AdaStereo: A Simple and Efficient Approach for Adaptive Stereo Matching

Xiao Song\textsuperscript{1}*\textsuperscript{,} Guorun Yang\textsuperscript{1}*\textsuperscript{,} Xinge Zhu\textsuperscript{2}, Hui Zhou\textsuperscript{1}, Zhe Wang\textsuperscript{1}, Jianping Shi\textsuperscript{1}

\textsuperscript{1} SenseTime Group Limited
\textsuperscript{2} The Chinese University of Hong Kong
\{songxiao,yangguorun,zhouhui,wangzhe,shijianping\}@sensetime.com
zhuxinge123@gmail.com

Abstract. In this paper, we attempt to solve the domain adaptation problem for deep stereo matching networks. Instead of resorting to black-box structures or layers to find implicit connections across domains, we focus on investigating adaptation gaps for stereo matching. By visual inspections and extensive experiments, we conclude that low-level aligning is crucial for adaptive stereo matching, since main gaps across domains lie in the inconsistent input color and cost volume distributions. Correspondingly, we design a bottom-up domain adaptation method, in which two particular approaches are proposed, i.e. color transfer and cost regularization, that can be easily integrated into existing stereo matching models. The color transfer enables transferring a large amount of synthetic data to the same color spaces with target domains during training. The cost regularization can further constrain both the lower-layer features and cost volumes to domain-invariant distributions. Although our proposed strategies are simple and have no parameters to learn, they do improve the generalization ability of existing disparity networks by a large margin. We conduct experiments across multiple datasets, including Scene Flow, KITTI, Middlebury, ETH3D and Driving-Stereo. Without whistles and bells, our synthetic-data pretrained models achieve state-of-the-art cross-domain performance compared to previous domain-invariant methods, even outperform state-of-the-art disparity networks fine-tuned with target domain ground-truths on multiple stereo matching benchmarks.

Keywords: Domain adaptation, stereo matching, color transfer, cost regularization

1 Introduction

Stereo matching is a fundamental problem in computer vision. The task aims to find corresponding pixels in stereo images, and the distance between corresponding pixels is known as disparity \cite{24}. Based on the epipolar geometry, stereo matching enables stable depth perception from estimated disparity, so that it

* indicates equal contribution.
Fig. 1: Examples of our AdaStereo. Left-right: left image, disparity map predicted by PSMNet [3] pretrained on the SceneFlow dataset [39], and disparity map predicted by our trained Ada-PSMNet. Top-down: Middlebury [46] and KITTI [40] examples.

supports further applications such as scene understanding [15,66,51,41], object detection [4,5,32,31], odometry [58,72] and SLAM [13,18] et al.

Recent stereo matching methods [39,29,43,34,3,62,52,23,67] adopt fully convolutional networks [36] to directly regress disparity map. These models have achieved pretty high accuracy on several benchmarks [16,40,46]. However, most of the methods are prone to overfit on the training dataset. As shown in Fig 1, the PSMNet [3] pretrained on the SceneFlow dataset [39], is able to predict high-quality disparity maps on the synthetic images, but fails to generate equally good results on the Middlebury [46] and KITTI [40] datatsets. Hence instead of designing powerful models for higher precision on specific datasets, how to obtain effective adaptive stereo matching models is more desirable and meaningful now.

In this paper, we address the important but less explored problem of adaptive stereo matching. Considering the fact that there are a great number of synthetic data but only a small amount of realistic data with ground-truths, the problem can be further limited to the adaptation from synthetic to realistic domains. Here, different from general utilization of adaptation methods, we first explore the inherent factors which hinder the adaptation performance of stereo matching models. Due to the low-level preference of stereo matching task, we investigate several low-level representations and calculate some statistics as shown in Fig. 2. On this basis, we find that low-level aligning across domains is crucial since the domain gaps mainly lie in the input color and cost volume distribution. Specifically, in terms of the image appearance, color differences are obvious across domains, which plays a crucial role in the domain gap. In terms of the cost volume computed from left and right lower-layer features, the distributions of cost values are inconsistent across domains as well. Therefore, we propose two particular approaches for domain-adaptive stereo matching: color transfer and cost regularization.

Firstly, for color transfer, a practical algorithm is proposed to align the colors of input images from different domains. Given a target image, the algorithm performs color transferring in the LAB space without extra learning parameters.
During model training, the algorithm can be seamlessly embedded into the input module of the stereo matching framework. Rather than the classical pipeline from pre-training to fine-tuning, our color transfer enables the training directly conducted on massive transferred data, consequently the model can naturally adapt to the expected domain.

Secondly, for cost regularization, we successively perform normalization on each channel and each spatial position in the lower-layer features. When the left and right lower-layer features used for building cost volumes are normalized, the generated cost volumes are regularized to a certain distribution as well, so that it makes the deep neural network more insensitive to potential domain shifts. Besides, since the cost regularization is operated on lower-layer features, it can be easily integrated into most of the existing stereo networks. Compared with other normalization approaches (e.g. IN [45], GN [59]), our regularization method is designed specifically for stereo matching task, which is imposed only once in the disparity network without any learning parameters.

