AdaStereo: A Simple and Efficient Approach for Adaptive Stereo Matching

Xiao Song[1⋆] Guorun Yang[1,3⋆] Xinge Zhu[2] Hui Zhou[1] Zhe Wang[1,4] Jianping Shi[1,5]
[1]SenseTime Research [2]The Chinese University of Hong Kong
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[4]Shanghai AI Laboratory [5]Qing Yuan Research Institute, Shanghai Jiao Tong University
{songxiao,yangguorun,zhouhui,wangzhe,shijianping}@sensetime.com zx018@ie.cuhk.edu.hk

Abstract

Recently, records on stereo matching benchmarks are constantly broken by end-to-end disparity networks. However, the domain adaptation ability of these deep models is quite poor. Addressing such problem, we present a novel domain-adaptive pipeline called AdaStereo that aims to align multi-level representations for deep stereo matching networks. Compared to previous methods for adaptive stereo matching, our AdaStereo realizes a more standard, complete and effective domain adaptation pipeline. Firstly, we propose a non-adversarial progressive color transfer algorithm for input image-level alignment. Secondly, we design an efficient parameter-free cost normalization layer for internal feature-level alignment. Lastly, a highly related auxiliary task, self-supervised occlusion-aware reconstruction is presented to narrow down the gaps in output space. Our AdaStereo models achieve state-of-the-art cross-domain performance on multiple stereo benchmarks, including KITTI, Middlebury, ETH3D, and DrivingStereo, even outperforming disparity networks finetuned with target-domain ground-truths.

1. Introduction

The stereo matching task aims to find all corresponding pixels in a stereo pair, and the distance between corresponding pixels is known as disparity [12]. Based on the epipolar geometry, stereo matching enables stable depth perception, hence it supports further applications such as scene understanding, object detection, odometry, and SLAM.

Recent stereo matching methods typically adopt fully convolutional networks [21] to regress disparity maps directly and have achieved state-of-the-art performance on stereo benchmarks [7, 25, 30]. However, the performance of these methods adapted from synthetic data to real-world scenes is limited. As shown in Fig. 1, the PSMNet [4] pretrained on the SceneFlow dataset [24] fails to produce good results on the Middlebury [30] and KITTI [25] datasets. Therefore, instead of designing complicated networks for higher accuracy on specific datasets, how to obtain effective domain-adaptive stereo matching models is more desirable now.

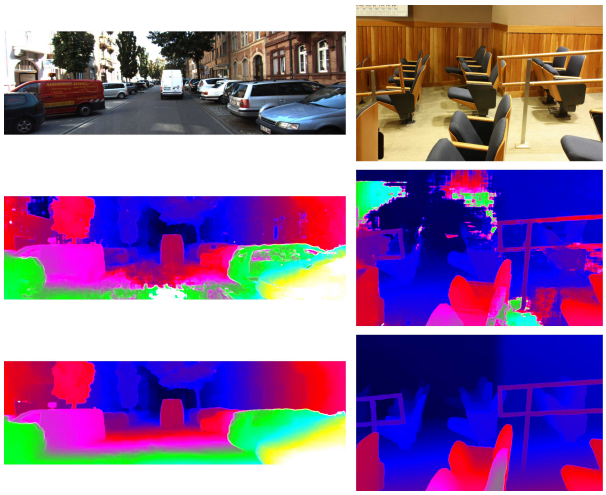

Figure 1. Overview examples. Left-right: KITTI [25] and Middlebury [30]. Top-down: left image, disparity maps predicted by the SceneFlow-pretrained PSMNet [4], and by our Ada-PSMNet.

In this work, we aim at the important but less explored problem of domain adaptation in stereo matching. Considering the fact that there are a great number of synthetic data but only a small number of realistic data with ground-truths, we focus on domain gaps between synthetic and realistic domains. We first analyse main differences between these two domains, as shown in Fig. 2. At the input image level, color and brightness are the obvious gaps. By making statistics on the internal cost volumes, we also find significant differences in distributions. Moreover, geometries of the output disparity maps are inconsistent as well. In order to bridge the domain gaps at these levels (input image, internal cost volume, and output disparity), we propose ***AdaStereo***, a standard and complete domain adaptation pipeline for stereo

⋆ indicates equal contribution.

matching, in which three particular modules are presented:

- For input image-level alignment, the **non-adversarial progressive color transfer** algorithm is presented to align input color space between source and target domains during training. It is the first attempt that adopts a non-adversarial style transfer method to align input-level inconsistency for stereo domain adaptation, avoiding harmful side-effects of geometrical distortions from GAN-based methods [56]. Furthermore, the proposed progressive update strategy enables capturing representative target-domain color styles during adaptation.

- For internal feature-level alignment, the **cost normalization** layer is proposed to align matching cost distribution. Oriented to the stereo matching task, two normalization operations are designed and embedded in this layer: (i) channel normalization reduces the inconsistency in scaling of each feature channel; and (ii) pixel normalization further regulates the norm distribution of pixel-wise feature vector for binocular matching. Compared to previous general normalization layers (*e.g.* IN [42], DN [54]), our cost normalization layer is parameter-free and adopted only once in the network.

- For output-space alignment, we conduct self-supervised learning through a highly related auxiliary task, **self-supervised occlusion-aware reconstruction**, which is the first proposed auxiliary task for stereo domain adaptation. Concretely, a self-supervised module is attached upon the main disparity network, to perform image reconstructions on the target domain. To address the ill-posed occlusion problem in reconstruction, we also design a domain-collaborative learning strategy for occlusion mask predictions. Through occlusion-aware stereo reconstruction, more informative geometries from target scenes are involved in model training, thus benefiting disparity predictions across domains.

