JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# 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*,
2. *Northwestern Polytechnical University*,
3. *Baidu Research*

∗ Corresponding author: Zhibo Rao (raoxi36@foxmail.com)
Xing Li (lixing36@foxmail.com)


南昌航空大学 NANCHANG HANGKONG UNIVERSITY

西北工業大學 NORTHWESTERN POLYTECHNICAL UNIVERSITY

Baidu 百度 Robotics and Autonomous Driving Lab

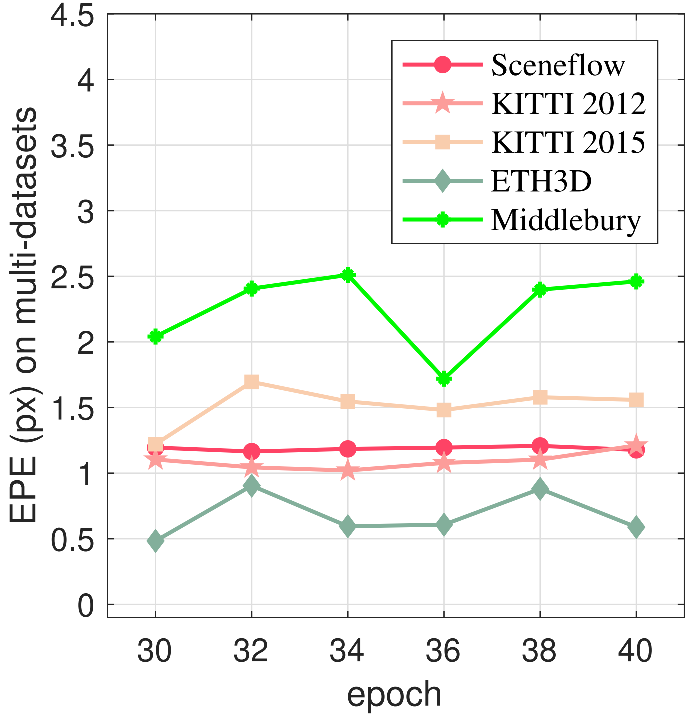
# Background and Motivation

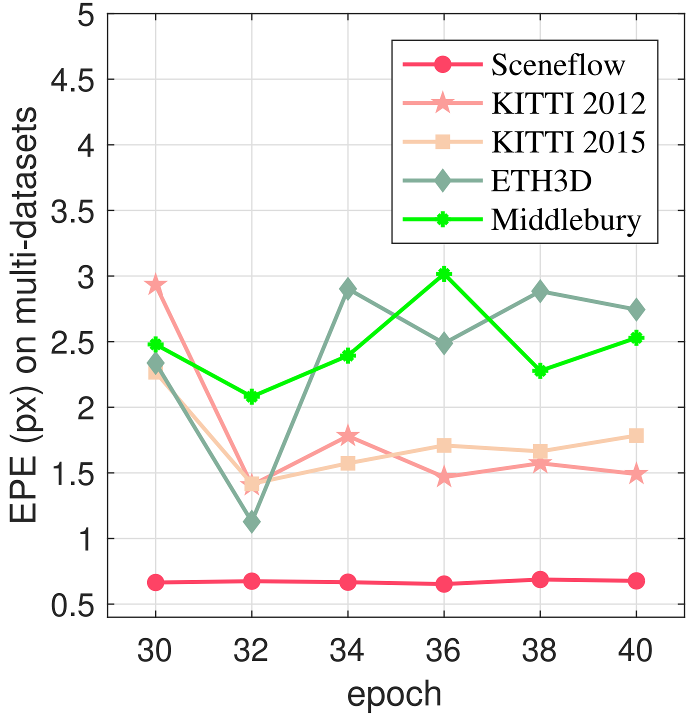
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

## Goal of cross-domain in stereo mathcing:

- The models are only trained in source domain (Sceneflow)
- The models are tested in target domains (KITTI, ETH3D, and Middlebury)

