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

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

Abstract

Although supervised deep stereo matching networks have made impressive achievements, the poor generalization ability caused by the domain gap prevents them from being applied to real-life scenarios. In this paper, we propose to leverage the feature of a model trained on large-scale datasets to deal with the domain shift since it has seen various styles of images. With the cosine similarity based cost volume as a bridge, the feature will be grafted to an ordinary cost aggregation module. Despite the broad-spectrum representation, such a low-level feature contains much general information which is not aimed at stereo matching. To recover more task-specific information, the grafted feature is further input into a shallow network to be transformed before calculating the cost. Extensive experiments show that the model generalization ability can be improved significantly with this broad-spectrum and task-oriented feature. Specifically, based on two well-known architectures PSMNet and GANet, our methods are superior to other robust algorithms when transferring from SceneFlow to KITTI 2015, KITTI 2012, and Middlebury. Code is available at https://github.com/SpadeLiu/Graft-PSMNet.

1. Introduction

As a low-cost means to acquire depth, stereo matching has been studied as a fundamental problem in the vision society for decades. Given a rectified image pair, the objective is to search for the corresponding points and calculate their disparities. Stereo matching algorithms generally involve four steps [27]: matching cost computation, cost aggregation, disparity optimization, and disparity refinement.

Although Convolutional Neural Network (CNN) based supervised stereo matching methods have achieved admirable performances, huge amounts of annotated data are required to train the models, which is cumbersome for real-life applications. Synthetic data [20] is sufficient while the domain gap between the source and the target images prevents the models from generalizing well. There are three solutions to this issue: the unsupervised image reconstruction loss [34, 35], domain adaptation techniques [18, 31], and domain generalized approaches [2, 45]. In this paper, we focus on the third situation which is more challenging since the target images are not available during training.

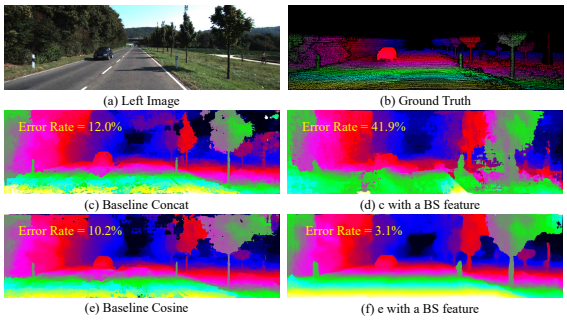


Figure 1. The toy experiment to validate the grafting operation. Two models with the cost volume formed by feature concatenation (Subfigure **c**) and cosine similarity (Subfigure **e**) are trained on SceneFlow, then their feature extraction modules are replaced with a **B**road-**S**pectrum feature (Subfigure **d** and **f**). For the four models, the 3-pixel error rates on a KITTI sample are labeled.

In domain generalized stereo matching, feature representation plays a crucial role [45] since the feature extraction module directly confronts images from different domains. Then a question is raised: can the goal be achieved by replacing the feature extraction module of an ordinary stereo matching network (*ordinary* means it is trained with synthetic data) with a broad-spectrum feature (*i.e.* the feature of a model trained on large-scale datasets)? Since this feature has seen various styles of images and learns to generalize well. In traditional algorithms, various feature descriptors [13] and cost aggregation methods [12, 42] could be

combined with each other to use. However in deep frameworks, the parameterized modules are entangled through end-to-end training, is this grafting operation (*i.e*. combining two trained modules without finetuning) practical?

To answer this question, we first conduct a toy experiment. With PSMNet [3] as the basic architecture, we train a model on a synthetic dataset SceneFlow [20], then its feature extraction module is replaced with the feature of VGG [30] trained on ImageNet [7]. Finally, the cross-domain performances are evaluated on KITTI 2015 [21]. As illustrated in Subfigure (c) and (d) of Figure 1, simply grafting a broad-spectrum feature to an ordinary cost aggregation module leads to a collapse of the disparity result. We analyze this is caused by the feature concatenation based cost volume, which forces the cost aggregation module to learn to measure the similarity based on the feature. When the feature is replaced, the learned metric will not be effective.

In order to disentangle the feature extraction module and the cost aggregation module, it is necessary to construct a generalized cost space [2]. On one hand, the cost volume should contain pure similarity information. In this way, the prior knowledge about the similarity metric is injected, preventing the cost aggregation module from overfitting to the used feature. Besides, the semantical information [10] which may interfere with cost aggregation due to the varied semantic classes of different domains is discarded. On the other hand, integrating the normalization of the cost value is beneficial for the generalization ability [31]. To this end, we utilize the elegant *cosine similarity* to construct the cost volume. In addition to satisfying the above demands, cosine similarity projects features with arbitrary channels to a scalar, making the cost accessible for various features.