Based on our proposed pipeline, we conduct effective domain adaptation from synthetic data to real-world scenes. In Fig. 1, our Ada-PSMNet pretrained on the synthetic dataset performs well on both indoor and outdoor scenes. In order to validate the effectiveness of each module, ablation studies are performed on diverse real-world datasets, including Middlebury [30], ETH3D [32], KITTI [7, 25], and DrivingStereo [47]. Our domain-adaptive models outperform other traditional / domain generalization / domain adaptation methods, and even finetuned models on multiple stereo matching benchmarks. Main contributions are summarized below:

- We locate the domain-adaptive problem and investigate domain gaps for deep stereo matching networks.

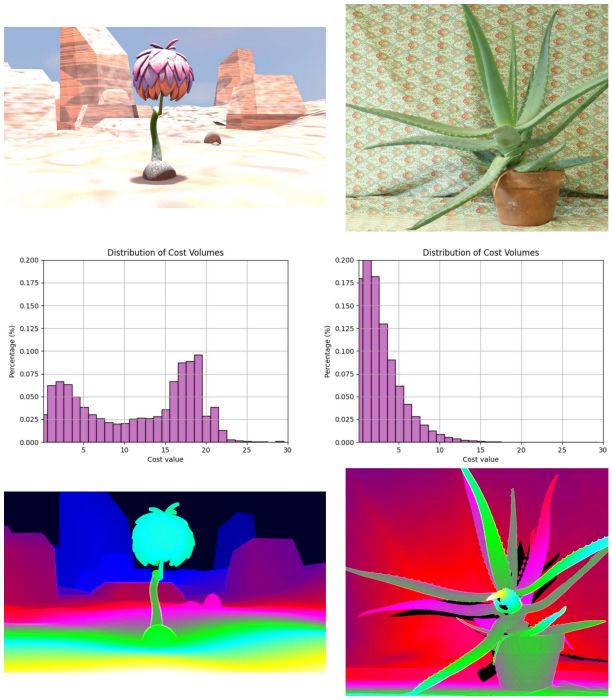

Figure 2. Comparisons of input images, internal cost volumes, and output disparity maps between synthetic and real-world datasets. Left-right: SceneFlow [24] and Middlebury [30]. Top-down: input image, internal cost volume, and output disparity map. Disparity maps are rendered by the same color map.

- We propose a novel domain adaptation pipeline, including three modules to narrow down the gaps at input image level, internal feature level, and output space.

- Our domain-adaptive models outperform other domain-invariant methods, and even finetuned disparity networks on multiple stereo matching benchmarks.

2. Related Work

2.1. Stereo Matching

Recently, end-to-end stereo matching networks have achieved state-of-the-art performance, which can be roughly categorized into two types: correlation-based 2-D stereo networks and cost-volume based 3-D stereo networks. On the one hand, Mayer *et al.* [24] proposed the first end-to-end disparity network DispNetC. Since then, based on color or feature correlations, more advanced models were proposed, including CRL [27], iResNet [18], HD^3 [49], SegStereo [48], and EdgeStereo [36, 35]. On the other hand, 3-D convolutional neural networks show the advantages in regularizing cost volume for disparity estimation, including GC-Net [15], PSMNet [4], GWCNet [11], GANet [53], *etc.* Our proposed domain adaptation pipeline for stereo matching can be easily applied on both 2-D and 3-D stereo networks.

2.2. Domain Adaptation

Prior works on domain adaptation can be roughly divided into two categories. The general idea of the first category is aligning source and target domains at different levels, including: (1) input image-level alignment [2, 14], using image-to-image translation methods such as CycleGAN [56]; (2) internal feature-level alignment, based on feature-level domain adversarial learning [41, 23]; (3) conventional discrepancy measures, such as MMD [22] and CMD [52]; and (4) output-space alignment [39, 43], based on adversarial learning. For the second category, self-supervised learning based domain adaptation methods [37] achieve great progress, in which simple auxiliary tasks generated automatically from unlabeled data are utilized to train feature representations, such as rotation prediction [8], patch-location prediction [45], *etc*. In this paper, we explicitly implement domain alignments at input level and internal feature level, while incorporating self-supervised learning into output-space alignment through a specifically designed auxiliary task.

2.3. Domain-Adaptive Stereo Matching

Although records on public benchmarks are constantly broken, few attention has been paid to the domain adaptation ability of deep stereo models. Pang *et al.* [28] proposed a semi-supervised method utilizing the scale information. Guo *et al.* [10] presented a cross-domain method using knowledge distillation. MAD-Net [38] was designed to adapt a compact stereo model online. Recently, StereoGAN [20] utilized CycleGAN [56] to bridge domain gaps by joint optimizations of image style transfer and stereo matching. However, no standard and complete domain adaptation pipeline was implemented in these methods, and their adaptation performance is quite limited. Contrarily, we propose a more complete pipeline for deep stereo models following the standard domain adaptation methodology, in which alignments across domains are conducted at multiple levels thereby remarkable adaptation performance is achieved. In addition, we do not conduct any adversarial learning, hence the training stability and semantic invariance are guaranteed.

3. Method

In this section, we first describe the problem of domain-adaptive stereo matching. Then we introduce the motivation and give an overview of our domain adaptation pipeline. After that, we detail the main components in the pipeline, *i.e*. non-adversarial progressive color transfer, cost normalization, and self-supervised occlusion-aware reconstruction.