**Question:** Generalization performance has fluctuations on target datasets.

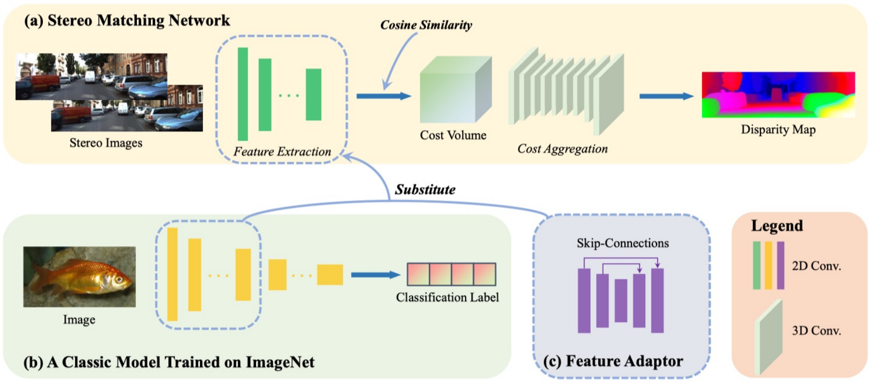


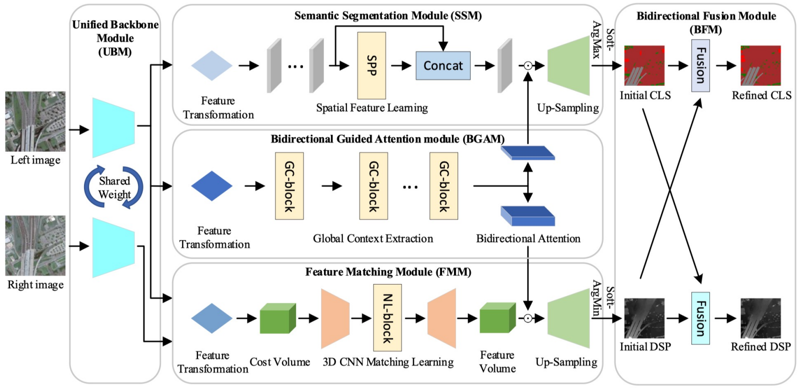

The generalization performance of CFNet among different epochs on multi-datasets.


[1] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In CVPR, pages 13906–13915, 2021.

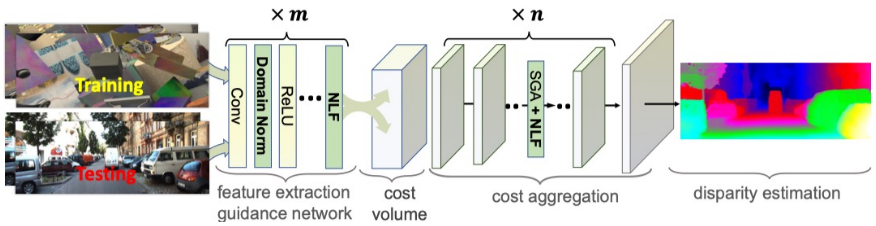
# Motivation




Graftnet shows **the feature extraction is the key for cross-domain.**




BGA-Net shows **multi-task learning can help model learn better feature.**




DSMNet shows **the structural information is also the key for cross-domain.**

Thus, How can we apply multi-task learning to help the feature extraction to obtain structural information?

[1] Biyang Liu, Huimin Yu, and Guodong Qi. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In CVPR, pages 13012–13021, 2022.
[2] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In ECCV, pages 420–439, 2020.
[3] Zhibo Rao, Mingyi He, Zhidong Zhu, Yuchao Dai, and Renjie He. Bidirectional guided attention network for 3d semantic detection of remote sensing images. TGRS, 59(7):6138–6153, 2020.