Owing to the generalized cost space, when the feature extraction module trained with synthetic data is replaced with a broad-spectrum feature, the cross-domain performance is improved significantly, as shown in Subfigure (e) and (f) of Figure 1. This also experimentally validates that a broad-spectrum feature can be employed to handle the domain shift. However, grafting such a low-level feature of the classification model is still suboptimal since it contains much general information that serves various tasks. It is necessary to adapt the grafted feature to our stereo matching task. Inspired by the researches in multi-task learning [15] and transfer learning [24], we build a shallow network and force it to recover more task-specific information from the grafted feature. Although this training process is conducted on the source domain, the feature adaptor is robust since its input, the broad-spectrum feature has weakened the influence of the image style. Besides, the small amount of the parameters will reduce the risk of overfitting [38].

In summary, there are two fundamental steps in our domain generalized stereo matching network GraftNet. Firstly, grafting a broad-spectrum feature (*i.e*. the feature of a model trained on large-scale datasets) to the cost aggregation module of an ordinary stereo matching network. Secondly, transforming the feature with a shallow network to recover the task-specific information. In practice, we find retraining the cost aggregation module with this transformed feature can further improve the performance. It is worth noting that our method can be built upon arbitrary stereo matching networks, the only modification is to construct the cost volume with cosine similarity. Without bells and whistles, our models based on PSMNet [3] and GANet [44] are superior to other robust and domain generalized algorithms when transferring from a synthetic dataset SceneFlow [20] to some realistic datasets such as KITTI 2015 [21], KITTI 2012 [8], and Middlebury [26].

2. Related Work

2.1. Deep Stereo Matching Networks

MC-CNN [43] firstly introduced CNN to stereo matching, where a siamese network was built to compute the matching cost of two patches. The subsequent studies involved multi-scale feature representation [5] and the acceleration of the similarity calculation [19]. Although a deep embedding is powerful, these works are limited by the traditional cost aggregation and disparity refinement steps.

DispNetC [20] was the first end-to-end stereo matching network, where the disparity was regressed from the correlation maps through 2D convolutions. This pipeline has been widely adopted since then. SegStereo [41] and EdgeStereo [32] designed multi-task frameworks to exploit the semantic clues and the edge information. AANet [40] integrated deformable convolution to adaptively aggregate the cost. Despite the low complexities, the performances of 2D-CNN based stereo matching networks are not superior.

Another common fashion is to build the cost volume by concatenating the left and the right features and aggregate the cost with 3D convolutions, which is initially proposed in GCNet [14]. PSMNet [3] further introduced the spatial pyramid pooling module to deal with textureless regions. For more effective cost aggregation, image content guided layers were designed in GANet [44]. Currently, LEAStereo [6], an architecture searched by a deep network, ranks top on the KITTI benchmarks. While 3D-CNN based stereo matching networks perform well on several datasets, the poor generalization abilities hinder their applications in real-life scenes. In this work, we show how to alleviate this problem with a domain-invariant and task-oriented feature.

2.2. Domain Generalized Stereo Matching

In domain generalized stereo matching, the model is agnostic to the image style of the target domain, thus it is more challenging than domain adaptation [18,31,34,35]. To achieve the goal, DSMNet [45] proposed domain normal-

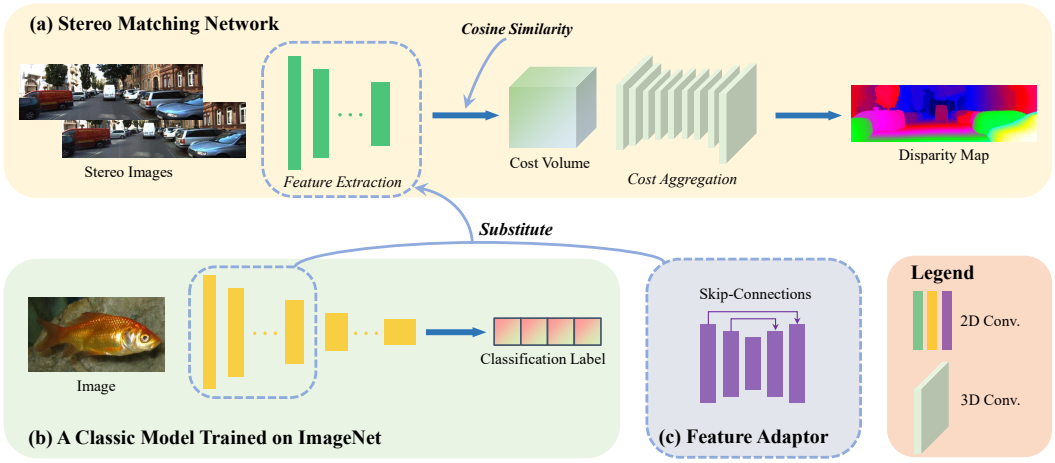


Figure 2. The overall architecture of GraftNet, consisting of a broad-spectrum feature, a feature adaptor, and a cost aggregation module. To build GraftNet, we first graft the feature from a classic model trained on the large-scale dataset ImageNet to the cost aggregation module of an ordinary stereo matching network. Then the feature is input to a shallow U-shape network to be transformed to recover more task-related information. Note the cost volume is formed by cosine similarity rather than feature concatenation to obtain a generalized cost space.

ization and a structure-preserving graph-based filter. CFNet [29] adaptively adjusted the search space to deal with the unbalanced disparity distribution across different domains. STTR [16] and RAFT-Stereo [17] introduced novel architectures which showed strong robustness. Our method is most similar to MS-Net [2], where traditional descriptors are utilized to construct the cost. However, in our work, the generalized matching space is realized with a deep feature which is more discriminative.

