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

1. Overview

In this supplementary material, we present more experimental results and analysis, including:

- We test different popular networks among training epochs to validate volatility.
- We show more visualization results of our method.
- We discuss the reasons for volatility and the issue of the unseen domain.

2. More Experiments

In this section, we first compare our method with other methods in peak results, and then we compare the volatility of these methods, as shown in follows.

Method	KT-12 (> 3px)	KT-15 (> 3px)	MB (> 2px)	ET (> 1px)
PSMNet	15.1	16.3	26.9	23.8
CFNet	4.7	5.8	15.3	5.8
LacGwcNet	6.0	5.7	18.3	6.3
GF-PSMNet∗	5.3	4.6	10.9	6.2
GF-PSMNet†	5.0	5.3	17.6	11.4
Mask-PSMNet	5.3	6.0	15.8	10.6
Mask-LacGwcNet	5.7	5.6	16.9	5.3
Mask-CFNet	4.8	5.8	13.7	5.7

Table 1. The peak cross-domain generalization evaluation. ∗ means only test on limited disparity range(0-192) and limited area ($mask\_occ$), † means we use pre-trained models they provided on all pixels (gt>0), and other methods test on all pixels (gt>0).

(1) Comparison of peak results

We list the peak results of our method with LacGwcNet [4], CFNet [6], and PSMNet [2], as shown in Tab. 1. Compared with baselines (LacGwcNet, CFNet, and PSMNet), our methods can effectively improve the cross-domain generalization of peak results. Meanwhile, our method achieves better performance improvements compared with the state-of-the-art method (GF-PSMNet [5]).

(2) Comparison of volatility

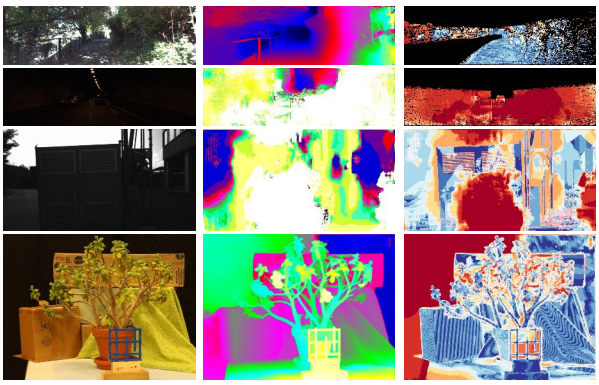
First, we download the code and pre-trained models of these methods (PSMNet [2], GANet [7], DSMNet [8], LacGwcNet [4], GF-PSMNet [5], and CFNet [6]). Then, we load the pre-trained models and continue to train the models for 15 epochs in four NVIDIA 1080 Ti. Finally, we compute the volatility (mean ± std) of these methods in the last 10 epochs, as shown in Tab. 2. **Note that our hardware is worse than what they used in their paper. Thus, we used a smaller batch size and cropped size, leading to a slight difference in cross-domain performance compared with what their paper claimed.** However, it did not affect that we can obtain the following conclusions.

Method	KT-12 (> 3px)	KT-15 (> 3px)	MB (> 2px)	ET (> 1px)
PSMNet	10.4 ± 3.50	14.7 ± 1.03	22.4 ± 7.54	17.6 ± 1.46
GANet	8.90 ± 0.75	12.1 ± 1.19	19.1 ± 9.24	12.1 ± 1.23
DSMNet	5.90 ± 1.14	6.60 ± 0.18	19.7 ± 3.64	7.12 ± 1.65
GF-PSMNet	6.00 ± 0.89	5.70 ± 0.58	17.8 ± 1.87	12.3 ± 1.59
LacGwcNet	9.17 ± 10.46	8.37 ± 9.24	18.28 ± 0.51	7.99 ± 1.37
CFNet	5.82 ± 0.13	6.56 ± 0.19	15.16 ± 0.90	7.24 ± 0.30
Mask-PSMNet	6.66 ± 0.66	6.36 ± 0.31	16.56 ± 0.94	11.4 ± 0.91
Mask-LacGwcNet	6.57 ± 0.30	6.08 ± 0.23	17.30 ± 0.89	6.57 ± 1.03
Mask-CFNet	5.03 ± 0.03	6.08 ± 0.07	12.82 ± 0.37	6.63 ± 0.21

Table 2. The volatility evaluation of cross-domain generalization. All methods test on all pixels (gt>0). We use mean ± std to estimate volatility.