# Contributions

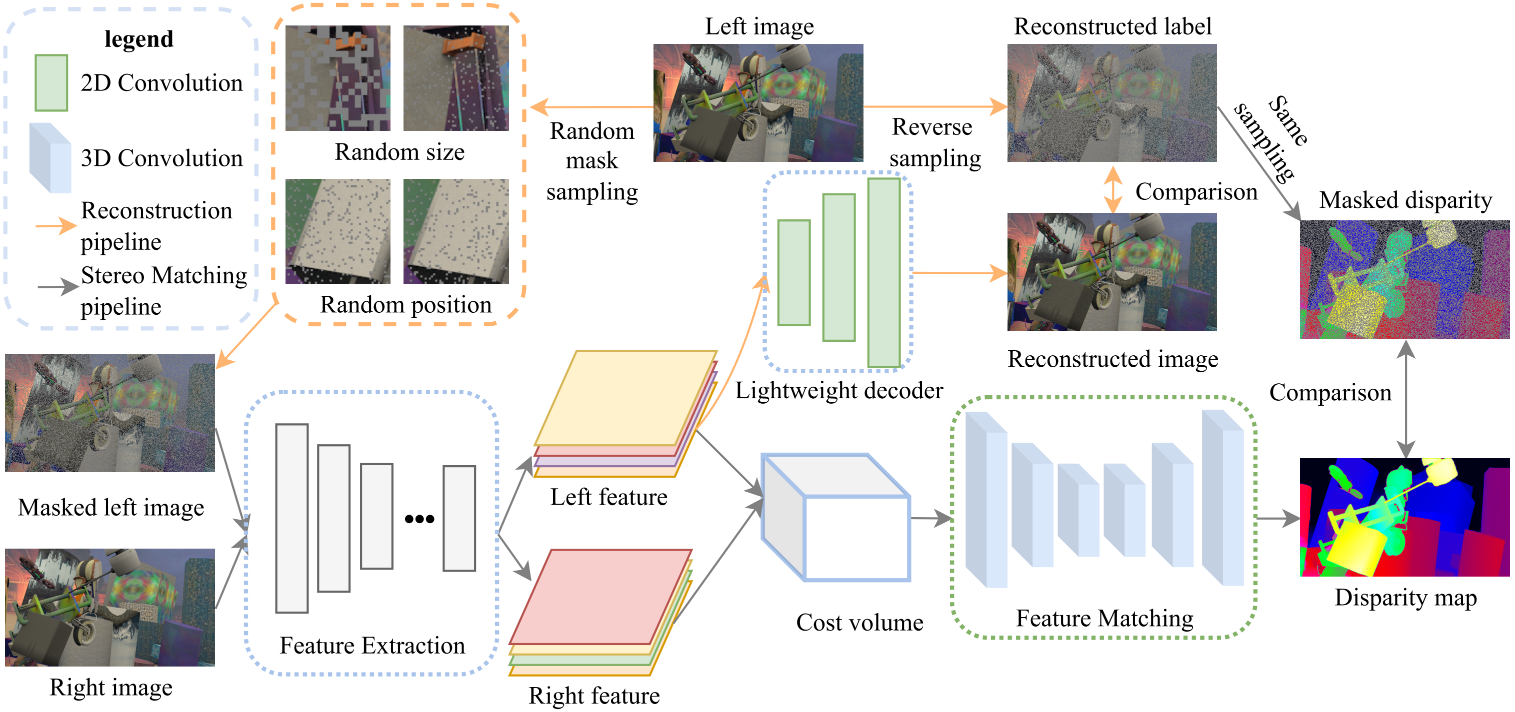


- We build **a pseudo-multi-task learning framework** to increase generalization.
- Our methods can **improve cross-domain accuracy and reduce the volatility**.
- We find that **cross-domain results varies significantly** among different epochs.

# Method



Inspired by masked representation and multi-task learning, we build a pseudo-multi-task learning framework for better structural information.




First, we randomly mask the part of the left image.

Second, we add a decoder to recover the left image.

Finally, we train models with two tasks as a pseudo-multi-task learning framework.

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll´ar, and Ross Girshick. Masked autoencoders are scalable vision learners. In CVPR, pages 16000–16009, 2022.