In addition, there are researchers devoted to tackling the problem from other perspectives. Poggi *et al*. [23] modulated the cost distribution with the sparse depth measurements obtained from some devices. Watson *et al*. [39] proposed an approach to generate labeled data from single images and showed that models trained on their MfS transferred better than those trained on SceneFlow [20].

2.3. Broad-Spectrum Features

The features of the classic architectures (*e.g*. ResNet [11], VGG [30]) trained on ImageNet [7] have been widely utilized to initialize the model parameters in several tasks [4, 9]. These pretrained classic models can be easily loaded from the libraries, *e.g*. PyTorch [22]. In our work, the feature is leveraged to obtain a robust representation since ImageNet covers various domains. To preserve the property of the feature, we keep the parameters fixed and build a network to transform the feature [15, 24].

3. Method

In this section, we will describe how to build our domain generalized stereo matching model GraftNet, whose key component is a broad-spectrum and task-oriented feature. The overall architecture is illustrated in Figure 2. Specifically, we first train a stereo matching network with the cost volume formed by cosine similarity (Section 3.1). Then we graft the feature from a classic model trained on ImageNet to the cost aggregation module of this stereo matching network and further transform it with a shallow network to recover the task-related information (Section 3.2). Finally, we empirically retrain the cost aggregation module with the transformed feature (Section 3.3).

3.1. Stereo Matching Network

In a typical deep stereo matching network [3, 6, 14, 44], the left and the right images are first passed through the feature extraction module, then a cost volume is constructed by concatenating the left and the right features at different displacements. After that, the cost is aggregated with several 3D convolutions, followed by *softmax* and *weighted average* to calculate the final disparity.

As demonstrated in [45], feature representation plays a crucial role in the generalization ability of the model. To this end, we intend to achieve the domain generalized stereo matching with a broad-spectrum feature. At the same time, the other parameterized part of a stereo matching network, the cost aggregation module, can only be trained with the synthetic data. Therefore, it is necessary to construct a generalized cost space [2] to disentangle the feature extraction module and the cost aggregation module.

In our model, the elegant *cosine similarity* is utilized to build the cost volume. Compared with feature concatenation, it has three advantages: 1) The semantical information [10] which is susceptible to the domain shift is eliminated, resulting in a cost volume with pure similarity in-

formation. 2) The normalization ensures the numerical stability of the cost values, which is beneficial for the cross-domain evaluation performance [31, 45]. 3) Features with arbitrary channels could be taken as the input since all of them will be projected to a scalar. Formally, the cosine similarity cost volume is expressed as:

$$\mathbf{CV}_{cos}(:,d,x,y) = \frac{\langle F^l(:,x,y), F^r(:,x-d,y)\rangle}{||F^l(:,x,y)||_2 \cdot ||F^r(:,x-d,y)||_2} \tag{1}$$

where d is the disparity index, and (x, y) denotes the pixel coordinate. F^l and F^r are the left and the right features, both with C channels. The calculated cost is a 4D tensor with only one channel, thus the input channel of the first 3D convolutional layer of the cost aggregation module should be modified to 1.

Without other adjustments, this basic stereo matching network is trained on the source domain with the cross entropy loss [36] and the smooth $L1$ loss [3] to supervise the disparity probability distributions and the final disparity values respectively:

$$L_{ce}(\hat{P}(d), P(d)) = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=0}^{d_{max}} -\hat{P}_i(d)\cdot \log P_i(d) \tag{2}$$

$$L_{sm}(\hat{D}, D) = \frac{1}{N}\sum_{i=1}^{N} smooth_{L_1}(\hat{D}_i, D_i) \tag{3}$$

where $\hat{P}(d)$ is the predicted distribution from *softmax* and $P(d)$ is the ground truth distribution, a normalized Laplacian distribution centered at the disparity ground truth D. $\hat{D}$ is the predicted disparity calculated by *weighted average*. N denotes the number of the pixels in an image.

Normally, there are multiple disparity results output from the cost aggregation module in the training phase [3, 10, 44]. In our model, each result is supervised with the above two loss functions, then the total loss is:

$$L = \sum_{m=1}^{M} \lambda_m (L_{ce} + \mu L_{sm}) \tag{4}$$

where M is the number of the disparity outputs. As for the balance weights, λ_m is set as same as in the adopted basic architecture, and μ is set to 0.1 heuristically.