In essence, both of the color transfer and cost regularization are proposed to bridge the gap between two distinct domains and finally benefit stereo matching models. Compared with other domain-invariant approaches, our method focuses on stereo matching task and realize domain adaptation in an intuitive way. We validate the effectiveness of our method on diverse datasets including Middle-Burry [46], ETH3D [48], KITTI [16,40] and DrivingStereo [61], which covers indoor, outdoor and driving scenarios. Furthermore, our method outperforms other domain-invariant methods and even fine-tuned models on multiple stereo benchmarks. The main contributions are summarized below:

– We locate the domain-adaptive problem for deep stereo matching networks that low-level aligning across domains is crucial for adaptive stereo matching.
– Two approaches, color transfer and cost regularization, are proposed to narrow the domain gaps in terms of the color space and cost volume respectively.
– Our domain-adaptive models outperform other domain-invariant methods, and even state-of-the-art disparity networks fine-tuned with target domain ground-truths on multiple stereo matching benchmarks.

2 Related Work

2.1 Stereo Matching

Deep neural networks first achieved great success in the stereo matching task by representing image patches with deep features and computing patch-wise similarity scores [65,38,49,9] from a siamese network. However, due to the limited receptive field and hand-engineered post-processing steps, performance of these methods is unsatisfactory especially in ill-posed regions, while their generalization ability is also poor because of the poor capacity of feature representations.

Recently end-to-end disparity networks achieved state-of-the-art performance, which could be roughly categorized into two types: correlation-based 2-D stereo networks and cost-volume based 3-D stereo networks. For the first category, Mayer
et al. [39] proposed the first end-to-end disparity network DispNetC, in which warping between features of two views is conducted for matching cost calculation, and per-pixel disparity is directly regressed without any post-processing steps. Based on color or feature correlations, more advanced methods were proposed, including CRL [43], iResNet [33], HD^3 [63], SegStereo [62] and EdgeStereo [52, 50].

For the second category, 3-D convolutional neural networks show the advantages in regularizing cost volume for disparity estimation. GC-Net [29] first introduced the 4-D cost volume without collapsing the feature dimension when assembling a stereo cost volume. PSMNet [3] further applied a pyramid feature extraction module and stacked 3-D hourglass blocks to improve the accuracy of stereo matching. Recently, more advanced methods were proposed to further optimize the cost volume regularization, including GwcNet [23], EMCAU [42], CSPN [10] and GANet [67]. Our work can be seen as complementary to the above methods because the proposed approaches can be easily applied to the existing disparity networks to improve their performance on new domains.

2.2 Domain adaptation

Domain Adaptation for High-level Tasks. Conventional methods include Maximum Mean Discrepancy (MMD) [17, 37], geodesic flow kernel [19], sub-space alignment [14], asymmetric metric learning [30], et al. For semantic segmentation domain adaptation, the pioneering work is [27], which combined global and local alignment methods with domain adversarial training. Another work [69] applied the curriculum learning to solve domain adaptation from easy to hard. In [6], authors proposed an unsupervised learning method to adapt road-scene segmenters across different cities. In [56, 7, 20, 74, 53], authors performed output-space adaptation at feature level by an adversarial module. [8, 73] further incorporated the adversarial alignment into the domain adaptation for object detection.

Domain Adaptation for Low-level Tasks. Much less attention has been given to domain adaptation for low-level tasks. Some works [22, 70, 54] about depth estimation domain adaptation were proposed, including meta-learning for pre-training, offline style-transfer from synthetic to real and pseudo-label based fine-tuning. However, most of these methods directly adopt high-level adaptation techniques, without in-depth analyses of key differences in low-level tasks.

2.3 Domain-Adaptive Stereo Matching

Although the performance on KITTI stereo benchmarks achieved by SOTA end-to-end disparity networks [67, 10, 50] is remarkable, the generalization ability of these models is quite poor. When the target dataset is small or even without ground-truth disparities for finetuning, which is quite common in real scenarios, a synthetic-data pretrained deep neural network cannot generalize well due to the domain gap and the over-fitting problem. Hence instead of achieving a tiny delta over existing methods, developing an effective domain adaptation method for deep stereo networks would be more meaningful, which is little explored in the field of stereo matching.
For stereo domain adaptation, Pang et al. [44] proposed a semi-supervised approach utilizing the scale information in disparity regression. Guo et al. [22] proposed a cross-domain method using knowledge distillation. The recent developed MAD-Net [55] aimed at adapting a small stereo model online and in real-time. However, the performance of these methods on target domains is poor even after domain adaptation. Recently, a domain generalization network DSMNet [68] was proposed, in which a domain normalization approach and a non-local graph-based filter was designed based on GANet [67]. Although the performance of DSMNet is good, the key components to improve generalization ability are not intuitive. In this paper, we focus on the domain adaptation problem for deep stereo models, where only synthetic training data and images in the target domain are required. Two simple but effective approaches are proposed to improve stereo domain adaptation in an intuitive way, which are applied on both a correlation-based baseline model and a cost-volume based baseline model for validation. As can be seen, our proposed domain-adaptive stereo models outperform other domain-invariant methods across multiple stereo matching datasets including KITTI [16,40], Middlebury [46], ETH3D [48] and DrivingStereo [61].

