StereoGAN: Bridging Synthetic-to-Real Domain Gap by Joint Optimization of Domain Translation and Stereo Matching

Rui Liu[1] Chengxi Yang[2] Wenxiu Sun[2] Xiaogang Wang[1] Hongsheng Li[1]
[1]CUHK-SenseTime Joint Laboratory, Chinese University of Hong Kong [2]SenseTime Research
ruiliu@link.cuhk.edu.hk {yangchengxi, sunwenxiu}@sensetime.com
{xgwang, hsli}@ee.cuhk.edu.hk

Abstract

Large-scale synthetic datasets are beneficial to stereo matching but usually introduce known domain bias. Although unsupervised image-to-image translation networks represented by CycleGAN show great potential in dealing with domain gap, it is non-trivial to generalize this method to stereo matching due to the problem of pixel distortion and stereo mismatch after translation. In this paper, we propose an end-to-end training framework with domain translation and stereo matching networks to tackle this challenge. First, joint optimization between domain translation and stereo matching networks in our end-to-end framework makes the former facilitate the latter one to the maximum extent. Second, this framework introduces two novel losses, i.e., bidirectional multi-scale feature re-projection loss and correlation consistency loss, to help translate all synthetic stereo images into realistic ones as well as maintain epipolar constraints. The effective combination of above two contributions leads to impressive stereo-consistent translation and disparity estimation accuracy. In addition, a mode seeking regularization term is added to endow the synthetic-to-real translation results with higher fine-grained diversity. Extensive experiments demonstrate the effectiveness of the proposed framework on bridging the synthetic-to-real domain gap on stereo matching.

1. Introduction

With the fast development of deep neural networks [23, 12] and large-scale benchmarks [31, 13, 7], deep learning-based stereo matching methods have made great progress in the past decade [29, 19]. These methods, however, relying on a large quantity of high-quality *left-right-disparity* training data. Although the input images to the stereo matching networks (*i.e.*, left and right images) are relatively easy to collect using stereo rigs in the real world, their corresponding ground-truth disparities are very difficult to collect. Instead, researchers tend to create synthetic training datasets [29, 31, 13] with perfect disparities. In this way, the demand of large quantity of training data is alleviated. However, the non-negligible domain gaps between synthetic and real must be considered when generalizing to real domains. In order to mitigate the domain gaps, some of the previous works [1, 40] train their models in two stages. Firstly the model is trained on synthetic dataset and then fine-tuned on a particular real dataset in either supervised [30, 1, 11] or unsupervised manner [38, 39]. In this paper, we focus on the latter one, a more challenging task with no ground-truth for the real target-domain data.

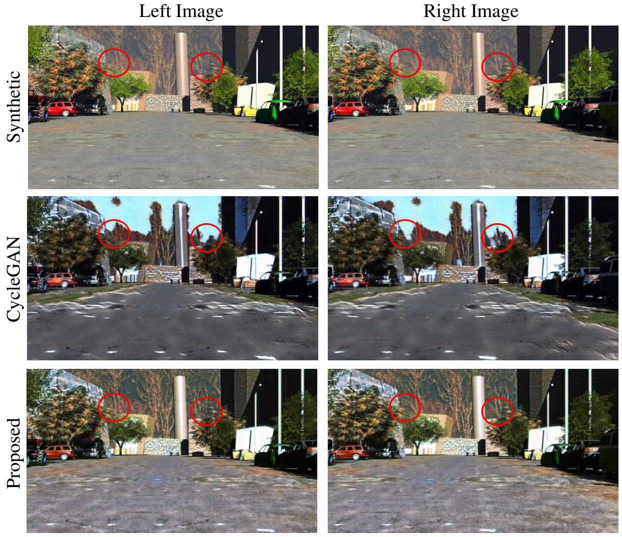

Figure 1. Domain translation results. Top row: stereo images from synthetic domain. Middle row: synthetic-to-real translated results by CycleGAN. Bottom row: synthetic-to-real translated results by our proposed model.

Existing unsupervised online adaptation methods advanced the research progress, however, still have difficulties on handling the domain gaps between source and target domains [38, 39]. Moreover, these methods introduce extra computation compared to a feed-forward neural network, although they have striven to reduce the computation complexity of updating network parameters [40].

Recently, unsupervised image-to-image translation models achieved great success [47, 25, 24] and thus were adopted in domain adaptation methods to tackle many applications such as semantic segmentation, person re-identification and object detection [15, 41, 32, 3]. However, it is non-trivial to generalize this series of methods to stereo matching. The middle row of Figure 1 reveals two main challenges for translation in stereo matching. 1) The general image-to-image translation does not take epipolar constraints into consideration, which leads to inconsistent textures and thus ambiguity of disparity, as emphasized by red circles. 2) It only attempts to transfer domain styles while neglecting the fact that its purpose should be serving the stereo matching networks. For instance, since most background of our synthetic images is brown mountains while that of real images in the training set is blue sky, the vanilla CycleGAN [47] regards this to be domain style and tries to translate from brown mountains to blue sky as shown in the first two rows of Figure 1. This would confuse stereo matching network, because the useful textures for stereo matching in the sky is definitely much less than those in the mountains. In this paper, we successfully addressed these two challenges by properly designed stereo constraints and joint training scheme. The intermediate image translation results are shown in the bottom row of Figure 1.