After training, the feature extraction module is discarded since it is susceptible to the domain shift, while the cost aggregation module is reserved for grafting other features. Owing to the generalized cost space, the cost aggregation module is less affected by the domain gap.

3.2. Broad-Spectrum and Task-Oriented Feature

In this work, we employ the feature of a model trained on large-scale datasets to resist the domain shift since it has seen various styles of images and has learned to generalize well. In the meanwhile, such a feature is easy to acquire, *e.g.* the classic models [11, 30] trained on ImageNet [7] can be directly loaded from the PyTorch library [22]. Rather than utilizing the pretrained parameters to initialize the model backbones [4, 9], we keep the module fixed to preserve the inherent property of the feature.

Specifically, the broad-spectrum feature will be grafted to the trained ordinary cost aggregation module in Section 3.1. To keep consistency, we adopt the feature that has the same resolution as the one used in the original stereo matching network. For example, if the basic architecture is PSMNet [3] and the grafted feature is from VGG [30], then the feature before the third pooling layer that has the quarter resolution of the image will be employed.

Although the influence of the domain shift has been weakened by a broad-spectrum feature, a simple grafting operation is ill-considered. The reason is that the feature is relatively low-level, containing much general information that serves various downstream tasks. It is necessary to extract more information specific to our stereo matching task.

To this end, we build a feature adaptor before calculating the cost, *i.e.* the shallow U-shape network [25] illustrated in Figure 2 (c). The feature adaptor is trained as a part of the stereo matching network, in which process the parameters of the broad-spectrum feature and the cost aggregation module are fixed and only serve as the intermediums to propagate the gradients. Although this training process is conducted on the source domain, the feature adaptor is effective on the target domain for two reasons: 1) Its input is a broad-spectrum representation which will weaken the influence of the image style. 2) The small amount of the parameters will reduce the risk of overfitting [38].

3.3. GraftNet

With the broad-spectrum and task-oriented feature output from the feature adaptor, we find retraining the cost aggregation module can further improve the performance. In this step, our method is similar to [2], *i.e.* constructing a generalized matching space and training a cost aggregation module with the synthetic data. However, experimental results in Section 4.5 show that an appropriate deep feature is more representative than traditional descriptors [2].

From the perspective of the model architecture, GraftNet consists of three components: a broad-spectrum feature, a feature adaptor, and a cost aggregation module. Although we are inspired by the toy grafting experiment in Figure 1, can the feature adaptor and the cost aggregation module be trained together? In practice, we find jointly training is not as effective as separately training (Please refer to the supplementary material). We conjecture that when the two modules are optimized individually, a trained module can provide a beneficial initialization for the other one.

Model	Step	KITTI 2015		KITTI 2012		Middlebury		ETH3D	
		EPE (px)	>3px	EPE (px)	>3px	EPE (px)	>2px	EPE (px)	>1px
PSMNet	Baseline	3.24	19.5%	2.59	18.6%	6.69	22.6%	2.20	12.1%
	Cosine Similarity CV	2.98	15.4%	2.30	14.3%	6.83	22.3%	1.17	**10.6%**
	Graft VGG's Feature	1.86	6.39%	1.28	5.90%	5.67	18.9%	1.81	11.9%
	+ Feature Adaptor	1.47	5.60%	1.16	5.20%	2.96	12.0%	1.66	12.6%
	Retrain CA Module	**1.32**	**5.34%**	**1.09**	**4.97%**	**2.34**	**10.9%**	**1.16**	10.7%
GANet	Baseline	2.76	17.1%	2.35	12.8%	7.33	20.7%	0.46	7.80%
	Cosine Similarity CV	1.80	8.78%	1.70	8.57%	6.09	21.3%	0.46	7.08%
	Graft VGG's Feature	1.91	7.12%	1.91	8.37%	7.75	24.3%	0.86	13.2%
	+ Feature Adaptor	1.31	5.55%	1.19	5.11%	1.96	10.9%	0.45	6.67%
	Retrain CA Module	**1.30**	**5.35%**	**1.07**	**4.60%**	**1.87**	**8.89%**	**0.43**	**6.17%**

Table 1. Quantitative results of the ablation experiment. PSMNet and GANet-11 are the used two basic architectures. Models are trained on SceneFlow and evaluated on four realistic datasets. CV represents cost volume and CA denotes the cost aggregation module.

Since grafting is the first and fundamental step in the whole pipeline, our domain generalized stereo matching network is termed GraftNet. Moreover, we wish the grafting operation could provide a novel viewpoint: Can parts of two trained CNNs be integrated without finetuning to obtain a new model? This question is worth exploring, especially for scenarios where training data is not available.