# Method



**Reconstruction loss function:** $$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{N}(I_o(i) - I_r(i))^2,$$

**Matching loss function:** we follow the loss functions of previous matching works.

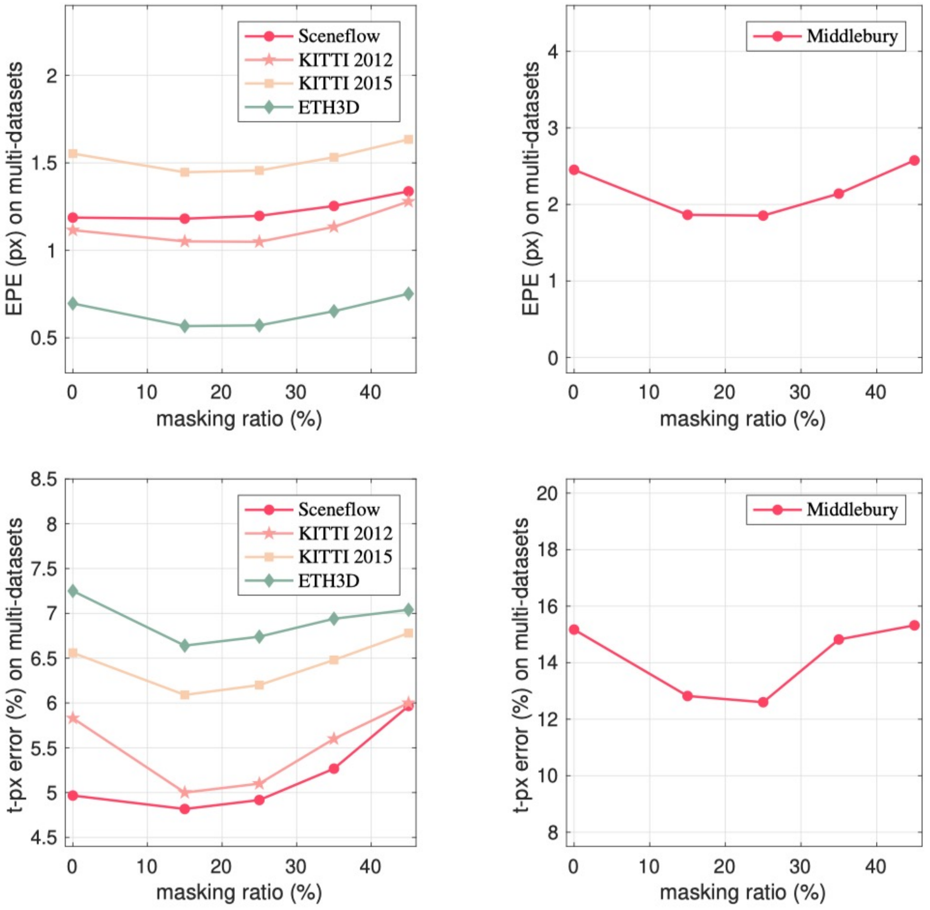
**The total loss:** $$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_m.$$

## Advantage:

- Our method works for all current matching algorithms, not just a few.
- Our method does not need an additional training process or access to the target domain data.
- Our image reconstruction branch does not participate in testing.

# Experiments






CFNet with different masking ratio.

| 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 |

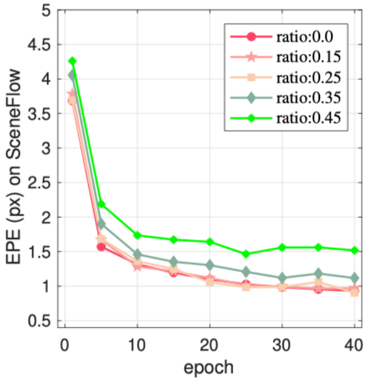
The influence of the masking ratio

*Conclusion:*

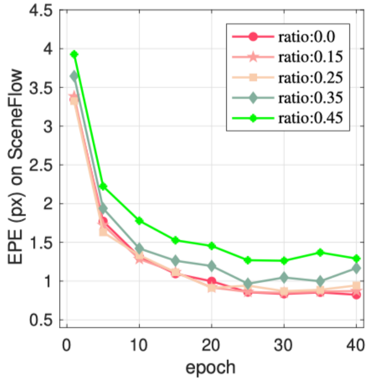
1. When the masking ratio is low, it does not affect the performance (source domain) but improves generalization performance.
2. As the masking ratio increases, the performance (source domain) gradually declines, and generalization performance rises first and then falls.

# Experiments

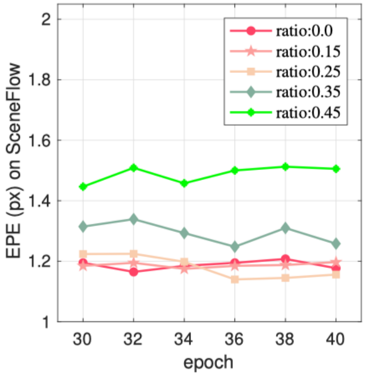
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA



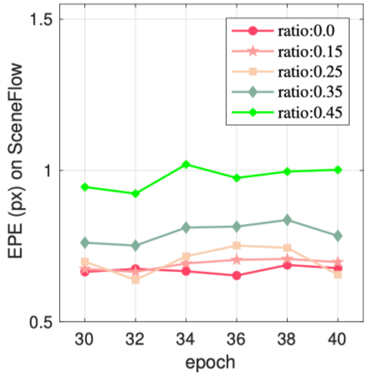

(a) CFNet (training)