3.1. Problem Description

In this paper, we focus on the domain adaptation problem for stereo matching. Different from domain generalization where a method needs to perform well on unseen scenes, domain adaptation allows using target-domain images without ground-truths during training. Specifically for stereo matching, since there are a great number of synthetic data [24] but only a small number of realistic data with ground-truths [25, 30, 32], the problem can be further limited to the adaptation from virtual to real-world scenarios. Given stereo image pairs (I_s^l, I_s^r) and (I_t^l, I_t^r) on source synthetic and target realistic domains, and the ground-truth disparity map $\hat{d}_s^l$ on the source synthetic domain, we train the model to predict the disparity map d_t^l on the target domain.

3.2. Motivation

As shown in Fig. 2, two images from the SceneFlow [24] and Middlebury [30] datasets are selected to describe inherent inconsistencies between two domains. (i) These two images own observable differences in their color and brightness. Moreover, according to the statistics on the *whole datasets*, the mean values of RGB channels are $(107, 102, 92)$ in SceneFlow and $(148, 132, 102)$ in Middlebury. Therefore, significant color variances are found between synthetic and realistic domains. (ii) For cost volumes computed from the 1D-correlation layer [24], we calculate the proportion of matching cost values in each interval and find the distributions between two domains are inconsistent as well. (iii) Although these two images have similar plants as foreground, the generated disparity maps still vary in scene geometries. Both the foreground objects and background screens have quite different disparities. Therefore, we conclude that the inherent differences across domains for stereo matching lie in the image color at input level, cost volume at feature level, and disparity map at output level.

Correspondingly, to solve the domain adaptation problem for stereo matching, we propose the progressive color transfer, cost normalization, and self-supervised reconstruction to handle the domain gaps in three levels respectively. The former two methods are presented to narrow down the differences in color space and matching cost distribution directly. The latter reconstruction is appended as an auxiliary task to impose extra supervision on disparity regression for target domain, finally benefiting the domain adaptation ability.

3.3. Framework Overview

Fig. 3 depicts the training pipeline of our stereo domain adaptation framework, in which three proposed modules are involved. For input, a randomly selected target-domain pair, and a randomly selected source-domain pair which adapts to target-domain color styles through our progressive color transfer algorithm, are simultaneously fed into a shared-weight disparity network with our cost normalization layer. The source-domain branch is under the supervision of the given ground-truth disparity map, while the target-domain branch is regulated by the proposed auxiliary task: self-supervised occlusion-aware reconstruction.

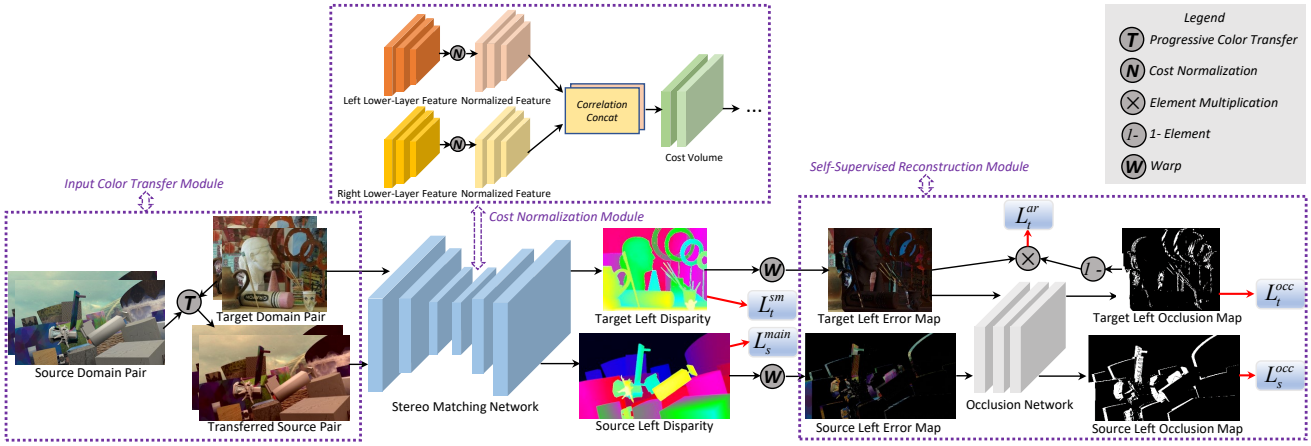

Figure 3. The training diagram of AdaStereo, with the adaptation from SceneFlow to Middlebury as an example. Color transfer and self-supervised occlusion-aware reconstruction modules are only adopted during training. (L_s^{main}, L_s^{occ}, L_t^{ar}, L_t^{occ}, L_t^{sm}) are the five training loss terms specified in Eq. 5.

3.4. Non-adversarial Progressive Color Transfer

As mentioned in Sec. 3.2, color difference plays a major role in the input image-level inconsistency across domains. Hence, we present an effective and stable algorithm for color transfer from source domain to target domain in a non-adversarial manner. During training, given a source-domain image I_s and a target-domain image I_t, the algorithm outputs a transferred source-domain image $I_{s\to t}$, which preserves the contents of I_s and owns the target-domain color styles. As in Alg. 1, color transfer is performed in the LAB color space. $T_{RGB\to LAB}(.)$ and $T_{LAB\to RGB}(.)$ denote the color space transformations. Under the LAB space, the mean value μ and standard deviation σ of each channel are first computed by $S(\cdot)$. Then, each channel in the source-domain LAB image $\tilde{I}_s$ is subtracted by its mean μ_s and multiplied by the standard deviation ratio λ. Finally, the transferred image $I_{s\to t}$ is obtained through the addition of the progressively updated μ_t and color space conversion. During training, two images in a source-domain pair are simultaneously transferred with the same μ_t and σ_t.