4. Experiment

4.1. Datasets & Evaluation Metrics

Source Domain. In the experiment, all of the stereo matching networks are trained on **SceneFlow** [20], a synthetic dataset containing 35454 training pairs and 4370 testing pairs, both with dense disparity ground truth. Since only the generalization ability is concerned in the domain generalized problem, the test set will not be used.

Target Domain. The models trained on SceneFlow are evaluated on the following realistic datasets:

- **KITTI** datasets consist of KITTI 2015 [21] and KITTI 2012 [8], whose ground truth disparity maps are sparse. On KITTI 2015, there are 200 training pairs and 200 testing pairs. On KITTI 2012, there are 194 training pairs and 195 testing pairs.
- **Middlebury 2014** [26] provides 15 training pairs and 15 testing pairs, where some samples are under inconsistent illumination or color conditions. All of the images are available in three different resolutions, we select the half-resolution ones.
- **ETH3D** [28] is a gray-scale dataset with 27 training pairs and 20 testing pairs.

For all the realistic datasets, we use their training sets to evaluate the cross-domain performance. The utilized metrics are **EPE** (End Point Error, the mean average error) and **τ-pixel error rate** (percentage of the points with absolute error larger than τ pixel).

Cost	Normal.	P. Simi.	EPE (px)	>3px
Concat	✗	✗	3.24	19.5%
N_Concat	✓	✗	3.14	18.3%
L2 Distance	✗	✓	**2.86**	16.2%
Cosine Simi.	✓	✓	2.98	**15.4%**

Table 2. Effects of the manner of building the cost volume. **Normal.**: whether the features are normalized before building the cost. **P. Simi.**: whether the cost contains pure similarity information. **N_Concat** means the cost is formed by concatenating the normalized features. Results are evaluated on KITTI 2015.

4.2. Implementation Details

The framework is implemented on PyTorch [22], with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) as the optimizer. For the basic stereo matching architecture, we train it for 8 epochs with a learning rate of 0.001. Then a broad-spectrum feature is grafted to the cost aggregation module, in which process no training is involved. After grafting, the feature adaptor is trained for 1 epoch with a learning rate of 0.001. Finally, the cost aggregation module is retrained for 10 epochs with the learning rate set as 0.001 for the first 5 epochs and 0.0001 for the remaining epochs.

For all experiments, PSMNet [3] is adopted as the basic architecture. In the ablation study (Section 4.3) and the comparison experiment with other robust algorithms (Section 4.5), GANet-11 [44] is additionally utilized to demonstrate the effectiveness and versatility of our method. The grafted feature is from VGG16 [30] which is trained on ImageNet [7], and in Section 4.4 more features are explored.

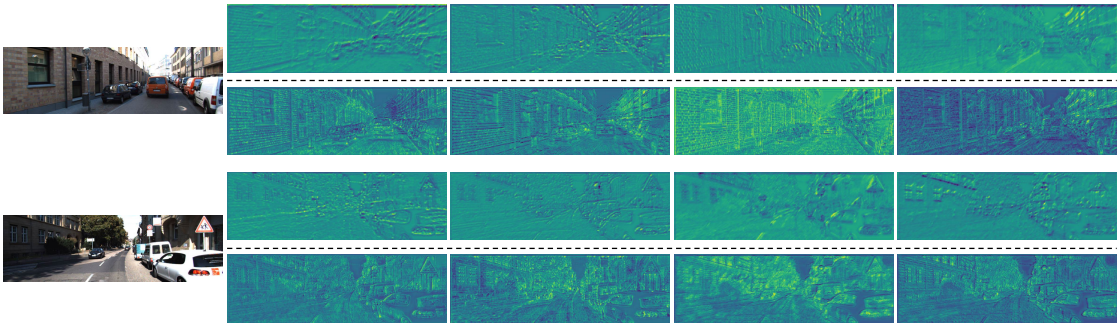


Figure 3. Illustration of the adapted features. For each image, the top row shows four channels of the features before adaption (*i.e*. VGG's feature) and the bottom row shows four channels of the features after adaption. Images are from KITTI 2015.

4.3. Ablation Study

In this section, we study the effects of the components in GraftNet, the evaluation results are listed in Table 1. First of all, although feature concatenation is commonly used to build the cost in supervised frameworks, cosine similarity is more suitable for domain generalized stereo matching. There are two reasons: 1) the normalization keeps the values stable, 2) the semantical information which is susceptible to the domain shift is discarded.

For the purpose of deeply investigating the influences of these two aspects, we compare several manners of building the cost in Table 2. Results show that a cost volume containing pure similarity information (*e.g*. calculated by cosine similarity or $L2$ distance) is better when generalization ability is considered. We also emphasize that cosine similarity allows us to graft features from other models. If the cost volume is constructed by feature concatenation, the disparity result of the assembled model will be collapsed.