(b) LacGwcNet (training)




(c) CFNet (testing)




(d) LacGwcNet (testing)

The convergence process with or without mask.

| Model | Mask | Training | Resolution | Runtime (s) |
|---|---|---|---|---|
| CFNet | ✔ | ✔ | $576 \times 320$ | 0.89 |
| CFNet | ✔ | ✘ | $960 \times 576$ | 0.052 |
| CFNet | ✘ | ✔ | $576 \times 320$ | 0.84 |
| CFNet | ✘ | ✘ | $960 \times 576$ | 0.051 |
| LacGwcNet | ✔ | ✔ | $576 \times 320$ | 1.63 |
| LacGwcNet | ✔ | ✘ | $960 \times 576$ | 0.264 |
| LacGwcNet | ✘ | ✔ | $576 \times 320$ | 1.61 |
| LacGwcNet | ✘ | ✘ | $960 \times 576$ | 0.264 |

The runtime with different resolutions.

*Conclusion:*

(1) the results are very stable on the source domain;

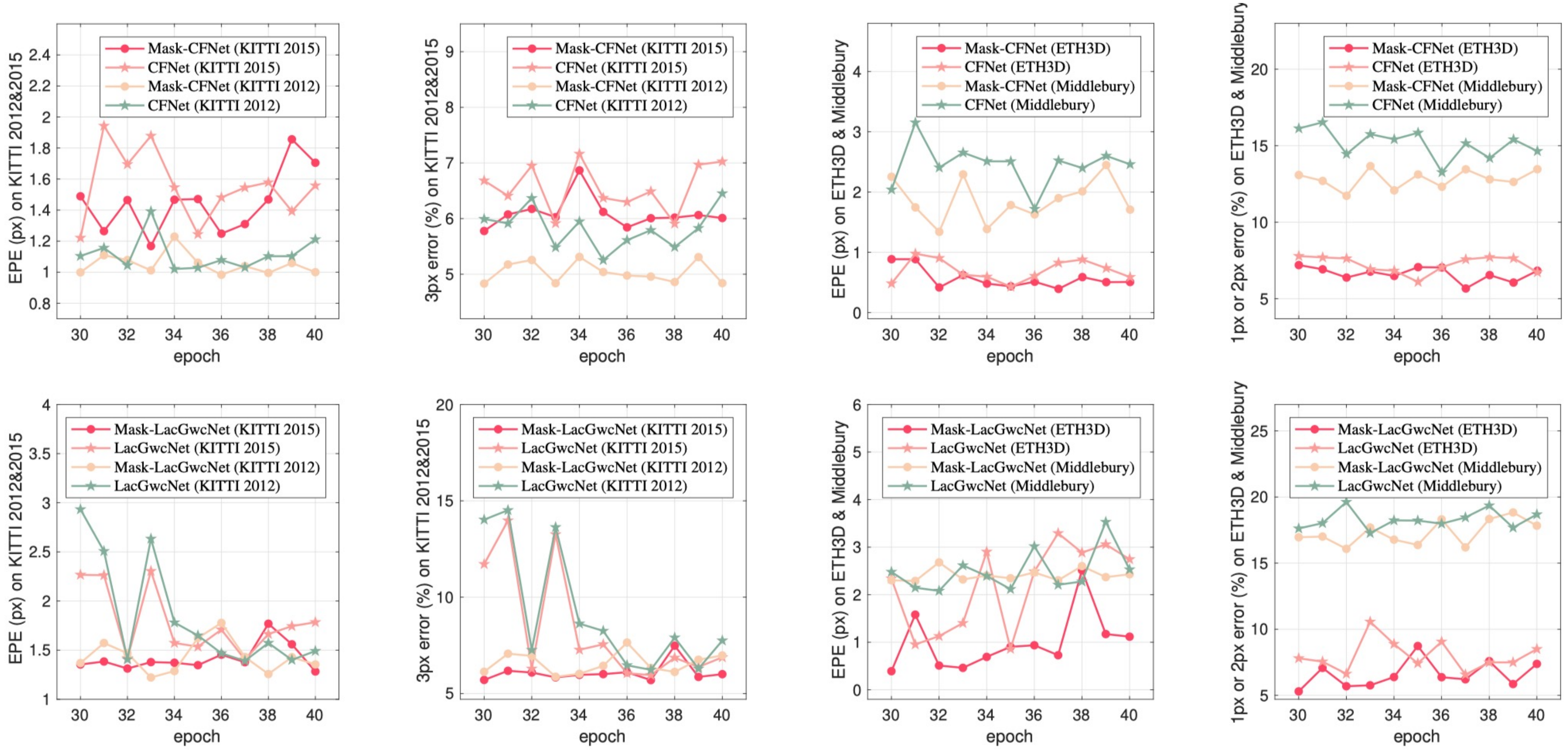
(2) our method does not change the convergence process for the low mask ratio;

