficult reconstruction tasks and less reference information in stereo matching.

After masking the left image, we feed the masked left and unmasked right images to the feature extraction module to obtain the left and right features. Then, we design a simple and efficient decoder to predict the missing pixels by the left features, introduced next.

Decoder Design. We treat the feature extraction module as an encoder to encode the masked left image and obtain the left features. In stereo matching, all methods use down-sampling in the feature extraction process to reduce computation and memory utilization. Therefore, we get the features with a compressed size (e.g., $H/4 \times W/4 \times F$) after the feature extraction module. To reconstruct the target image, we design a lightweight decoder to decode the features to the original size, as shown in Fig. 2.

We only employ the decoder module to cope with the image reconstruction task during training. In other words, the decoder module does not work in the testing process. Therefore, the decoder module will not affect the runtime of the existing stereo matching models in the testing process. In this paper, our decoder design is very lightweight and only consists of three 2D convolutions and two up-sampling operations. Thus, the training process will not be significantly prolonged in our design.

Reconstruction Target. The decoder module reconstructs the left image by predicting the missing pixels. The last layer of the decoder module is a line projection without batch normalization (BN) or ReLU Activation Function (ReLU). Following MAE, we use normalized pixels as the reconstruction target to improve representation quality, and we only compute the loss on the masked pixel of the left image. Our idea is to use image reconstruction and stereo matching to build a pseudo-multi-task learning framework, promoting structural representation learning to improve cross-domain performance further.

3.2. Loss Functions

We apply image reconstruction and stereo matching to construct pseudo-multi-task learning. Hence, the loss functions contain two parts, as follows:

Reconstruction loss function. We employ the mean squared error (MSE) to compute the loss $\mathcal{L}_r$ between reconstructed and original left images in the pixel space. The reconstruction loss function $\mathcal{L}_r$ can be defined as follows:

$$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{N}(I_o(i) - I_r(i))^2, \tag{1}$$

where N demotes the number of masked pixels, i is the identifier of masked pixels, I_o represents the original left image, and I_r indicates the reconstructed masked pixels.

Matching loss function. The matching loss $\mathcal{L}_m$ in different papers is different. Most of these methods use the mean absolute error (MAE) or $\mathrm{Smooth}\ L_1$ loss as based loss functions. Then, they add photometric or other loss functions to improve performance. In our method, we follow the loss functions of previous matching works and continue to adopt the same manner as the loss following their paper. Therefore, the total loss can be presented as follows:

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_m. \tag{2}$$

3.3. Algorithm applicability and advantage

Algorithm applicability. Although many stereo matching methods have been proposed recently, they do not break away from the existing framework (four parts mentioned in Sec. 1). All models have feature extraction modules, although the style of the modules is variable. Hence, our method can be plugged into the current matching methods by treating the feature extraction modules as encoders and adding a simple decoder. Our method works for all current matching algorithms, not just a few.

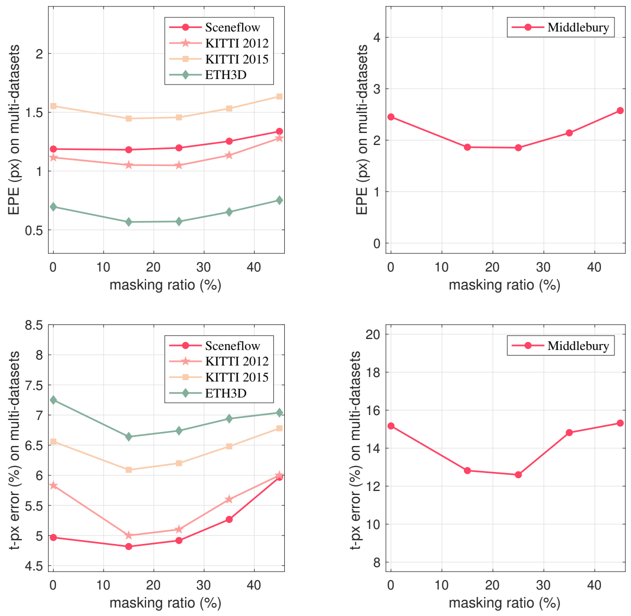


Figure 4. **CFNet with different masking ratio.** As the masking ratio rises, performance rises first and then declines.