In particular, we propose an end-to-end deep learning framework consisting of domain translation and stereo matching networks to estimate stereo disparity on the target domain, using only source-domain synthetic stereo image pairs with ground-truth disparity and target-domain real stereo image pairs without any annotation. The stereo image translation is constrained by a novel bidirectional multi-scale feature re-projection loss and a correlation consistency loss. The former one is realized by a multi-scale feature re-projection module. For feature maps at each layer of domain translation networks, the inverse warping [17] of the right feature map according to the given disparity should be as close as its corresponding left feature map. Both ground-truth disparity for synthetic data and estimated disparity for real data would contribute to joint training in a bidirectional manner. We also introduce a correlation consistency loss to ensure that the reconstructed stereo images should maintain consistent correlation feature maps, which are extracted from the stereo matching network, with those original images.

In addition, we observed that real stereo pairs usually do not exactly match each other due to different camera configurations and settings. To this end, inspired by successful applications of using noise to manipulate image [18, 28], we propose a mode seeking regularization term to ensure the fine-grained diversity in synthetic-to-real translation, as shown in Figure 2. As we could observe as circled in red, the local intensity between the left image and right image varies, which simulates the real data. With such augmentation, the domain translation makes the stereo matching in the real domain more robust and effective.

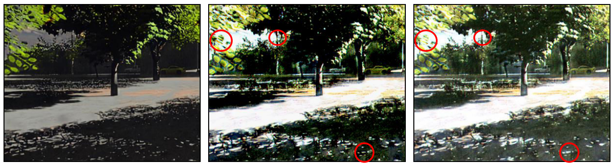

Figure 2. The effect of mode seeking regularization term. Leftmost image is from synthetic domain, and middle image and rightmost image are translated from leftmost image with different random maps. Red circles emphasize the fine-grained difference between middle image and rightmost image. Please zoom in to observe more details.

In summary, our contributions are listed as follows:

- We for the first time combine unsupervised domain translation with disparity estimation in an end-to-end framework to tackle the challenging problem of stereo matching in the absence of real ground-truth disparities.

- We propose novel stereo constraints including the bidirectional multi-scale feature re-projection loss and the correlation consistency loss, which better regularizes this joint framework to achieve stereo-consistent translation and accurate stereo matching. The additional mode seeking regularization endows the synthetic-to-real translation with higher fine-grained diversity.

- Extensive experiments demonstrate that our proposed model outperforms the state-of-the-art unsupervised adaptation approaches for stereo matching.

2. Related Work

Stereo matching conventionally follows a four-step pipeline including matching cost computation, cost aggregation, disparity optimization and post-processing [33]. Local descriptors such as absolute difference (AD), sum of squared difference (SAD) and so on are usually adopted for measuring left-right inconsistency, so as to calculate matching costs for all possible disparities. Cost aggregation and disparity optimization are usually treated as a 2D graph partitioning problem, which could be optimized by graph cut [22] or belief propagation [37, 21]. Semi-global matching (SGM) [14] approximates the global optimization with dynamic programming.

Deep learning-based stereo matching methods have achieved great progress due to the rise of deep neural networks [23, 12] and large-scale benchmarks [8, 7] in the last decade. Among them, Zbontar and LeCun [45] for the first time presented the computation of stereo matching costs by a deep Siamese network. Luo *et al.* [27] accelerated the computation of matching costs by correlating unary features. Recently, many end-to-end neural networks were developed to directly predict the whole disparity maps from stereo image pairs [29, 30, 34, 43, 19, 1, 44, 11]. Among them, DispNet [29] is a pioneer work which for the first time uses an end-to-end deep learning framework to directly regress disparity maps. The follow-up work GCNet [19] introduces 3D convolutional networks to aggregate contextual information for obtaining better cost volumes.

Domain adaptation methods have shown great potential in filling the gap between synthetic and real domains. Previous works attempted to solve this problem by either learning domain-invariant representations [4, 5] or pushing two domain distributions to be close [9, 42, 35, 36]. For example, the gap between source and target domain could be filled by matching the distribution [10, 26] or statistics [35, 36] of deep features.

Recently, unsupervised image-to-image translation models achieved great success under unpaired setting [47, 25, 24] and thus were applied as domain adaptation methods in many applications including semantic segmentation, person re-identification and object detection [15, 41, 32, 3].

