# GraftNet: Towards Domain Generalized Stereo Matching with a Broad-Spectrum and Task-Oriented Feature – Supplementary Material

Biyang Liu [1,2], Huimin Yu [1,2,3,4], Guodong Qi [1,2]
[1]College of Information Science and Electronic Engineering, Zhejiang University
[2]ZJU-League Research & Development Center, [3]State Key Lab of CAD&CG, Zhejiang University
[4]Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking
{biyangliu, yhm2005, guodong_qi}@zju.edu.cn

| Strategy | Finetune | Adaptor | EPE (px) | >3px |
|---|---|---|---|---|
| Jointly | ✗ | ✓ | 1.69 | 8.27% |
| Jointly | lr $10^{-3}$ | ✓ | 2.14 | 12.0% |
| Jointly | lr $10^{-4}$ | ✓ | 3.09 | 15.4% |
| - | lr $10^{-3}$ | ✗ | 1.87 | 10.3% |
| - | lr $10^{-4}$ | ✗ | 2.42 | 12.6% |
| Ours | ✗ | ✓ | **1.32** | **5.34%** |

Table 1. Effects of the training strategy of GraftNet. **Jointly** means the feature adaptor and the cost aggregation module are jointly trained. **Finetune**: whether the broad-spectrum feature is finetuned. If it is finetuned, the learning rate is reported. **Adaptor**: whether the feature adaptor is built in the model. The result of the separately training strategy used in our model is presented in the bottom. Results are evaluated on KITTI 2015.

| Dataset | KT-15 >3px | KT-12 >3px | MB >2px | ET >1px |
|---|---|---|---|---|
| SF | 5.3% | 5.0% | 10.9% | 10.7% |
| SF + vKT | **4.7%** | **4.9%** | **9.8%** | **9.8%** |

Table 2. Experimental results when integrating more training data. The first column represents the used source datasets, and the others are the evaluation results on the corresponding target datasets. **SF**: SceneFlow. **vKT**: Virtual KITTI 2. **KT-15**: KITTI 2015, **KT-12**: KITTI 2012, **MB**: Middlebury, **ET**: ETH3D.

| Resolution | w/o Adaptor | | w/ Adaptor | |
|---|---|---|---|---|
| | EPE (px) | >3px | EPE (px) | >3px |
| 288 × 288 | **1.82** | **6.05%** | **1.53** | 5.72% |
| 224 × 224 | 1.87 | **6.05%** | 1.61 | **5.68%** |
| 160 × 160 | 1.89 | 6.08% | 1.55 | 5.84% |

Table 3. Effects of the input resolution for training the broad-spectrum feature on the performance of GraftNet with or without the feature adaptor. Results are evaluated on KITTI 2015.

## 1. Supplementary Experiments

In the following supplementary experiments, PSMNet [2] is utilized as the basic stereo matching architecture and other details are as same as those discussed in the paper.

### 1.1. Different Training Strategies

As discussed in Sectionn 3.3 of the paper, separately training the feature adaptor and the cost aggregation module is more effective than jointly training. In this section, more detailed results are provided. We also evaluate the model performances when the broad-spectrum feature is finetuned with different learning rates (lr). When finetuned, the feature will learn to recover the task-related information, thus we consider discarding the feature adaptor.

Comparing the first row and the last row of Table 1, the separately training strategy used in our model is more effective. We analyze when the feature adaptor and the cost aggregation module are optimized individually, a trained module can provide a beneficial initialization for the other one. As shown in the *2nd* row to the *5th* row, finetuning the broad-spectrum feature will affect the robustness, although the feature adaptor is not necessary in this condition.

### 1.2. More Source Images

In the paper, the source images are only from SceneFlow [4]. In this section, we additionally introduce Virtual KITTI 2 [1], which is a synthetic dataset simulating the KITTI scenes. Virtual KITTI 2 contains 1594 training pairs with dense disparity ground truth.

As shown in Table 2, integrating more synthetic images can improve the cross-domain performance, even for the scenes (**Middlebury** & **ETH3D**) that are not close to the training datasets (**Virtual KITTI 2**). This indicates that richer source data might also be beneficial to our method.

### 1.3. Different Input Resolutions When Training the Broad-spectrum Feature

The input resolution for training a classification model on ImageNet [3] is normally set as $224 \times 224$, while the feature of a stereo matching network is trained with a larger resolution (*e.g*. $512 \times 256$ for PSMNet). In this section, we study whether the input resolution for training the broad-spectrum feature will affect the performance of GraftNet. Specifically, we retrain VGG [5] with 3 input resolutions, *i.e*. $288 \times 288$, $224 \times 224$, and $160 \times 160$.

The evaluation results on KITTI 2015 are presented in Table 3. From the table, training the broad-spectrum feature with a larger input resolution is beneficial for the subsequent stereo matching task only when the feature adapter is not utilized. This is exactly what we expect the feature adaptor to do, *i.e*. transforming the feature to make it more suitable for our stereo matching task.

## References


[1] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 1

[2] 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

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[4] 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. 1

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2