Algorithm advantage. Compared with current domain adaptation and domain generalization methods in stereo matching, our approach has many advantages regarding runtime and convenience, as shown in follows. (1) Unlike domain adaptation technologies, our method does not need an additional training process or access to the target domain data. Meanwhile, our decoder module is a lightweight structure. Thus, it will not significantly prolong the training process of the existing algorithms. (2) Unlike domain generalization methods, our image reconstruction branch does not participate in testing. Therefore, our approach does not affect the runtime of the current matching methods.

4. Experiments

4.1. Datasets & Evaluation metrics

The current domain generalized evaluation manner is that the models are trained on the source domain and tested on the target domain. Thus, we first briefly describe these datasets based on different domains and then introduce evaluation metrics on these datasets.

Source domain. We train all models on the Sceneflow dataset [19] in the experiment with the same schedule.

Target domain. Following previous works [12, 16, 30], we evaluate all models on the KITTI 2012&2015 (KT-12 and KT-15) [7, 20], ETH3D (ET) [29], and Middlebury (MB) [28] without fine-tuning.

Evaluation metric. Following previous works [16, 40], we apply end-point-error (EPE) and the percentage of error pixels larger than t pixels ($> t$ px) as evaluation metrics. According to evaluated websites, we set $t = 1, 2, 3$, which correspond to the default thresholds of $t = 1$ (*bad 1.0*) on the ETH3D, and $t = 2$ (*bad 2.0*) on the Middlebury, and $t = 3$ (D_1) on the KITTI 2012&2015, respectively.

4.2. Implementation Details

In this paper, we implement our strategy with state-of-the-art methods (CFNet [30] and LacGwcNet [15]) by PyTorch. According to our goal, we need to decide on the hyper-parameters and training schedule, as presented below.

Hyper-parameters. We set the maximum disparity $D = 256$ in CFNet and LacGwcNet to ensure that nearly all possible disparity values in the images could be detected. We apply Adam Optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) to optimize the model. In training, we set a mini-batch size of 1 image pair per GPU (six on 6 GPUs) and all loss weights following their papers.

Training schedule. We only train models on the SceneFlow dataset. The learning rate is initially set to 1×10^{-3} for 30 epochs and then reduced to 1×10^{-4} for the other 10 epochs. After this process, we obtain the final models. For data augmentation, we use random cropping, random resize, random shift, and random chromatic transformations to cope with the input image on the SceneFlow dataset.

4.3. Main Properties

Ratio	Type	KT-12	KT-15	ET	MB
0	EPE	1.11	1.55	0.69	2.45
0.15	EPE	**1.05**	**1.44**	**0.56**	1.86
0.25	EPE	**1.05**	1.45	0.57	**1.85**
0.35	EPE	1.13	1.53	0.65	2.14
0.45	EPE	1.27	1.63	0.75	2.57
0	t-px error	5.83	6.56	7.25	15.17
0.15	t-px error	**5.01**	**6.09**	**6.64**	12.82
0.25	t-px error	5.12	6.20	6.74	**12.60**
0.35	t-px error	5.63	6.48	6.94	14.81
0.45	t-px error	4.00	6.79	7.04	15.32

Table 1. The generalization performance of CFNet with different masking ratios. A low ratio improves generalization performance, while a high ratio hurts generalization performance.

Masking ratio. Fig. 4 and Tab. 1 show the influence of the masking ratio (we use an average of ten epochs as generalization performance.). When the masking ratio is low, it does not affect the performance (source domain) but improves generalization performance. As the masking ratio increases, the performance (source domain) gradually declines, and generalization performance rises first and then falls. Thus, a high masking ratio is not unsuited for this pseudo-multi-task framework due to two factors: (1) a high masking ratio will increase the difficulty of reconstruction, which is a disadvantage to the learning process; (2) excessive masking will reduce the learnable matching area in the following modules, such as the feature matching module. Some qualitative results of our final models on the four realistic datasets are displayed in Fig. 5.