Compared with the Reinhard's color transfer method [29], the main contribution of our algorithm is the proposed progressive update strategy that proves to be more beneficial for domain adaptation. Considering color inconsistencies might exist across different images in the same target-domain dataset while the previous method [29] only allows one-to-one transformations, the source-domain images can not adapt to meaningful color styles that are representative for the whole target-domain dataset during adaptation. The progressive update strategy is proposed to address such problem. To be specific, target-domain μ_t and σ_t are progressively re-weighted by current inputs $({\mu_t}^i, {\sigma_t}^i)$ and historical records (μ_t, σ_t) with a momentum γ, simultaneously ensuring the diversity and representativeness of target-domain color styles during adaptation. Experimental results further validate its effectiveness over the previous color transfer method [29].

In a larger sense, we are the first to use a non-adversarial style transfer method to align input-level inconsistency for stereo domain adaptation. Unlike GAN-based style/color transfer networks [16, 56] that cause harmful side-effects of geometrical distortions for the low-level stereo matching task, our algorithm is more stable and training-efficient, which can be easily embedded in the training framework of stereo domain adaptation. Experimental results further validate its superiority over other adversarial transfer methods.

Algorithm 1 Progressive Color Transfer

Input: Source-domain dataset D_s, target-domain dataset D_t, $\mu_t = 0$, $\sigma_t = 0$

Randomly shuffle D_s and D_t
for $i \in [0, len(D_s))$ **do**
  Select $I_s \in D_s, I_t \in D_t$
  $\tilde{I}_s \Leftarrow T_{RGB\to LAB}(I_s)$, $\tilde{I}_t \Leftarrow T_{RGB\to LAB}(I_t)$
  $(\mu_s, \sigma_s) \Leftarrow S(\tilde{I}_s)$, $({\mu_t}^i, {\sigma_t}^i) \Leftarrow S(\tilde{I}_t)$
  $\mu_t \Leftarrow (1-\gamma) * \mu_t + \gamma * {\mu_t}^i$
  $\sigma_t \Leftarrow (1-\gamma) * \sigma_t + \gamma * {\sigma_t}^i$
  $\tilde{I}_s \Leftarrow \tilde{I}_s - \mu_s, \lambda \Leftarrow \sigma_t/\sigma_s$
  $\tilde{I}_{s\to t} \Leftarrow \lambda * \tilde{I}_s + \mu_t$
  $I_{s\to t} \Leftarrow T_{LAB\to RGB}(\tilde{I}_{s\to t})$
end for

3.5. Cost Normalization

Cost volume is the most important internal feature-level representation in a deep stereo network, encoding all necessary information for succeeding disparity regression. Hence for domain-adaptive stereo matching, an intuitive way is to directly narrow down the deviations in matching cost distributions across domains. Correspondingly, we design the

cost normalization layer, which is compatible with all cost volume building patterns (correlation and concatenation) in stereo matching, as shown in Fig. 3.

Before cost volume construction, the left lower-layer feature $\mathcal{F}^{\mathcal{L}}$ and right feature $\mathcal{F}^{\mathcal{R}}$ with the same size of $N \times C \times H \times W$ (N: batch size, C: channel, H: spatial height, W: spatial width), are both successively regularized by two proposed normalization operations: channel normalization and pixel normalization. Specifically, the channel normalization is applied across all spatial dimensions (H, W) per channel per sample individually, which is defined as:

$$\mathcal{F}_{n,c,h,w} = \frac{\mathcal{F}_{n,c,h,w}}{\sqrt{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}||\mathcal{F}_{n,c,h,w}||^2 + \varepsilon}}, \quad (1)$$

where $\mathcal{F}$ denotes the 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 across all channels per spatial position per sample individually, which is defined as:

$$\mathcal{F}_{n,c,h,w} = \frac{\mathcal{F}_{n,c,h,w}}{\sqrt{\sum_{c=0}^{C-1}||\mathcal{F}_{n,c,h,w}||^2 + \varepsilon}}. \quad (2)$$

Through channel normalization which reduces the inconsistency in norm and scaling of each feature channel, and pixel normalization which further regulates the norm distribution of pixel-wise feature vector for binocular matching, inter-domain gaps in matching cost distributions due to varied image contents and geometries are greatly reduced.

In a nutshell, our parameter-free cost normalization layer is indeed a normalization layer designed specifically for stereo domain adaptation, which is adopted only once before cost volume construction. On the contrary, previous normalization layers (*e.g.* BIN [26], IN [42], CN [5] (just IN), and DN [54]) are general normalization approaches, which contain learnable parameters and are repeatedly adopted in the network's feature extractor. Hence regulations on cost volume from these general normalization layers are not direct and effective enough, requiring extra implicit learning process. Moreover, our cost normalization layer does not use zero-centralization to prevent extra disturbances in matching cost distributions. Experimental results further validate its superiority over other general normalization layers.

3.6. Self-Supervised Occlusion-Aware Reconstruction

Self-supervised auxiliary tasks are demonstrated to be beneficial for domain adaptation on high-level tasks [8, 45]. However, such methodology has not been explored for the low-level stereo matching task. In this subsection, we propose an effective auxiliary task for stereo domain adaptation: self-supervised occlusion-aware reconstruction. As shown in Fig. 3, a self-supervised module is attached upon the main disparity network, to perform image reconstructions on the target domain. To address the ill-posed occlusion problem in reconstruction, we design a domain-collaborative learning strategy for occlusion mask predictions. Through occlusion-aware stereo reconstruction, more informative geometries from target scenes are involved in training.