From the *2nd* and the *3rd* rows of the two subtables in Table 1, grafting the feature of VGG trained on ImageNet is beneficial for KITTI datasets. This indicates the overfitting problem of current stereo matching networks does exist, and a broad-spectrum feature can improve the generalization ability. However, this feature does not work for Middlebury (when equipped with GANet) and ETH3D. We analyze although a domain-generalized representation is obtained, it does not fit the task, the low-level feature from the classification model contains seldom detailed information. Coincidentally, on KITTI, the ground truth is sparse especially at the disparity discontinuities, thus a feature with more global context information is effective as well.

To restore the task-specific information, we build a shallow network to transform the broad-spectrum feature. From the *4th* and the *9th* rows of Table 1, the evaluation performances on Middlebury and ETH3D are improved significantly with the feature adaptor. This result suggests that both the *broad-spectrum* property and the *task-oriented* property of the feature is important. In Figure 3, we further exhibit the features before and after adaption. As it can be seen, rich texture information which is essential for stereo matching is recovered with the adaptor.

Architecture	EPE (px)	>3px
✗	1.86	6.39%
Linear	1.50	6.06%
Non-Linear	**1.46**	6.20%
U-Net	1.47	**5.60%**

Table 3. Effects of the architecture of the feature adaptor. The first row means no adaptor is utilized. The linear adaptor is a single convolutional layer and the non-linear adaptor consists of two convolutional layers and an activation layer. U-Net is the architecture shown in Figure 2 (c). Results are evaluated on KITTI 2015.

In Table 3, we compare different architectures of the feature adaptor. Although a linear layer is enough in [15], in our work a more complex network is needed since not only the task gap but also the feature level should be considered. In the meanwhile, the parameter number of the adaptor can not be too large to prevent overfitting to the source data. Therefore, a shallow U-shape network [25] is adopted.

Finally, in Table 1, retraining the cost aggregation module with the adapted feature can further improve the evaluation performance. The reason might be that compared with the original feature trained on the source domain, the broad-spectrum and task-oriented feature provides a more robust cost volume, guiding the optimization of the cost aggregation module towards the goal of domain generalized stereo matching. Some qualitative results of our final model on the four realistic datasets are displayed in Figure 4.

4.4. Grafting Various Features

In this section, we attempt to graft various features to further investigate the impact of the feature. Six features are adopted: VGG16 [30], ResNet18 [11], and ResNet50 [11] trained for Classification (C) on ImageNet, ResNet18 trained for Monocular Depth Estimation (MDE) [9] on

Feature	Task	Dataset	EPE (px)	>3px
VGG16	C	ImageNet	1.86	6.39%
ResNet18	C	ImageNet	1.90	6.62%
ResNet18	MDE	KITTI	1.73	6.54%
ResNet50	C	ImageNet	2.06	**6.19%**
ResNet50	DCL	ImageNet	2.20	9.17%
ResBlocks	OFE	KITTI	**1.59**	6.22%

Table 4. Experimental results of grafting features from various models. **C**: classification, **MDE**: monocular depth estimation, **DCL**: dense contrastive learning, **OFE**: optical flow estimation. Results are evaluated on KITTI 2015, the best is shown in **bold** and the second is underlined.

KITTI, ResNet50 trained by Dense Contrastive Learning (DCL) [37] on ImageNet, stacked ResBlocks trained for Optical Flow Estimation (OFE) [33] on KITTI.

The qualitative results evaluated on KITTI 2015 are presented in Table 4. Comparing the *2nd* row and the *3rd* row, a broad-spectrum feature performs close to a domain-specific feature, meaning that the domain shift can be handled with the feature of a model trained on large-scale datasets. From the *3rd* and the *5th* rows, although MDE and DCL are dense prediction tasks, their features cannot satisfy the needs of stereo matching. The feature trained for a closer task OFE might be more helpful, while there is still a performance gap between it and the feature used in our model (**6.22%** vs **5.60%**). These conclusions once again stress the importance of the task-oriented property of the feature.

In addition, considering image classification models and stereo matching models are usually trained with different input resolutions, and the input resolution is vital for the pixelwise task stereo matching, we study the effect of the input resolution when training the broad-spectrum feature. Please refer to the supplementary material for more results.

4.5. Comparison with Robust Algorithms

In this section, we compare our models with other robust and domain generalized methods. As reported in [16, 39], augmenting images with a random color and brightness transform can improve the model generalization ability. Therefore, for a fair comparison, the algorithms are separated into two categories according to whether data augmentation strategies including *color jitter* are involved.

As shown in Table 5, among the methods that do not utilize the random color transform strategy, our Graft-PSMNet and Graft-GANet are superior, especially on KITTI and Middlebury. When integrating more data augmentation approaches, the model performance can be further boosted. On ETH3D, our models are not the best, we analyze the reason is that ImageNet contains few gray-scale images, making the grafted feature difficult to express well on ETH3D. This inspires us that more styles of images should be collected for an absolutely domain-invariant representation.