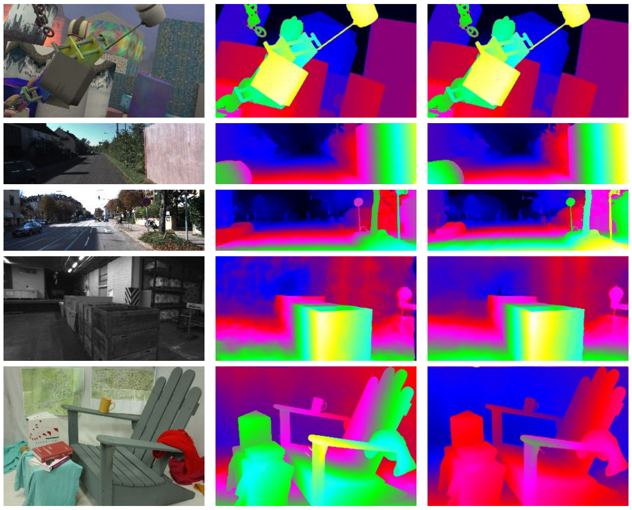
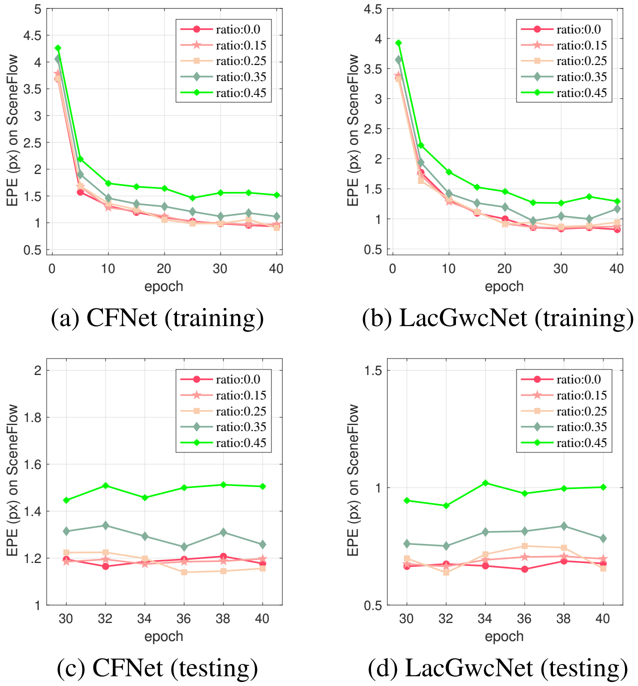

(a) Left image (b) Mask-CFNet (C) Mask-LacGwcNet

Figure 5. The examples of different models on the source and target datasets (without fine-tuning). From top to bottom are SceneFlow (source dataset), KITTI 2012&2015, ETH3D, and Middlebury (four target datasets).

Convergence process. In Fig. 6, we compare the performance and the convergence process with or without masked representation learning. Because CFNet and LacGwcNet have many outputs, we use the matching accuracy of the final output as the performance on the training dataset. Meanwhile, we test the performance of these methods on the source domain. We conclude that: (1) the results are very stable on the source domain; (2) compared with baselines, our method does not change the convergence process for the low mask ratio; (3) a high mask ratio will affect the learning process and reduce the matching accuracy. Stereo matching is a task highly related to location information. Mask ratio means more missing pixels, and it will affect the models to learn the correlation based on adjacent pixels in the feature matching module.

Runtime. As shown in Tab. 2, we list the runtime of two models at the training and testing processes. Our method only adds a lightweight decoder to reconstruct the left image in the training process, which does not participate in the test process. Therefore, we draw the following conclusions: (1) for the training process, our method does not significantly prolong training time compared with baselines; (2) for the testing process, our approach is no different from baselines.