During the self-supervised learning, stereo reconstruction is firstly measured by differences between the input target-domain left image I_t^l and the corresponding warped image $\overline{I_t^l}$ (based on the right image I_t^r and the produced disparity map d_t^l). Then, a small fully-convolutional occlusion prediction network takes the concatenation of d_t^l, I_t^r, and the pixel-wise error map $e_t^l = |I_t^l - \overline{I_t^l}|$ as input, and produces a pixel-wise occlusion mask O_t^l whose element denotes per-pixel occlusion probability from 0 to 1. Next, the reconstruction loss L_t^{ar} is re-weighted by the occlusion mask O_t^l and error map e_t^l on each pixel. Furthermore, we introduce the disparity smoothness loss (L_t^{sm}) to avoid possible artifacts. To guide the occlusion mask learning on the target domain more effectively, the shared-weight occlusion prediction network simultaneously learns an occlusion mask O_s^l on the source domain, under the supervision of the ground-truth occlusion mask $\hat{O}_s^l$ generated from the ground-truth disparity map $\hat{d}_s^l$. More details are provided in the supplementary material.

Our self-supervised occlusion-aware reconstruction task is the first proposed auxiliary task for stereo domain adaptation. In addition, our design enables collaborative occlusion mask learning on both source and target domains, acting as another domain adaptation on occlusion prediction that ensures the quality of the target-domain occlusion mask and explicitly improves the effectiveness of the target-domain reconstruction loss. Experimental results further validate its superiority over other high-level auxiliary tasks.

3.7. Training Loss

On the source domain, we train the main task of disparity regression using the per-pixel smooth-$L1$ loss: $L_s^{main} = Smooth_{L_1}(d_s^l - \hat{d}_s^l)$. In addition, the per-pixel binary cross entropy loss is adopted for occlusion mask training on the source domain: $L_s^{occ} = BCE(O_s^l, \hat{O}_s^l)$.

On the target domain, the occlusion-aware appearance reconstruction loss is defined as:

$$\begin{aligned} L_t^{ar} = \alpha \frac{1 - SSIM(I_t^l \odot (1 - O_t^l), \overline{I_t^l} \odot (1 - O_t^l))}{2} \\ + (1-\alpha)||I_t^l \odot (1 - O_t^l) - \overline{I_t^l} \odot (1 - O_t^l)||_1 \end{aligned} \quad (3)$$

where $\odot$ denotes element-wise multiplication, $SSIM$ denotes a simplified single scale SSIM [44] term with a 3×3 block fiter, and α is set to 0.85. Besides, we apply a $L1$-regularization term on the produced target-domain occlusion mask: $L_t^{occ} = ||O_t^l||_1$. Last but not least, we adopt an edge-aware term as the target-domain disparity smoothness loss,

where ∂I and ∂d denote image and disparity gradients:

$$L_t^{sm} = |\partial_x d_t^l| e^{-|\partial_x I_t^l|} + |\partial_y d_t^l| e^{-|\partial_y I_t^l|} \quad (4)$$

Finally, the total training loss is a weighted sum of five loss terms mentioned above, where λ_s^{occ}, λ_t^{ar}, λ_t^{occ}, and λ_t^{sm} denote corresponding loss weights:

$$L = L_s^{main} + \lambda_s^{occ} L_s^{occ} + \lambda_t^{ar} L_t^{ar} + \lambda_t^{occ} L_t^{occ} + \lambda_t^{sm} L_t^{sm} \quad (5)$$

4. Experiment

To prove the effectiveness of our domain adaptation pipeline, we extend the 2-D stereo baseline network ResNetCorr [46, 35] as **Ada-ResNetCorr**, and the 3-D stereo baseline network PSMNet [4] as **Ada-PSMNet**. We first conduct detailed ablation studies on multiple datasets including KITTI [7, 25], Middlebury [30], ETH3D [32], and DrivingStereo [47]. Next, we compare the cross-domain performance of our domain-adaptive models with other traditional / domain generalization / domain adaptation methods. Finally, we show that our domain-adaptive models achieve remarkable performance on public stereo matching benchmarks.

4.1. Datasets

The SceneFlow dataset [24] is a large synthetic dataset containing $35k$ training pairs with dense ground-truth disparity maps, acting as the source-domain dataset for training.

The KITTI dataset includes two subsets, *i.e.* KITTI 2012 [7] and KITTI 2015 [25], providing 394 stereo pairs of outdoor driving scenes with sparse ground-truth disparities for training, and 395 pairs for testing. The Middlebury dataset [30] is a small indoor dataset containing less than 50 stereo pairs with three different resolutions. The ETH3D dataset [32] includes both indoor and outdoor scenarios, containing 27 gray-image pairs with dense ground-truth disparities for training, and 20 pairs for testing. The DrivingStereo dataset [47] is a large-scale stereo matching dataset covering a diverse set of driving scenarios, containing over $170k$ stereo pairs for training and 7751 pairs for testing. These four real-world datasets act as different target domains that are adopted for cross-domain evaluations.

We adopt the bad pixel error rate ($D1$-error) as the evaluation metric, which calculates the percentage of pixels whose disparity errors are greater than a certain threshold.

4.2. Implementation Details

Each model is trained end-to-end using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) on eight NVIDIA Tesla-V100 GPUs. The learning rate is set to 0.001 for training from scratch, and we train each model for 100 epochs with a batch size of 16 using 624×304 random crops. The momentum factor γ in Alg. 1 is set to 0.95. The weights of different loss terms (λ_s^{occ}, λ_t^{ar}, λ_t^{occ}, λ_t^{sm}) in Eq. 5 are set to (0.2, 1.0, 0.2, 0.1). An individual domain-adaptive stereo model is trained for each target domain. The model specifications of the Ada-ResNetCorr and Ada-PSMNet are provided in the supplementary material.