3 Method

In this section, we describe our AdaStereo method. At first, we introduce the motivation behind the proposed domain-adaptive stereo matching method. Then, we detail the formulations of color transfer and cost regularization.

Fig. 2: Comparisons of input images and internal representations from different datasets. Top-down: SceneFlow [39] and Middlebury [46] examples. Left-right: input image, lower-layer features, cost volumes and high-level features.
3.1 Motivation

Compared with other computer vision tasks, stereo matching has its own properties. For the model structure, the network always adopts a siamese feature extractor for stereo inputs, and the cost volume is computed by correlation or concatenation. On this basis, in order to identify the inherent factors which impair the adaptive performance of deep stereo network, our exploration is not only limited to disparity results, but also internal representations under distinct domains.

Fig. 3 depicts a typical stereo matching model. The model is composed of three parts: shallow feature extractor, matching feature aggregator and disparity encoder-decoder, whose outputs correspond to the lower-layer features, cost volumes and high-level representations, respectively. In the forward pass, we extract these three kinds of representations and analyze their differences under distinct domains. As shown in Fig. 2, two images with sort of similar contents from SceneFlow [39] and Middlebury [46] are selected and fed to the network. Firstly, these two images have observable variances on the color and brightness. The visualization of lower-layer features reveals the differences as well, where the values of SceneFlow features (highlighted areas) are much greater than Middlebury features. Next, for cost volumes computed by the 1D-correlation layer [39], we calculate the proportion of cost values in each interval and find the distributions between two domains are inconsistent as well. Finally, no much differences are exposed in the high-level representations since few activations are recognized in the map. Except for above observable comparisons, we also calculate the mean values of RGB channels in two datasets, and find significant color variances between the synthetic and realistic domains: (107, 102, 92) for SceneFlow [39] and (148, 132, 102) for Middlebury [46]. Hence we conclude that the inherent differences between domains for stereo matching mainly lie in the input color and cost volumes resulting from the inconsistent lower-layer features.

Correspondingly, in order to solve the adaptation problem for stereo matching task, we propose the color transfer and cost regularization to handle the domain gaps in the level of input color space and cost volume respectively.

3.2 Color Transfer

As mentioned above, the color difference plays a crucial role in the low-level inconsistency. We thus present an efficient algorithm for color transfer that is different from those complex and unstable style transfer methods. Given the source image $I_s$ and the referenced image $I_r$, our algorithm outputs the transferred image $I_t$ which owns the content of $I_s$ and color style of $I_r$. As detailed in Alg. 1, the transfer is performed in the LAB color space. The $T_{RGB\rightarrow LAB}(\cdot)$ and $T_{LAB\rightarrow RGB}(\cdot)$ denote the color space transformations between RGB and LAB. Under the LAB space, the mean value $\mu$ and standard deviation $\sigma$ in each channel are first computed by $S(\cdot)$. Then each channel in the referenced image $I_r$ is subtracted by its mean value and multiplied by the factor $\lambda$ which is scaled
Fig. 3: The training diagram of our AdaStereo. For the color transfer, the left and right images from the SceneFlow dataset [39] are transferred to the Middlebury [46] style, which is only adopted during training without extra costs. For the cost regularization, the lower-layer features are normalized before correlation or concatenation.

by standard deviations. Finally the transferred image $I_t$ is acquired through the converse addition of mean values and color space conversion.

The algorithm enables fast and effective color transfer across different datasets without any learning parameters in training. Currently, stereo matching models need to be pretrained on the large amount of synthetic images [39] and fine-tuned on the small amount of realistic images [46,40]. With the assistance of our transfer algorithm, the massive virtual images can be directly transferred to the target scenes, which makes the pre-trained model directly adapt to new domains. As shown in Fig. 3, our algorithm can be seamlessly embedded into the input module of stereo matching framework during training. In order to improve the diversity of training data, the referenced image $I_r$ is randomly selected from target scenarios in each training iteration. Experiments in Sec. 4.3 will provide more qualitative examples to further validate our algorithm.