Training epochs and generalization performance. As presented in Fig. 7, we draw two models' curves with or without masked representation learning. When the models converge (> 30 epochs, as shown in Fig. 6), the results are stable on the source dataset. However, the generalization performance varies significantly between adjacent training epochs. Compared with the baselines, the models with masked representation learning can perform better and more stable. As shown in Tab. 3, we apply mean and variance to measure these fluctuations among different training epochs. By quantifying the changes, we find that the structural information can help the models achieve a relatively stable and better generalization performance.



(a) CFNet (training) (b) LacGwcNet (training)
(c) CFNet (testing) (d) LacGwcNet (testing)

Figure 6. The convergence process with or without masked representation learning. It proves that a low ratio does not change the convergence process and the matching accuracy, while a high ratio affects the learning process and reduces the matching accuracy.

Model	Mask	Training	Resolution	Runtime (s)
CFNet	✓	✓	576 × 320	0.89
CFNet	✓	✗	960 × 576	0.052
CFNet	✗	✓	576 × 320	0.84
CFNet	✗	✗	960 × 576	0.051
LacGwcNet	✓	✓	576 × 320	1.63
LacGwcNet	✓	✗	960 × 576	0.264
LacGwcNet	✗	✓	576 × 320	1.61
LacGwcNet	✗	✗	960 × 576	0.264

Table 2. The runtime with different resolutions. It presents that our method does not significantly prolong training time and is no different from the baseline at the test runtime.

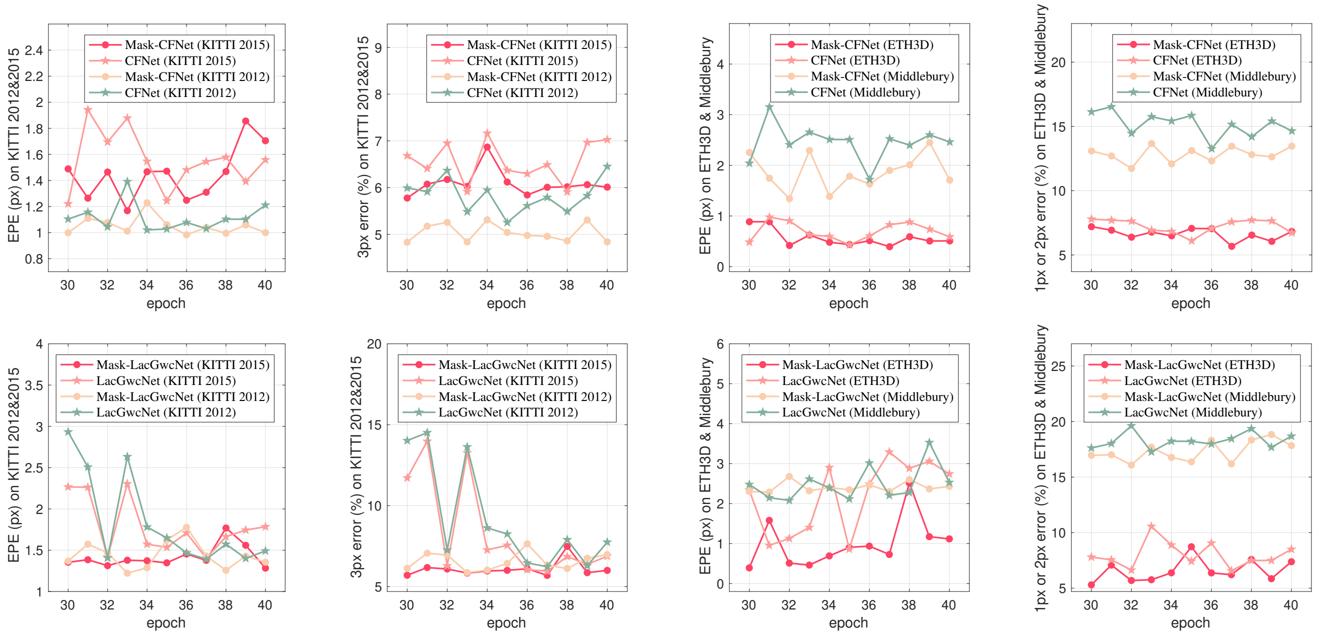


Figure 7. The generalization performance with or without masked representation among different epochs. It shows that the generalization performance varies significantly between adjacent training epochs, and our method achieves better performance and is more stable.