4.3. Ablation Studies

In Tab. 1, we conduct detailed ablation studies on four real-world datasets to evaluate the key components in our domain adaptation pipeline, based on Ada-PSMNet and Ada-ResNetCorr. As can be seen, applying the progressive color transfer algorithm during training can significantly reduce error rates on multiple target domains, *e.g.* 8.3% on KITTI, 8.6% on Middlebury, 4.7% on ETH3D, and 13.5% on DrivingStereo from Ada-PSMNet, benefiting from massive color-aligned training images without geometrical distortions. We also provide qualitative results of color transfer in the supplementary material. In addition, compared with baseline models, error rates are reduced by $1\% \sim 4\%$ on varied target domains by integrating the proposed cost normalization layer, which also works well when implemented together with the input color transfer module. Furthermore, adopting the self-supervised occlusion-aware reconstruction can further reduce error rates by $0.5\% \sim 1.5\%$ on varied target domains, though the adaptation performance is already remarkable through color transfer and cost normalization. Finally, both Ada-PSMNet and Ada-ResNetCorr significantly outperform the corresponding baseline model on all target domains, especially an accuracy improvement of 15.8% from Ada-PSMNet on the large-scale DrivingStereo dataset.

In order to further demonstrate the effectiveness of each module, we perform exhaustive comparisons with other alternative methods respectively. As shown in Tab. 2, our specifically designed cost normalization layer which is parameter-free and adopted only once in the network, outperforms other general and learnable normalization layers (AdaBN [17], BIN [26], IN [42], and DN [54]) which are repeatedly adopted in the network's feature extractor. In Tab. 3, our progressive color transfer algorithm far outperforms three popular color/style transfer networks (WCT^2 [50], WaterGAN [16], and CycleGAN [56]), indicating that geometrical distortions from such GAN-based color/style transfer models are harmful for the low-level stereo matching task. Moreover, our method outperforms the Reinhard's color transfer method [29] by about 1% in $D1$-error, revealing the effectiveness of the proposed progressive update strategy. In Tab. 4, our proposed self-supervised occlusion-aware reconstruction is demonstrated to be a more effective auxiliary task for stereo domain adaptation, while other alternatives all hurt the domain adaptation performance.

Table 1. Ablation studies on the KITTI, Middleburry, ETH, and DrivingStereo training sets. $D1$-error (%) is adopted for evaluation.

Model	cost normalization	color transfer	self-supervised reconstruction	KITTI 2012	KITTI 2015	Middlebury half	Middlebury quarter	ETH	DrivingStereo
Ada-PSMNet	✗	✗	✗	13.6	12.1	18.6	11.5	10.8	20.9
	✓	✗	✗	11.8	9.1	16.8	10.1	9.0	16.7
	✗	✓	✗	5.3	5.4	10.0	5.8	6.1	7.4
	✓	✓	✗	4.5	4.7	9.0	5.1	5.2	6.4
	✓	✓	✓	**3.6**	**3.5**	**8.4**	**4.7**	**4.1**	**5.1**
Ada-ResNetCorr	✗	✗	✗	9.8	9.4	22.5	12.8	15.8	17.2
	✓	✗	✗	8.1	8.4	19.7	10.9	13.4	15.2
	✗	✓	✗	6.7	6.7	15.1	8.3	7.1	10.2
	✓	✓	✗	6.0	5.9	13.7	7.4	6.6	9.2
	✓	✓	✓	**5.1**	**5.0**	**12.7**	**6.6**	**5.8**	**8.0**

Table 2. Comparisons with existing normalization layers on the KITTI and DrivingStereo training sets. $D1$-error (%) is adopted.

Methods	KITTI	DrivingStereo
PSMNet Baseline	12.1	20.9
+Adaptive Batch Norm [17]	11.8	20.3
+Batch-Instance Norm [26]	11.2	19.5
+Instance Norm [42]	10.7	18.6
+Domain Norm [54]	9.5	17.2
+Our Cost Norm	**9.1**	**16.7**

Table 3. Comparisons with color/style transfer methods on the KITTI and DrivingStereo training sets. $D1$-error (%) is adopted.

Methods	KITTI	DrivingStereo
PSMNet Baseline	12.1	20.9
+WCT^2 [50]	10.2	17.3
+WaterGAN [16]	8.7	11.5
+CycleGAN [56]	8.0	10.6
+Color Transfer [29]	6.2	8.3
+Our Progressive Color Transfer	**5.4**	**7.4**

Table 4. Comparisons with other auxiliary tasks for stereo domain adaptation on the KITTI training set. $D1$-error (%) is adopted.

Methods	KITTI
PSMNet + Our Color Transfer + Our Cost Norm (Baseline)	4.7
+Patch Localization Task [45]	6.5
+Rotation Prediction Task [8]	6.1
+Flip Prediction Task [45]	6.0
+Our Self-Supervised Reconstruction Task	**3.5**

4.4. Cross-Domain Comparisons

In Tab. 5, we compare our proposed domain-adaptive stereo models with other traditional stereo methods, domain generalization, and domain adaptation stereo networks on three real-world datasets. Firstly, both Ada-ResNetCorr and Ada-PSMNet show great superiority over traditional stereo methods. Secondly, for comparisons with domain generalization networks, unfairness may exist since our domain-adaptive models use target-domain images during training. It is caused by the problem definition of domain adaptation as mentioned in Sec. 3.1. However, as can be seen in Tab. 5, our Ada-PSMNet achieves tremendous gains rather than small deltas compared with all domain generalization networks, including the state-of-the-art DSMNet [54] and its baseline network GANet [53]. Lastly, among the few published domain adaptation networks, only the StereoGAN [20] reported such cross-domain performance, while our Ada-PSMNet achieves a 3.5 times lower error rate than StereoGAN [20] on the KITTI training set. Hence, our proposed multi-level alignment pipeline successfully address the domain adaptation problem for stereo matching. In Fig. 4, we provide qualitative results of our method on different real-world datasets, in which Ada-PSMNet predicts accurate disparity maps on both outdoor and indoor scenes.