Algorithm 1 Color transfer algorithm

**Input:** $I_s, I_r$

**Output:** $I_t$

\[
\begin{align*}
\tilde{I}_s & \leftarrow T_{\text{RGB}\rightarrow \text{LAB}}(I_s), \quad \tilde{I}_r \leftarrow T_{\text{RGB}\rightarrow \text{LAB}}(I_r) \\
(\mu_s, \sigma_s) & \leftarrow S(\tilde{I}_s), \quad (\mu_r, \sigma_r) \leftarrow S(\tilde{I}_r) \\
\tilde{I}_r & \leftarrow \tilde{I}_r - \mu_r \\
\lambda & \leftarrow \sigma_s / \sigma_r \\
\tilde{I}_s & \leftarrow \lambda \cdot \tilde{I}_r \\
\tilde{I}_t & \leftarrow \tilde{I}_s + \mu_s \\
I_t & \leftarrow T_{\text{LAB}\rightarrow \text{RGB}}(\tilde{I}_t)
\end{align*}
\]
3.3 Cost Regularization

After the color transfer, we propose the cost regularization to constrain the distribution of cost volume by pixel normalization and channel normalization, aiming at further narrowing the deviations in cost volumes across different domains. As mentioned in Sec. 2, current models compute cost volumes by two patterns: correlation and concatenation. In order to satisfy both patterns, our regularization is conducted only once on the lower-layer features just before building the cost volume.

Before correlation or concatenation, the left feature $\mathcal{F}^L$ and right feature $\mathcal{F}^R$ of size $N \times C \times H \times W$ ($N$: batch size, $C$: channel, $H$: spatial height, $W$: spatial width), are successively regularized by the channel normalization and pixel normalization. Specifically, the channel normalization is applied across the spatial dimensions ($H, W$) for each channel per sample separately, which is defined as:

$$F_{n,c,h,w} = \frac{F_{n,c,h,w}}{\sqrt{\sum_{h=0}^{H-1} \sum_{w=0}^{W-1} |F_{n,c,h,w}|^2 + \varepsilon}}$$

where $F$ denotes the extracted lower-layer feature, $h$ and $w$ denote the spatial position, $c$ denotes the channel, and $n$ denotes the batch index. After the channel normalization, the pixel normalization is further applied to each spatial position per sample separately, which is defined as:

$$F_{n,c,h,w} = \frac{F_{n,c,h,w}}{\sqrt{\sum_{c=0}^{C-1} |F_{n,c,h,w}|^2 + \varepsilon}}$$

By the pixel normalization and channel normalization, the cost volume generated subsequently is regularized to a fixed range, which can help reduce the gaps of matching costs due to varied contents across domains.

Compared with previous normalization methods [45,59], our proposed regularization method does not contain any learning parameters, which is imposed exactly on the lower-layer features for building cost volume. The cost regularization is able to reduce the variations in the lower-layer features and cost volumes under different inputs, making the model more robust to domain shifts. In addition, this operation can be seamlessly integrated into 2-D or 3-D disparity networks. In the following experiments, we validate the effectiveness of cost regularization for popular models across multiple datasets.

4 Experiment

To prove the effectiveness of our method, we extend the 2-D stereo baseline network ResNetCorr [60,62,50] as Ada-ResNetCorr and the 3-D stereo baseline network PSMNet [3] as Ada-PSMNet. Based on our domain-adaptive stereo models, we first conduct detailed ablation studies across multiple datasets including KITTI [16,40], Middlebury [46], ETH3D [48] and DrivingStereo [61]. Next
we compare the cross-domain performance of our method with other domain-invariant methods and state-of-the-art end-to-end disparity networks. Finally, we show that our domain-adaptive stereo models achieve remarkable performance on three public stereo matching benchmarks, even outperforming several state-of-the-art disparity networks fine-tuned on real datasets.

4.1 Datasets

Five public stereo matching datasets are used, one synthetic dataset for training and other four real datasets for cross-domain evaluation.

The SceneFlow dataset [39] is a large synthetic dataset containing 35k training pairs with dense ground-truth disparities, including three subsets that are very different with real scenes in color and contents. We use all synthetic stereo pairs with ground-truth disparities for training.

The KITTI dataset includes two subsets, i.e., KITTI 2012 [16] and KITTI 2015 [40], providing 394 stereo pairs of outdoor driving scenes with sparse ground-truth disparities for training, and 395 image pairs for testing. The Middlebury dataset [46] is a small indoor dataset containing no more than 50 stereo pairs with three different resolutions. The ETH3D dataset [48] includes both indoor and outdoor scenarios, containing 27 gray image pairs with dense ground-truth disparities for training, and 20 image pairs for testing. The DrivingStereo dataset [61] is a large-scale stereo dataset covering a diverse set of driving scenarios, containing over 170k stereo pairs for training and 7751 pairs for testing with sparse ground-truth disparities.

To evaluate the results, we use the bad pixel error ($D_1$-error), which calculates the percentage of disparity errors above a threshold. The KITTI, Middlebury and ETH3D training sets and the DrivingStereo test set are used for cross-domain evaluations.