M.	Mask	Data	EPE (Mean)	EPE (Var.)	D_1 (Mean)	D_1 (Var.)
CFNet	✔	KT-12	**1.44**	**0.04**	**5.03**	**0.03**
	✘	KT-12	1.55	0.05	5.82	0.13
	✔	KT-15	**1.05**	0.01	**6.08**	**0.07**
	✘	KT-15	1.11	0.01	6.56	0.19
	✔	ET	**0.56**	**0.02**	**6.63**	**0.21**
	✘	ET	0.69	0.03	7.24	0.30
	✔	MB	**1.86**	0.13	**12.82**	**0.37**
	✘	MB	2.45	0.13	15.16	0.90
LacGwcNet	✔	KT-12	**1.43**	**0.03**	**6.57**	**0.30**
	✘	KT-12	1.83	0.32	9.17	10.46
	✔	KT-15	**1.41**	**0.02**	**6.08**	**0.23**
	✘	KT-15	1.78	0.11	8.37	9.24
	✔	ET	**1.00**	**0.37**	**6.57**	**1.03**
	✘	ET	2.18	0.84	7.99	1.37
	✔	MB	**2.40**	**0.02**	**17.30**	0.89
	✘	MB	2.49	0.19	18.28	**0.51**

Table 3. Volatility comparison of with or without masked representation. It demonstrates that our method can help the models achieve a relatively stable and better generalization performance.

4.4. Comparisons

Comparisons with generalization method. We compare our methods with other stereo matching methods, including well-known and generalized methods. Tab. 4 shows that our approach can help models improve generalization ability, indirectly proving that masked representation learning can promote high-level representation feature learning. Note that all methods have volatility among training epochs, and we list the volatility as shown in the supplemental material. It prompted us to suggest that the scope of fluctuation should be attached to future work in stereo matching.

Method	KT-12 > 3px	KT-15 > 3px	MB > 2px	ET > 1px
PSMNet [3]	15.1	16.3	26.9	23.8
GWCNet [8]	12.0	12.2	34.2	11.0
GANet [39]	10.1	11.7	20.3	14.1
DSMNet [40]	6.2	6.5	21.8	6.2
FC-DSM [41]	5.5	6.2	12.0	6.0
CFNet [30]	4.7	5.8	15.3	5.8
GF-PSMNet [16]	5.3	4.6	10.9	6.2
Mask-CFNet	4.8	5.8	13.7	5.7
Mask-LacGwcNet	5.7	5.6	16.9	5.3

Table 4. Cross-domain generalization evaluation (peak results) on four target datasets. All methods are only trained on the Scene-Flow dataset and tested on training images of four real datasets.

Comparisons with fine-tuning. We fine-tune the models (CFNet and LacGwcNet) on the KITTI datasets to test the performance after fine-tuning. Tab. 5 shows that the results are similar to the trend of the SceneFlow dataset. When the masking ratio is low ($< 25\%$), our method does not affect the performance and produces nearly the same results. As the ratio rises, the performance gradually declines.