(3) a high mask ratio will affect the learning process and reduce the matching accuracy;

(4) for the training process, our method does not significantly prolong training time compared with baselines;

(5) for the testing process, our approach is no different from baselines.

# Experiments


The generalization performance with or without masked representation among different epochs.

# Experiments



| 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** |

Volatility comparison of with or without masked representation.

*Conclusion:*

1. the generalization performance varies significantly between adjacent training epochs.
2. the models with masked representation learning can perform better and more stable.

# Experiments



| 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 |

Cross-domain generalization evaluation (peak results) on four target datasets.

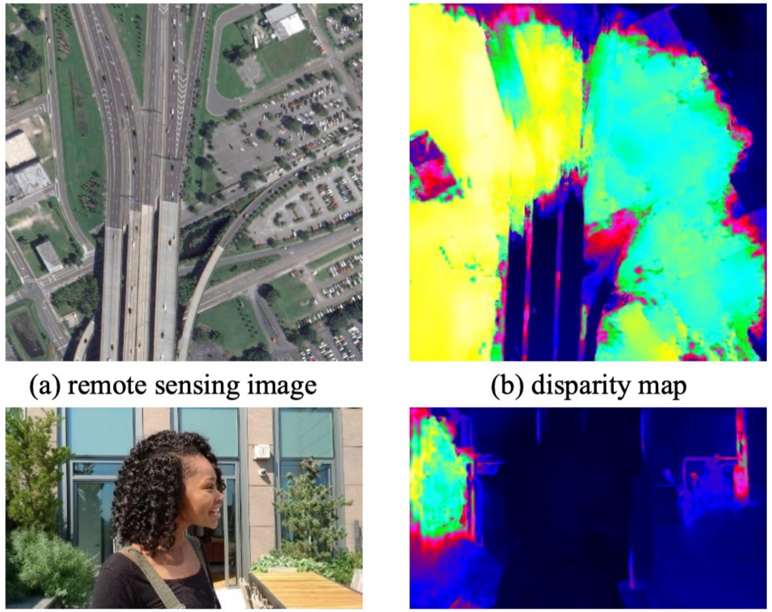
| Method | Ratio | KT-12 (Out-Noc) | KT-15 (D1-all) |
|---|---|---|---|
| LacGwcNet [15] | 0 | 1.13 | 1.77 |
| LacGwcNet [15] | 0.15 | 1.15 | 1.78 |
| LacGwcNet [15] | 0.25 | 1.16 | 1.77 |
| LacGwcNet [15] | 0.35 | 1.27 | 1.95 |
| LacGwcNet [15] | 0.45 | 1.39 | 2.21 |
| CFNet [30] | 0 | 1.23 | 1.88 |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

The fine-tuning results on the KITTI dataset.

*Conclusion:* Our method can help model improve cross-domain performance, but it seems no help for fine-tuning.

# Discussion





(a) remote sensing image (b) disparity map

(c) daily image (d) disparity map

The examples of failures in the real unseen domain

Today, many cross-domain stereo methods can do better in KITTI, ETH3D, Middlebury.
Does it mean we can use these algorithms in practice?
The answer is no.

Nearly all papers used KITTI, ETH 3D, and Middlebury datasets as the unseen domains.
However, can these datasets be represented the unseen domain?

# Conclusion

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

We have proposed a simple masked representation method to address the problem of unstable generalization performance among different training epochs.

- Our approach is more stable and better generalization performance.
- The experiments proved that the current evaluation manner is unsuitable, and we consider the stability should be evaluated in cross-domain methods.
- We discussed the failure of our approach and the topic about unseen domain.

# THANK YOU FOR WATCHING
# &
# Q. A