Table 5. Cross-domain comparisons with other traditional / domain generalization / domain adaptation stereo methods on the KITTI, Middlebury, and ETH3D training sets. $D1$-error (%) is adopted. The second and third columns indicate whether the method is trained on the SceneFlow dataset, and whether the method uses target-domain images during training respectively.

Methods	Train SceneFlow	Target Images	Test KITTI	Test Middlebury	Test ETH
Traditional Stereo Methods					
PatchMatch [1]	✗	✗	17.2	38.6	24.1
SGM [13]	✗	✗	7.6	25.2	12.9
Domain Generalization Stereo Networks					
HD^3 [49]	✓	✗	26.5	37.9	54.2
GWCNet [11]	✓	✗	22.7	34.2	30.1
PSMNet [4]	✓	✗	16.3	25.1	23.8
GANet [53]	✓	✗	11.7	20.3	14.1
DSMNet [54]	✓	✗	6.5	13.8	6.2
Domain Adaptation Stereo Networks					
StereoGAN [20]	✓	✓	12.1	-	-
Ada-ResNetCorr	✓	✓	5.0	12.7	5.8
Ada-PSMNet	✓	✓	**3.5**	**8.4**	**4.1**

4.5. Evaluations on Stereo Benchmarks

We further compare our domain-adaptive stereo model Ada-PSMNet with several unsupervised/self-supervised methods and finetuned disparity networks on public stereo matching benchmarks: KITTI, Middlebury, and ETH3D. We directly upload the results from our SceneFlow-pretrained model to the online benchmark, and do not finetune using target-domain ground-truths before submitting test results.

Table 6. Performance on the ETH3D stereo benchmark. The 1-pixel error (%) and 2-pixel error (%) are adopted for evaluation.

Method	Deep-Pruner	iResNet	Stereo-DRNet	SGM-Forest	PSMNet	DispNet	*Ada-PSMNet*
Use ETH-gt	✓	✓	✓	✗	✓	✓	✗
Bad 1.0	3.52	3.68	4.46	4.96	5.02	17.47	**3.09**
Bad 2.0	0.86	1.00	0.83	1.84	1.09	7.91	**0.65**

Table 7. Performance on the Middlebury stereo benchmark. The 2-pixel error (%) is adopted for evaluation.

Method	EdgeStereo	CasStereo	iResNet	MCV-MFC	Deep-Pruner	PSMNet	*Ada-PSMNet*
Use Mid-gt	✓	✓	✓	✓	✓	✓	✗
Bad 2.0	18.7	18.8	22.9	24.8	30.1	42.1	**13.7**

Table 8. Performance on the KITTI 2015 stereo benchmark. The $D1$-error (%) is adopted for evaluation.

Method	GC-Net	L-ResMatch	SGM-Net	MC-CNN	DispNetC	Weak-Sup	MADNet	Unsupervised	*Ada-PSMNet*
Use KITTI-gt	✓	✓	✓	✓	✓	✓	✗	✗	✗
$D1$-error	**2.87**	3.42	3.66	3.89	4.34	4.97	8.23	9.91	3.08

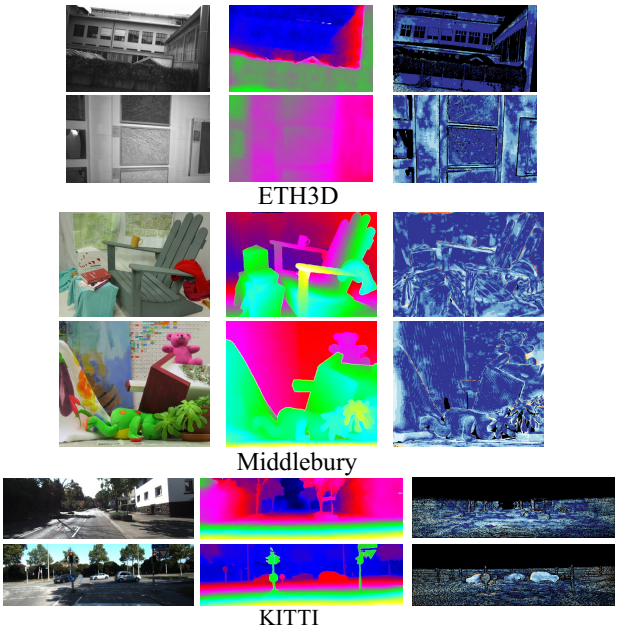

Figure 4. Disparity predictions from our SceneFlow-pretrained Ada-PSMNet on the ETH3D, Middlebury, and KITTI datasets. Left-right: left image, colorized disparity map, and error map.

4.5.1 Results on the ETH3D Benchmark

As can be seen in Tab. 6, SceneFlow-pretrained Ada-PSMNet outperforms the state-of-the-art patch-based model DeepPruner [6], end-to-end disparity networks (iResNet [18], PSMNet [4], and StereoDRNet [3]) finetuned with ground-truth disparities from the ETH3D training set, and state-of-the-art traditional method SGM-Forest [31]. By the time of the paper submission, AdaStereo ranks 1^{st} on the ETH3D benchmark in terms of the 2-pixel error metric.