Method	Ratio	KT-12 (Out-Noc)	KT-15 (D1-all)
LacGwcNet [15]	0	1.13	1.77
LacGwcNet [15]	0.15	1.15	1.78
LacGwcNet [15]	0.25	1.16	1.77
LacGwcNet [15]	0.35	1.27	1.95
LacGwcNet [15]	0.45	1.39	2.21
CFNet [30]	0	1.23	1.88
CFNet [30]	0.15	1.23	1.89
CFNet [30]	0.25	1.27	1.91
CFNet [30]	0.35	1.36	2.05
CFNet [30]	0.45	1.48	2.28

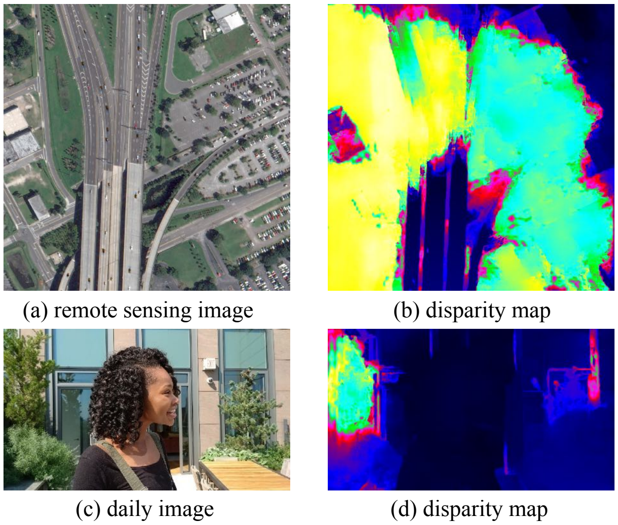
Table 5. The fine-tuning results on the KITTI dataset. It indicates the results are similar to the trend of the SceneFlow dataset.

5. Discussion

This paper has proposed a simple approach to alleviate the significant volatility of generalization performance. This section will discuss two topics: (1) When does our approach fail? (2) What is the real unseen domain?

When does our approach fail? (1) In our study, we have added the channels of the feature extraction module to get a better reconstruction image. However, the generalization performance did not obtain remarkable improvement. Although larger models can handle two tasks well, it is not conducive to the fusion of multi-task learning because larger modules may learn features of two tasks separately. Similarly, it will weaken the effect of masked representation. (2) Excessive masking will cause the learnable matching area to shrink in the feature matching module, affecting matching accuracy or generalization. Thus, large-size masking is not recommended to improve generalization.

What is the real unseen domain? In previous studies, nearly all papers used KITTI, ETH 3D, and Middlebury datasets as the unseen domains. However, can these datasets be represented the unseen domain? Although the color distribution, illumination, and authenticity are different between the source domain and the target domains, similar objects are included in the source domain, such as the street scenes in the Driving dataset (sub-dataset of SceneFlow), pedestrian in the Driving dataset, daily supplies in FlyingThing3D (sub-dataset of SceneFlow), etc. Herefore, we can get good results without fine-tuning models on the target dataset. Does it mean we can use these algorithms in practice? The answer is no. We use these models that work well on target datasets to deal with daily images collected by the zed camera or remote sensing images collected by satellite. As shown in Fig. 8, we can not get the expected results for these images. We consider that the main reason is that these elements or scenes do not appear in the source dataset, such as the human body, buildings under the top view angle, etc. When the real unseen domain appears in the images, matching methods based on deep learning can not predict disparity correctly. Thus, we doubt the cross-domain performance of these datasets (KITTI, ETH 3D, and Middlebury) can represent actual cross-domain performance.


(a) remote sensing image

(b) disparity map

(c) daily image

(d) disparity map

Figure 8. The examples of failures in the real unseen domain. When elements do not appear in the source dataset, the models can not predict the expected results.

6. Conclusion

In this paper, we have proposed a simple masked representation method to address the problem of unstable generalization performance among different training epochs for stereo matching. Our approach is inspired by masked representation learning and multi-task learning to construct a pseudo-multi-task learning framework, which helps models learn better features in the feature extraction module to improve generalization performance and stability. Extensive experiments have proved the effectiveness of our method. Meanwhile, it also demonstrated that the current evaluation manner is unsuitable for measuring generalization performance. In the last, we discussed the failure of our approach and doubted the current definition of the unseen domain.

Acknowledgements. This research was supported in part by Jiangxi Natural Science Foundations (20224BAB212009, 20212BBE53017), National Natural Science Foundation of China (61866027, 62271410, and 62001396), the Foundation of Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition (ET202204304), and the Doctor's Foundation of Nanchang Hangkong University (EA202204228, EA202204257). We thank the anonymous area chairs and reviewers for their helpful feedback.