4.2 Model Specifications and Implementation Details

We first introduce the network structures of Ada-ResNetCorr and Ada-PSMNet. Ada-ResNetCorr is extended on ResNetCorr [60,62,50], which is a correlation-based 2-D baseline network. We extend the ResNetCorr, where the layers from “conv1_1” to “conv1_3” in ResNet-50 [25] are adopted as the shallow feature extractor, and we also replace the output convolutional layer with the soft-argmin block in 3-D stereo networks [29,3] for disparity regression. The proposed cost regularization is simply applied on the extracted lower-layer features before correlation, which are of 1/2 spatial size to the input image. The maximum displacement in the correlation layer is set to 128. Ada-PSMNet is extended on PSMNet [3], which is an effective cost-volume based 3-D baseline network. We follow the structure of PSMNet except that the maximum disparity is set to 256. The cost regularization is simply applied on the extracted features of two views before constructing the 4-D cost volume. As can be seen, the proposed cost regularization is concise and efficient, which can be seamlessly integrated in
Table 1: Ablation study. An individual model is trained for each target domain using synthetic data. $D_1$-error (%) is adopted for evaluation

| Model         | cost | regularization | color transfer | KITTI 2012 | KITTI 2015 | Middlebury half | Middlebury quarter | ETH | DrivingStereo |
|---------------|------|----------------|----------------|------------|------------|----------------|-------------------|-----|---------------|
| Ada-PSMNet    | ✓    | ✓              | ✓              | 13.6       | 12.1       | 18.6           | 11.5              | 10.8 | 20.9          |
| Ada-ResNetCorr| ✓    | ✓              | ✓              | 9.8        | 9.4        | 22.5           | 12.8              | 15.8 | 17.2          |

existing stereo networks without any new learning parameters. In addition, the batch normalization layer is kept in our domain-adaptive stereo models.

When conducting domain adaptation from synthetic to real, we treat each real dataset as a target domain individually. Correspondingly, a domain-adaptive stereo model is trained for each target domain, using all synthetic stereo pairs in the SceneFlow dataset, as well as image pairs without ground-truth disparities in the real dataset for color transfer. During training, a pair is randomly selected from the target dataset as color reference for each input synthetic image pair. The training is conducted on eight NVIDIA Titan-Xp GPUs. During training, the smooth $L_1$ loss is adopted. All models are optimized with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and the “poly” learning rate policy. For data augmentation, we use the color shift, saturation and contrast adjustments with factors between 0 and 0.4, as well as the style-PCA based lighting noises. For Ada-PSMNet, we train the model for 100 epochs with a batch size of 16 using $624 \times 304$ random crops from input images, and the initial learning rate is set to 0.002. For Ada-ResNetCorr, we train the model for 300 epochs with a batch size of 16 using $513 \times 321$ random crops, and the initial learning rate is set to 0.01.

4.3 Ablation Study

We conduct detailed ablation studies across four datasets to evaluate the key components in our domain-adaptive stereo matching method. As listed in Table 1, we validate using both Ada-PSMNet and Ada-ResNetCorr. As can be seen, applying color transfer during training can significantly improve accuracies on multiple target domains, e.g. 8.0% on KITTI, 8.2% on Middlebury, 4.2% on ETH3D and 13.0% on DrivingStereo for Ada-PSMNet, benefiting from massive color-aligned training data without any extra costs. In Fig. 4, we provide qualitative results of color transfer from the synthetic SceneFlow dataset to three real datasets. As can be seen, the generated images are quite vivid hence the gaps to target domains are significantly reduced in terms of the color space.
Furthermore, compared with baseline models, the accuracies are improved by $2\% \sim 5\%$ on real datasets by integrating cost regularization, which also works well when implemented along with color transfer. Finally, the full-setting Ada-DSMNet and Ada-ResNetCorr outperform the baseline models by a notable margin on all target domains, especially an accuracy improvement of 14.1\% for Ada-PSMNet on the large-scale DrivingStereo dataset, which fully demonstrates the effectiveness of our domain-adaptive stereo matching method.

In DSMNet [68], the authors also applied their proposed domain normalization approach and non-local filter layers on the PSMNet for verification, and an error rate of 8.5\% is achieved on the KITTI 2015 training set. By comparison, our Ada-PSMNet achieves an error rate of 5.0\% based on the same baseline network, while using more intuitive and effective approaches for domain adaptation.

**Comparison with Style Transfer Method.** To verify the role of color space in domain gap and the effectiveness of our color transfer approach, we also adopt the popular style transfer method WCT$^2$ [64], to perform the offline transfer as an alternative. We test the Ada-ResNetCorr with WCT$^2$ on KITTI 2015 and Middlebury training sets, and the error rates are 9.8\% and 12.6\% respectively. By comparison, the error rates of Ada-ResNetCorr with our color transfer method are 7.1\% and 8.6\% respectively, achieving notable gains over WCT$^2$. Two advantages of color transfer lead to the improvement: (1) The simple and stable color transfer focuses on the color space transferring, which is the crucial part of domain gap while WCT$^2$ generates unstable target-style images often with blurry boundaries and missing details; (2) Our method is an online transfer method, which delivers sufficient diversity in training, while WCT$^2$ performs transfer in a fixed one-to-one manner.
4.4 Cross-domain Comparisons with State-of-the-Art Methods