Model	KT-15 >3px	KT-12 >3px	MB >2px	ET >1px
GwcNet [10]	22.7%	20.2%	37.9%	54.2%
PSMNet [3]	16.3%	15.1%	34.2%	23.8%
GANet [44]	11.7%	10.1%	20.3%	14.1%
MS-PSMNet [2]	7.8%	14.0%	19.8%	16.8%
MS-GCNet [2]	6.2%	5.5%	18.5%	8.8%
DSMNet [45]	6.5%	6.2%	13.8%	**6.2%**
Graft-PSMNet	**5.3%**	5.0%	10.9%	10.7%
Graft-GANet	5.4%	**4.6%**	**8.9%**	**6.2%**
CFNet* [29]	6.0%	5.1%	15.4%	5.3%
RAFT-Stereo* [17]	5.7%	-	12.6%	**3.3%**
SGM+NDR [1]	5.5%	6.0%	12.4%	4.8%
Graft-PSMNet*	**4.8%**	4.3%	**9.7%**	7.7%
Graft-GANet*	4.9%	**4.2%**	9.8%	6.2%

Table 5. Comparison of robust and domain generalized stereo matching methods, ours are listed at the bottom of the two subtables. * means the color jitter data augmentation strategy is leveraged during training. **KT-15**: KITTI 2015, **KT-12**: KITTI 2012, **MB**: Middlebury, **ET**: ETH3D. The best result is shown in **bold** and the second result is underlined.

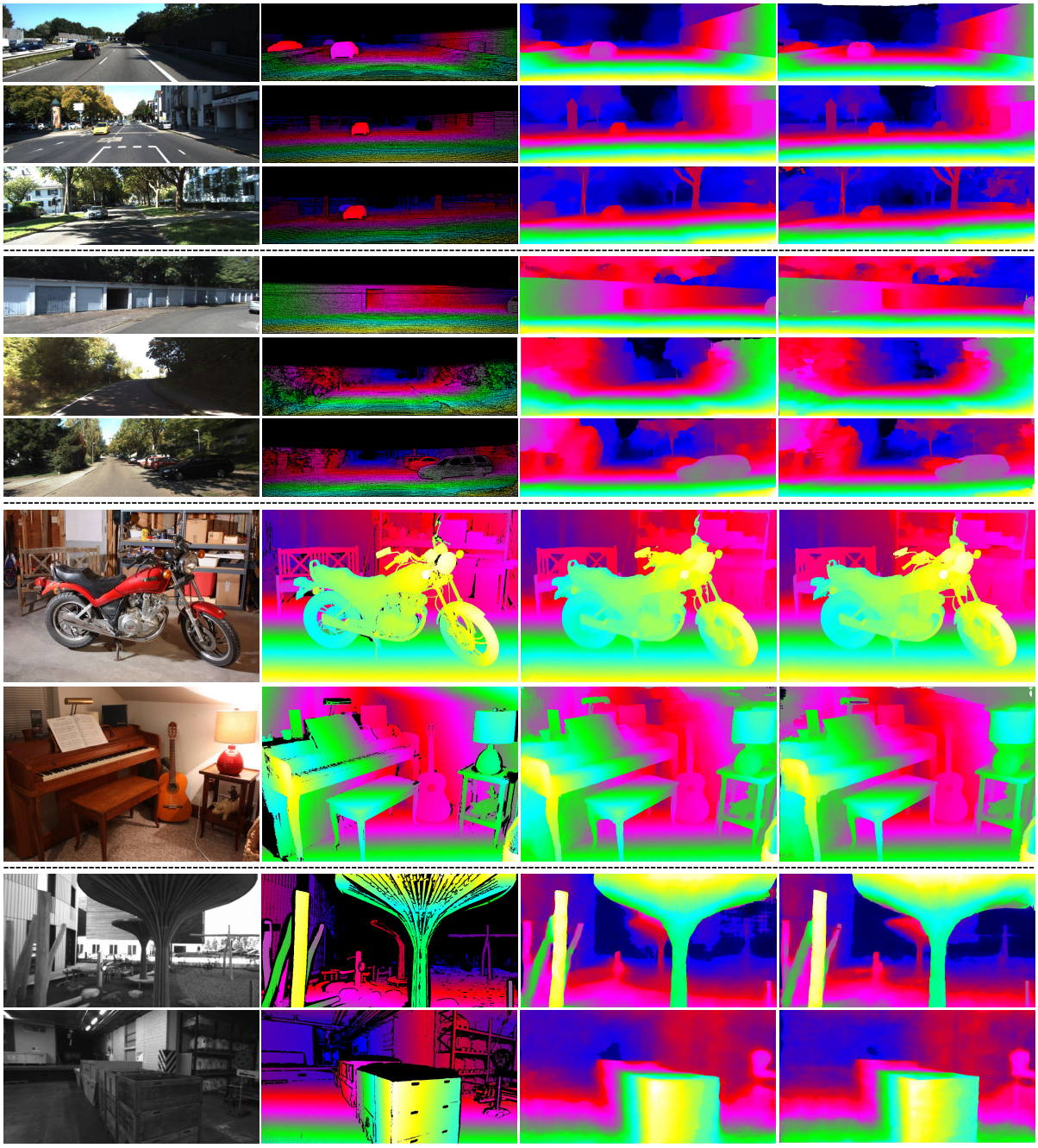
5. Limitation & Future Work

There are two main limitations of our work: 1) As discussed in Section 4.5, the grafted feature is not perfectly domain-invariant. 2) The annotations for image classification are implicitly used through loading the parameters of the model trained on ImageNet, meaning that additional labeled data (but not limited to stereo matching) are needed.