References

[1] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization. In *International Conference on 3D Vision (3DV)*, pages 364–373, 2020. 2

[2] Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, and Henry Fuchs. Stereodrnet: Dilated residual stereonet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11786–11795, 2019. 1

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 1, 2, 3, 7

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1691–1703, 2020. 3

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019. 3

[6] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *arXiv preprint*, 2022. 3, 4

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 5

[8] 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 (CVPR)*, pages 3273–3282, 2019. 7

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3, 4

[10] 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 (ICCV)*, pages 66–75, 2017. 1, 2, 3

[11] 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 (CVPR)*, pages 16263–16272, 2022. 1, 2, 3

[12] Xing Li, Yangyu Fan, Zhibo Rao, Zhe Guo, and Guoyun Lv. Improving stereo matching generalization via fourier-based amplitude transform. *IEEE Signal Processing Letters (SPL)*, 2022. 2, 5

[13] Xing Li, Yangyu Fan, Zhibo Rao, Guoyun Lv, and Shiya Liu. Synthetic-to-real domain adaptation joint spatial feature transform for stereo matching. *IEEE Signal Processing Letters (SPL)*, 29:60–64, 2021. 2

[14] 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. 3

[15] 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 (AAAI)*, volume 36, pages 1647–1655, 2022. 1, 5, 8

[16] 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 (CVPR)*, pages 13012–13021, 2022. 2, 3, 5, 7

[17] Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, and Hongsheng Li. Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12757–12766, 2020. 2

[18] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880, 2019. 2, 3

[19] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 2, 5

[20] 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. 5

[21] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. Multi-level context ultra-aggregation for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3283–3291, 2019. 1

[22] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 3

[23] Zhibo Rao, Yuchao Dai, Zhelun Shen, and Renjie He. Rethinking training strategy in stereo matching. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, pages 1–14, 2022. 1

[24] Zhibo Rao, Mingyi He, Yuchao Dai, and Zhelun Shen. Patch attention network with generative adversarial model

for semi-supervised binocular disparity prediction. *The Visual Computer*, pages 1–17, 2020. 1

[25] Zhibo Rao, Mingyi He, Yuchao Dai, and Zhelun Shen. Sliding space-disparity transformer for stereo matching. *Neural Computing and Applications (NCAA)*, pages 1–14, 2022. 1

[26] Zhibo Rao, Mingyi He, Yuchao Dai, Zhidong Zhu, Bo Li, and Renjie He. Nlca-net: a non-local context attention network for stereo matching. *APSIPA Transactions on Signal and Information Processing*, 9, 2020. 1

[27] Zhibo Rao, Mingyi He, Zhidong Zhu, Yuchao Dai, and Renjie He. Bidirectional guided attention network for 3-d semantic detection of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):6138–6153, 2020. 2, 3

[28] 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, 2014. 5

[29] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3260–3269, 2017. 5

[30] 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. 1, 2, 3, 5, 7, 8

[31] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. Adastereo: a simple and efficient approach for adaptive stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10328–10337, 2021. 2

[32] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9661–9670, 2019. 1

[33] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 195–204, 2019. 1, 2

[34] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1096–1103, 2008. 3

[35] Bin Xu, Yuhua Xu, Xiaoli Yang, Wei Jia, and Yulan Guo. Bilateral grid learning for stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12497–12506, 2021. 1

[36] 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 (CVPR)*, pages 12981–12990, 2022. 1

[37] 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. 1

[38] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1592–1599, 2015. 2

[39] 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 (CVPR)*, pages 185–194, 2019. 1, 2, 3, 7

[40] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision (ECCV)*, pages 420–439, 2020. 1, 2, 3, 5, 7

[41] 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 (CVPR)*, pages 13001–13011, 2022. 2, 3, 7

[42] 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 (AAAI)*, pages 12926–12934, 2020. 1