In Table 2, we compare the proposed domain-adaptive stereo models: Ada-ResNetCorr and Ada-PSMNet, with other state-of-the-art end-to-end disparity networks and domain-invariant stereo methods on three real datasets. All deep neural networks are trained on the SceneFlow dataset. As can be seen, both Ada-ResNetCorr and Ada-PSMNet far outperform other SceneFlow-pretrained stereo networks and traditional stereo matching algorithms on all target domains. When comparing with the state-of-the-art domain-invariant stereo network DSMNet [68], our Ada-ResNetCorr achieves on-par cross-domain performance while our Ada-PSMNet outperforms the DSMNet on all real datasets, especially 4.7% lower in $D_1$-error on the Middlebury dataset. It is worth mentioning that our domain-adaptive models are not designed based on the most advanced disparity networks. Contrarily, we aim to provide more insights for domain adaptation and generalization in stereo matching using our simple but effective approaches, rather than designing non-intuitive black-box structures and tricks. In Fig. 5, we provide qualitative results of our method on real datasets.

4.5 Evaluations on Stereo Benchmarks

To further demonstrate the effectiveness of our domain adaptation method, we compare the proposed domain-adaptive stereo model Ada-PSMNet with several unsupervised/self-supervised methods and fine-tuned end-to-end disparity networks on public stereo matching benchmarks: KITTI, Middlebury and ETH3D. Our AdaStereo model is not fine-tuned using ground-truth disparities in target domains before submitting test results.
AdaStereo: A Simple and Efficient Approach for Adaptive Stereo Matching

Fig. 5: Disparity estimates of our synthetic-data pretrained AdaStereo model on the KITTI, Middlebury and ETH3D datasets. Left-right: left stereo image, colorized disparity map and error map.

**Results on the ETH3D Benchmark.** For the evaluation on the ETH3D stereo benchmark, 20 gray stereo pairs are provided with hidden ground-truth disparities. The results of the SceneFlow pre-trained Ada-PSMNet is directly uploaded to the online benchmark. As can be seen in Table 3, our synthetic-data pre-trained model outperforms the state-of-the-art patch-based model DeepPruner [12] and end-to-end networks (including iResNet [33], PSMNet [3] and StereoDRNet [2]) which are fine-tuned on the ETH domain, as well as the SOTA traditional stereo algorithm SGM-Forest [47]. It is interesting to find out that our Ada-PSMNet outperforms the fine-tuned PSMNet by simply using the proposed color transfer and cost regularization without any new learning parameters. By the time of the paper submission, AdaStereo ranks 1st on the ETH3D benchmark in 2-pixel error metric.

**Results on the Middlebury Benchmark.** For the evaluation on the Middlebury stereo benchmark, 15 high-resolution image pairs are provided with hidden
Table 3: Comparison with state-of-the-art methods on the ETH3D stereo benchmark. The 1-pixel error (%) and 2-pixel error (%) are adopted for evaluation.

| Models         | Deep-Pruner[12] | iResNet[33] | SGM-Forest[47] | PSMNet[3] | Stereo-DRNet[2] | DispNet[39] | AdaStereo |
|----------------|-----------------|-------------|----------------|-----------|-----------------|-------------|-----------|
| Training       | ETH-gt          | ETH-gt      | no gt          | ETH-gt    | ETH-gt          | ETH-gt      | Synthetic |
| Bad 1.0        | 3.52            | 3.68        | 4.96           | 5.02      | 5.59            | 17.47       | 3.41      |
| Bad 2.0        | 0.86            | 1.00        | 1.84           | 1.09      | 1.48            | 7.91        | 0.74      |

Table 4: Comparison with state-of-the-art end-to-end disparity networks on the Middlebury stereo benchmark. The 2-pixel error (%) is adopted for evaluation.

| Models         | EdgeStereo[50]  | CasStereo[21] | iResNet[33] | MCV-MFC[35] | Deep-Pruner[12] | PSMNet[3] | AdaStereo |
|----------------|-----------------|---------------|-------------|-------------|-----------------|-----------|-----------|
| Training       | Mid-gt          | Mid-gt        | Mid-gt      | Mid-gt      | Mid-gt          | Mid-gt    | Synthetic |
| Bad 2.0        | 18.7            | 18.8          | 22.9        | 24.8        | 30.1            | 42.1      | 16.0      |

ground-truth disparities. We also directly upload the results of our SceneFlow pre-trained model to the online benchmark. As can be seen in Table 4, our synthetic-data pre-trained model outperforms all other state-of-the-art end-to-end disparity networks (e.g. EdgeStereo[50], iResNet[33]) which are fine-tuned using Middlebury training data by a noteworthy margin. Our Ada-PSMNet achieves a remarkable 2-pixel error rate of 16.0% on the full-resolution test set, which is the best performance achieved by an end-to-end stereo network on the Middlebury benchmark.