4.5.2 Results on the Middlebury Benchmark

As can be seen in Tab. 7, SceneFlow-pretrained Ada-PSMNet significantly outperforms all other state-of-the-art end-to-end disparity networks (EdgeStereo [35], CasStereo [9], iResNet [18], MCV-MFC [19], and PSMNet [4]) which are finetuned using ground-truth disparities from the Middlebury training set. Our Ada-PSMNet achieves a remarkable 2-pixel error rate of 13.7% on the full-resolution test set, outperforming all other finetuned end-to-end stereo matching networks on the benchmark.

4.5.3 Results on the KITTI Benchmark

As can be seen in Tab. 8, SceneFlow-pretrained Ada-PSMNet far outperforms the online-adaptive model MADNet [38], weakly supervised [40] and unsupervised [55] methods, meanwhile achieving higher accuracy than some supervised disparity networks including MC-CNN-acrt [51], L-ResMatch [34], DispNetC [24], and SGM-Net [33]. Moreover, our Ada-PSMNet achieves comparable performance with the KITTI-finetuned GC-Net [15].

5. Conclusions

In this paper, we focus on the domain adaptation problem for deep stereo networks. Following the standard domain adaptation methodology, we propose a novel domain adaptation pipeline specifically for stereo matching task, in which multi-level alignments are conducted: a non-adversarial progressive color transfer algorithm for input-level alignment; a parameter-free cost normalization layer for internal feature-level alignment; a highly related self-supervised auxiliary task for output-space alignment. We verify our SceneFlow-pretrained domain-adaptive models on four real-world datasets, and state-of-the-art cross-domain performance is achieved on all target domains. Our AdaStereo model also achieves remarkable performance on multiple stereo matching benchmarks without finetuning.

References

[1] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *BMVC*, 2011. 7

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017. 3

[3] Rohan Chabra, Julian Straub, Christopher Sweeney, Richard Newcombe, and Henry Fuchs. Stereodrnet: Dilated residual stereonet. In *CVPR*, 2019. 8

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018. 1, 2, 6, 7, 8

[5] Zhenwei Dai and Reinhard Heckel. Channel normalization in convolutional neural network avoids vanishing gradients. *arXiv preprint arXiv:1907.09539*, 2019. 5

[6] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *CVPR*, 2019. 8

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 2, 6

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 3, 5, 7

[9] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, 2020. 8

[10] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *ECCV*, 2018. 3

[11] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, 2019. 2, 7

[12] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[13] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 2008. 7

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 3

[15] 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 *ICCV*, 2017. 2, 8

[16] Jie Li, Katherine A Skinner, Ryan M Eustice, and Matthew Johnson-Roberson. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *RAL*, 2017. 4, 6, 7

[17] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016. 6, 7

[18] Zhengfa Liang, Yiliu Feng, Yulan Guo, and Hengzhu Liu. Learning deep correspondence through prior and posterior feature constancy. In *CVPR*, 2018. 2, 8

[19] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. Stereo matching using multi-level cost volume and multi-scale feature constancy. *TPAMI*, 2019. 8

[20] 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 *CVPR*, 2020. 3, 7

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 3

[23] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 3

[24] 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 *CVPR*, 2016. 1, 2, 3, 6, 8

[25] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 1, 2, 3, 6

[26] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *NeurIPS*, 2018. 5, 6, 7

[27] Jiahao Pang, Wenxiu Sun, JS Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *ICCV Workshop*, 2017. 2

[28] Jiahao Pang, Wenxiu Sun, Chengxi Yang, Jimmy Ren, Ruichao Xiao, Jin Zeng, and Liang Lin. Zoom and learn: Generalizing deep stereo matching to novel domains. In *CVPR*, 2018. 3

[29] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *CGA*, 2001. 4, 6, 7

[30] 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 *GCPR*, 2014. 1, 2, 3, 6

[31] Johannes L Schonberger, Sudipta N Sinha, and Marc Pollefeys. Learning to fuse proposals from multiple scanline optimizations in semi-global matching. In *ECCV*, 2018. 8

[32] 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 *CVPR*, 2017. 2, 3, 6

[33] Akihito Seki and Marc Pollefeys. Sgm-nets: Semi-global matching with neural networks. In *CVPR*, 2017. 8

[34] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *CVPR*, 2017. 8

[35] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *IJCV*, 2020. 2, 6, 8

[36] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *ACCV*, 2018. 2

[37] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019. 3

[38] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *CVPR*, 2019. 3, 8

[39] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 3

[40] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Weakly supervised learning of deep metrics for stereo reconstruction. In *ICCV*, 2017. 8

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 3

[42] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2, 5, 6, 7

[43] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 3

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[45] Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *Access*, 7:156694–156706, 2019. 3, 5, 7

[46] Guorun Yang, Zhidong Deng, Hongchao Lu, and Zeping Li. Src-disp: Synthetic-realistic collaborative disparity learning for stereo matching. In *ACCV*, 2018. 6

[47] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*, 2019. 2, 6

[48] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *ECCV*, 2018. 2

[49] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *CVPR*, 2019. 2, 7

[50] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *ICCV*, 2019. 6, 7

[51] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *CVPR*, 2015. 8

[52] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017. 3

[53] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, 2019. 2, 7

[54] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *ECCV*, 2020. 2, 5, 6, 7

[55] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 8

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 3, 4, 6, 7