Aiming at these limitations, we intend to deeply combine self-supervised representation learning with our GraftNet in the future. By this means, only images are required and huge amounts of data from the Internet can be leveraged to improve the robustness of the learned feature.

6. Conclusion

This paper attempts to achieve domain generalized stereo matching from the perspective of *data*, where the key is a broad-spectrum and task-oriented feature. The former property comes from the various styles of images seen during training, and the latter property is realized by recovering task-related information from the broad-spectrum feature. By constructing a generalized cost space with cosine similarity, the feature is combined with an ordinary cost aggregation module. Experimental results on several datasets show that our Graft-PSMNet and Graft-GANet are superior to other robust and domain generalized algorithms. We hope our method can inspire subsequent studies, including multi-task learning and domain generalized approaches.



(a) Left Image (b) Ground Truth (c) Graft-PSMNet (d) Graft-GANet

Figure 4. Qualitative results of Graft-PSMNet and Graft-GANet when transferring from SceneFlow to KITTI 2015, KITTI 2012, Middlebury, and ETH3D (from top to bottom). Best viewed in color.

Societal Impact

The fundamental purpose of our work is to improve the quality of the depth obtained by stereo systems, which plays an essential role in many industries such as robot navigation and autonomous driving. On the one hand, the developments of these industries bring huge convenience to human activities. On the other hand, there are still lots of questions to be answered about security, emotion, etc. To handle these potential problems, we must not only carefully evaluate the security of the AI (Artificial Intelligence) systems, but also establish and complete some related laws.

References

[1] Filippo Aleotti, Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement for arbitrary resolution stereo. *arXiv preprint arXiv:2110.15367*, 2021. 7

[2] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization. In *2020 International Conference on 3D Vision (3DV)*, pages 364–373. IEEE, 2020. 1, 2, 3, 4, 7

[3] 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. 2, 3, 4, 5, 7

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3, 4

[5] Zhuoyuan Chen, Xun Sun, Liang Wang, Yinan Yu, and Chang Huang. A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 972–980, 2015. 2

[6] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *arXiv preprint arXiv:2010.13501*, 2020. 2, 3

[7] 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, 3, 4, 5

[8] 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, 5

[9] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 3, 4, 6

[10] 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. 2, 3, 4, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4, 6

[12] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005. 1

[13] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1

[14] 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. 2, 3

[15] Wei-Hong Li and Hakan Bilen. Knowledge distillation for multi-task learning. In *European Conference on Computer Vision*, pages 163–176. Springer, 2020. 2, 3, 6

[16] 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*, pages 6197–6206, 2021. 3, 7

[17] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. *arXiv preprint arXiv:2109.07547*, 2021. 3, 7

[18] 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*, pages 12757–12766, 2020. 1, 2

[19] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016. 2

[20] 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, 2, 3, 5

[21] 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, 5

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 3, 4, 5

[23] Matteo Poggi, Davide Pallotti, Fabio Tosi, and Stefano Mattoccia. Guided stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 979–988, 2019. 3

[24] Pierluigi Zama Ramirez, Alessio Tonioni, Samuele Salti, and Luigi Di Stefano. Learning across tasks and domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8110–8119, 2019. 2, 3

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4, 6

[26] 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. 2, 5

[27] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002. 1

[28] Thomas Schöps, Johannes L. Schönberger, 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[29] 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*, pages 13906–13915, 2021. 3, 7

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3, 4, 5, 6

[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*, pages 10328–10337, 2021. 1, 2, 4

[32] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, 128(4):910–930, 2020. 2

[33] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 7

[34] 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*, pages 9661–9670, 2019. 1, 2

[35] 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*, pages 195–204, 2019. 1, 2

[36] Stepan Tulyakov, Anton Ivanov, and François Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. *Advances in Neural Information Processing Systems*, 31:5871–5881, 2018. 4

[37] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 7

[38] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020. 2, 4

[39] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. In *European Conference on Computer Vision*, pages 722–740. Springer, 2020. 3, 7

[40] 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*, pages 1959–1968, 2020. 2

[41] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 636–651, 2018. 2

[42] Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):650–656, 2006. 1

[43] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015. 2

[44] 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. 2, 3, 4, 5, 7

[45] 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. 1, 2, 3, 4, 7