From Tab. 2, we can find that our method can get better cross-domain performance and be more stable. Meanwhile, it also proves that current popular methods have volatility in cross-domain generalization. Unlike fine-tuning manner on target datasets (the same fine-tuning manner obtains similar results), cross-domain performance is unstable. However, the existing evaluation manner chooses the peak per-

[1]✉ is the corresponding author.



(a) left image (b) disparity map (c) error map

Figure 1. The examples of failures in the KITTI2012&2015, ETH3D, and Middlebury datasets.

formance as cross-domain performance. It is not suitable for evaluating cross-domain performance containing significant volatility. Thus, we recommend that add the volatility evaluation to compare the robustness of the methods.

3. More Visualization Results

The results show that our method can achieve a better cross-domain performance in most data of the KITTI2012&2015, ETH3D, and Middlebury datasets, but stereo matching models still can not deal with some scenes. We also show the failed examples of CFNet in the KITTI2012&2015, ETH3D, and Middlebury datasets, as shown in Fig. 1. Note that most deep learning-based methods can not deal with these scenes, not just CFNet. Thus, it also encourages us to test more data to verify the cross-domain performance of the current methods. We discover that current well-generalized methods still can not deal with many scenes, such as US3D, ZED, etc. It has to make us suspect that can KITTI2012&2015, ETH3D, and Middlebury datasets represent the unseen domain. Furthermore, we also wonder if the cross-domain performance evaluated in these datasets (KITTI2012&2015, ETH3D, and Middlebury) can represent the actual cross-domain performance of stereo matching methods.

4. More Discussion

Volatility. It is worth to be discussed why volatility exists in generalized stereo matching models. We guess several reasons causing this problem, as shown in follows. In the cross-domain task, the target domain's attributes are unknown for the models and optimizers, such as the input style, disparity range, camera parameters, etc. Meanwhile, the learning direction is oriented to the source domain rather than the target domain. The learning process contains many random attributes, like (1) all methods use data augmentation that contains random attributes; (2) all methods use a mini-batch manner, also having random attributes. Thus, random attributes will increase or reduce the domain difference between the source and target domains in the training process, causing volatility. However, in a fine-tuning manner, the testing and training data belong to identical distribution, and the learning direction is also the same. Thus, the volatility is minimal in the same domain, while the generalization fluctuation problem is unavoidable in the cross-domain.

Real unseen domain. In Fig. 8, no methods based on deep learning can work, including GANet, PSMNet, DSMNet, etc. This phenomenon inspires us to think about the reason. To analyze this phenomenon, we summarize the experimental phenomenon, as shown in the following. (1) The domain's attributes (input style, disparity range, camera parameters, etc.) of the KITTI2012&2015, ETH3D, US3D, ZED, and Middlebury are all very different from SceneFlow. However, models can work in the KITTI2012&2015, ETH3D, and Middlebury while failing in ZED and US3D. (2) Traditional methods (SGM) can work on US3D or ZED [1, 3] while the learning method fails. (3) Not only US3D or ZED, but also KITTI2012&2015, US3D, and Middlebury have several stereo pairs that models cannot work ($D_1 > 45\%$), as shown in Fig. 1. Thus, we convince that poor cross-domain performance is related to image content. Meanwhile, it reveals that there is still a long way to achieve perfect cross-domain performance, so our next step is to make models adapt to new scenes dynamically.

References

[1] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1524–1532, 2019. 2

[2] 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

[3] Sunok Kim, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Feature augmentation for learning confidence measure in stereo matching. *IEEE Transactions on Image Processing (TIP)*, 26(12):6019–6033, 2017. 2

[4] 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] 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. 1

[6] 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

[7] 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

[8] 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