Results on the KITTI Benchmark. In order to further improve the performance of our domain-adaptive model on KITTI, we fuse the training data in the SceneFlow and Cityscapes [11] datasets for collaborative learning. Cityscapes is an urban scene understanding dataset, providing about 20k rectified stereo pairs, and their disparity maps are pre-computed by SGM[26]. The pre-computed disparity maps are regarded as labels in the pre-training in which the color transfer is still used, and other training settings remain unchanged. As can be seen in Table 5, our domain-adaptive model far outperforms the weak-supervised/unsupervised methods, and achieves close or higher accuracy than some supervised stereo models (e.g. MC-CNN[65], DispNetC[39] and Content-CNN[38]).

Table 5: Comparison with state-of-the-art methods on the KITTI 2015 stereo benchmark. The D1-error (%) is adopted for evaluation.

| Models          | MC-CNN-acr[65] | AdaStereo | DispNetC [39] | Content-CNN[38] | Weak-Sup [57] | MADNet [55] | Unsupervised [71] |
|-----------------|----------------|-----------|---------------|-----------------|---------------|-------------|-------------------|
| Training        | KITTI-gt       | Synthetic | KITTI-gt      | KITTI-gt        | no gt         | no gt       |                   |
| D1-error        | 3.89           | 3.93      | 4.34          | 4.54            | 4.97          | 8.23        | 9.91              |
5 Conclusions

In this paper, we focus on the domain adaptation problem for deep stereo matching networks. Two simple but effective approaches that are tightly connected with stereo matching task are proposed, i.e. color transfer and cost regularization, which can significantly reduce the gaps in color spaces, lower-layer features and cost volumes across different domains. The proposed modules can be easily integrated into both correlation based and cost-volume based stereo networks without any new parameters to learn and extra computational burdens. We verify our synthetic-data pretrained domain-adaptive stereo models on four real datasets, and the results show that our AdaStereo method achieves state-of-the-art cross-domain performance across multiple target domains. Our method also achieves remarkable accuracy on three public benchmarks, outperforming several state-of-the-art end-to-end disparity networks fine-tuned on target domains.

References

1. Bleyer, M., Rhemann, C., Rother, C.: Patchmatch stereo-stereo matching with slanted support windows. In: BMVC (2011)
2. Chabra, R., Straub, J., Sweeney, C., Newcombe, R., Fuchs, H.: Stereodrnet: Dilated residual stereonet. In: CVPR (2019)
3. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: CVPR (2018)
4. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: NIPS (2015)
5. Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals using stereo imagery for accurate object class detection (2017)
6. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Wang, Y.C.F., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: ICCV. pp. 2011–2020. IEEE (2017)
7. Chen, Y., Li, W., Gool, L.V.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. CVPR pp. 7892–7901 (2017)
8. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR. pp. 3339–3348 (2018)
9. Chen, Z., Sun, X., Wang, L., Yu, Y., Huang, C.: A deep visual correspondence embedding model for stereo matching costs. In: ICCV (2015)
10. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. IEEE transactions on pattern analysis and machine intelligence (2019)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
12. Duggal, S., Wang, S., Ma, W.C., Hu, R., Urtasun, R.: Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In: ICCV (2019)
13. Engel, J., Stückler, J., Cremers, D.: Large-scale direct slam with stereo cameras. In: IROS (2015)
14. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: ICCV. pp. 2960–2967 (2013)
15. Franke, U., Joos, A.: Real-time stereo vision for urban traffic scene understanding. In: IV (2000) 2
16. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012) 2, 3, 5, 8, 9
17. Geng, B., Tao, D., Xu, C.: Daml: Domain adaptation metric learning. IEEE Transactions on Image Processing 20(10), 2980–2989 (2011) 4
18. Gomez-Ojeda, R., Moreno, F.A., Zúñiga-Noèl, D., Scaramuzza, D., Gonzalez-Jimenez, J.: Pl-slam: A stereo slam system through the combination of points and line segments. IEEE Transactions on Robotics (2019) 2
19. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR. pp. 2066–2073. IEEE (2012) 4
20. Gong, R., Li, W., Chen, Y., Gool, L.V.: Dlow: Domain flow for adaptation and generalization. CVPR pp. 2472–2481 (2018) 4
21. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. arXiv preprint arXiv:1912.06378 (2019) 14
22. Guo, X., Li, H., Yi, S., Ren, J., Wang, X.: Learning monocular depth by distilling cross-domain stereo networks. In: ECCV. pp. 484–500 (2018) 4, 5
23. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: CVPR (2019) 2, 4, 12
24. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003) 1
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 9
26. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence 30(2), 328–341 (2007) 12, 14
27. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016) 4
28. Hosni, A., Rheinmann, C., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012) 12
29. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: ICCV (2017) 2, 4, 9
30. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: CVPR. pp. 1785–1792 (2011) 4
31. Li, P., Chen, X., Shen, S.: Stereo r-cnn based 3d object detection for autonomous driving. In: CVPR (2019) 2
32. Li, P., Qin, T., Shen, S.: Stereo vision-based semantic 3d object and ego-motion tracking for autonomous driving. In: ECCV (2018) 2
33. Liang, Z., Feng, Y., Guo, Y., Liu, H.: Learning deep correspondence through prior and posterior feature constancy. In: CVPR (2018) 4, 13, 14
34. Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Zhou, L., Zhang, J.: Learning for disparity estimation through feature constancy. In: CVPR (2018) 2
35. Liang, Z., Guo, Y., Feng, Y., Chen, W., Qiao, L., Zhou, L., Zhang, J., Liu, H.: Stereo matching using multi-level cost volume and multi-scale feature constancy. IEEE transactions on pattern analysis and machine intelligence (2019) 14
36. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 2
37. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105 (2015) 4
38. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: CVPR (2016) 3, 14
39. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR (2016) 2, 4, 5, 6, 7, 9, 11, 14
40. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015) 2, 3, 5, 7, 8, 9
41. Miclea, V.C., Nedevschi, S.: Real-time semantic segmentation-based stereo reconstruction. IEEE Transactions on Intelligent Transportation Systems (2019) 2
42. Nie, G.Y., Cheng, M.M., Liu, Y., Liang, Z., Fan, D.P., Liu, Y., Wang, Y.: Multi-level context ultra-aggregation for stereo matching. In: CVPR. pp. 3283–3291 (2019) 4
43. Pang, J., Sun, W., Ren, J., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: ICCV Workshop (2017) 2, 4
44. Pang, J., Sun, W., Yang, C., Ren, J., Xiao, R., Zeng, J., Lin, L.: Zoom and learn: Generalizing deep stereo matching to novel domains. In: CVPR. pp. 2070–2079 (2018) 5
45. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 3, 8
46. Scharstein, D., Hirschmuller, H., Kitajima, Y., Krathwohl, G., Nei, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: GCPR (2014) 2, 3, 5, 6, 7, 8, 9
47. Schönberger, J.L., Sinha, S.N., Pollefeys, M.: Learning to fuse proposals from multiple scanline optimizations in semi-global matching. In: ECCV (2018) 13, 14
48. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017) 3, 5, 8, 9
49. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning. In: CVPR (2017) 3
50. Song, X., Zhao, X., Fang, L., Hu, H., Yu, Y.: EdgStereo: An effective multi-task learning network for stereo matching and edge detection. IJCV 4, 8, 9, 14
51. Song, X., Zhao, X., Fang, L., Lin, T.: Discriminative representation combinations for accurate face spoofing detection. Pattern Recognition 85, 220–231 (2019) 2
52. Song, X., Zhao, X., Hu, H., Fang, L.: EdgStereo: A context integrated residual pyramid network for stereo matching. In: ACCV (2018) 2, 4
53. Sun, R., Zhu, X., Wu, C., Huang, C., Shi, J., Ma, L.: Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In: CVPR. pp. 4360–4369 (2019) 4
54. Tonioni, A., Poggi, M., Mattoccia, S., di Stefano, L.: Unsupervised domain adaptation for depth prediction from images. IEEE transactions on pattern analysis and machine intelligence (2019) 4
55. Tonioni, A., Tosì, F., Poggi, M., Mattoccia, S., Stefano, L.D.: Real-time self-adaptive deep stereo. In: CVPR. pp. 195–204 (2019) 5, 14
56. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018) 4
57. Tulyakov, S., Ivanov, A., Fleuret, F.: Weakly supervised learning of deep metrics for stereo reconstruction. In: ICCV (2017) 14
58. Wang, R., Schworer, M., Cremers, D.: Stereo dso: Large-scale direct sparse visual odometry with stereo cameras. In: ICCV (2017)
59. Wu, Y., He, K.: Group normalization. In: ECCV (2018)
60. Yang, G., Deng, Z., Lu, H., Li, Z.: Src-disp: Synthetic-realistic collaborative disparity learning for stereo matching. In: ACCV (2018)
61. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 899–908 (2019)
62. Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J.: Segstereo: Exploiting semantic information for disparity estimation. In: ECCV (2018)
63. Yin, Z., Darrell, T., Yu, F.: Hierarchical discrete distribution decomposition for match density estimation. In: CVPR. pp. 6044–6053 (2019)
64. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: ICCV (2019)
65. Zbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: CVPR (2015)
66. Zhang, C., Wang, L., Yang, R.: Semantic segmentation of urban scenes using dense depth maps. In: ECCV (2010)
67. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: CVPR (2019)
68. Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B., Torr, P.: Domain-invariant stereo matching networks. arXiv preprint arXiv:1911.13287 (2019)
69. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV. vol. 2, p. 6 (2017)
70. Zhao, S., Fu, H., Gong, M., Tao, D.: Geometry-aware symmetric domain adaptation for monocular depth estimation. CVPR pp. 9780–9790 (2019)
71. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)
72. Zhu, J.: Image gradient-based joint direct visual odometry for stereo camera. In: IJCAI (2017)
73. Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: CVPR. pp. 687–696 (2019)
74. Zhu, X., Zhou, H., Yang, C., Shi, J., Lin, D.: Penalizing top performers: Conservative loss for semantic segmentation adaptation. In: ECCV. pp. 568–583 (2018)