In the field of stereo matching, unsupervised online adaptation advanced great progress. These methods first train a disparity estimation network on synthetic data and then fine-tune it online using unsupervised loss such as re-projection loss when continuously accessing new stereo pairs from other domains [38, 40]. This unsupervised adaptation strategy is then incorporated in a meta-learning framework [39].

3. Method

Given a set of N synthetic *left-right-disparity* tuples $\{(x_l, x_r, x_d)_i\}_{i=1}^N$ in the source domain $\mathcal{X}$, where $(x_l, x_r, x_d) \in (\mathcal{X}_L, \mathcal{X}_R, \mathcal{X}_D) = \mathcal{X}$, and a set of M real stereo images $\{(y_l, y_r)\}_{j=1}^M$ in the target domain $\mathcal{Y}$ without any ground-truth disparity, where $(y_l, y_r) \in (\mathcal{Y}_L, \mathcal{Y}_R)$, our goal is to learn an accurate disparity estimation network F for estimating the disparity $\hat{y}_d = F(y_l, y_r)$ on the target domain.

For the sake of clear formulation, we define a paired set $(\mathcal{X}_L, \mathcal{X}_R) = \{(x_{l1}, x_{r1}), (x_{l2}, x_{r2}), ..., (x_{lN}, x_{rN})\}$ where (x_{li}, x_{ri}) stands for a paired stereo image, *i.e.*, a left image x_{li} and its corresponding right image x_{ri} (see Eqs. (4-7)). We also define an unpaired set $\{\mathcal{X}_L, \mathcal{X}_R\} = \{x_{l1}, x_{r1}, x_{l2}, x_{r2}, ..., x_{lN}, x_{rN}\}$ where we can only sample a single left or right image (see Eqs. (1-2)).

Different from previous works that directly train stereo matching network F with synthetic data [29, 19, 40], we propose a joint domain translation and stereo matching framework, which aims to translate synthetic-style stereo images into realistic ones with novel stereo constraints and thus better cooperate with the stereo matching network in an end-to-end manner, as shown in Figure 3.

3.1. Cycle-consistency Domain Translation for Stereo Matching

Cycle-consistency domain translation loss. To help synthetic-to-real translation network G_{x2y} capture the global domain style of the real datasets, we adopt a real domain discriminator D_y whose goal is to distinguish synthetic-to-real generated images from real-domain images. On the contrary, G_{x2y} learns to generate images that look similar to real-domain images to fool the real domain discriminator D_y. These two sub-nets constitute a minimax game that optimizes in an adversarial manner and achieves optimal when D_y cannot tell whether images are generated or not. The adversarial loss for synthetic-to-real generation is formulated as:

$$\begin{aligned}\mathcal{L}_{adv}(G_{x2y}, D_y, \mathcal{X}, \mathcal{Y}) = \mathbb{E}_{y\sim\{\mathcal{Y}_L,\mathcal{Y}_R\}}\left[\log D_y(y)\right] \\ + \mathbb{E}_{x\sim\{\mathcal{X}_L,\mathcal{X}_R\}}\left[\log\left(1 - D_y(G_{x2y}(x)\right)\right],\end{aligned} \tag{1}$$

where $y \sim \{\mathcal{Y}_L, \mathcal{Y}_R\}$ means a single real image y is sampled from the non-paired real-domain set $\{\mathcal{Y}_L, \mathcal{Y}_R\}$. We also introduce a similar adversarial loss for supervising the process of real-to-synthetic generation as $\mathcal{L}_{adv}(G_{y2x}, D_x, \mathcal{Y}, \mathcal{X})$.

Adversarial losses could only supervise G_{x2y} and G_{y2x} to produce images that are not distinguishable by domain discriminators, but any random permutation of outputs can happen without any other constraints. In order to regularize G_{x2y} and G_{y2x} to be one-to-one mapping, the cycle consistency loss is also adopted,

$$\begin{aligned}&\mathcal{L}_{cyc}(G_{x2y}, G_{y2x}) \\ &= \mathbb{E}_{y\sim\{\mathcal{Y}_L,\mathcal{Y}_R\}}\left[\|G_{x2y}(G_{y2x}(y)) - y\|_1\right] \\ &+ \mathbb{E}_{x\sim\{\mathcal{X}_L,\mathcal{X}_R\}}\left[\|G_{y2x}(G_{x2y}(x)) - x\|_1\right].\end{aligned} \tag{2}$$

To sum up, the cycle-consistency domain translation loss following the CycleGAN [47] can be defined as

$$\begin{aligned}\mathcal{L}_{cdt}(G_{x2y}, G_{y2x}, D_x, D_y) = \mathcal{L}_{adv}(G_{x2y}, D_y, \mathcal{X}, \mathcal{Y}) \\ + \mathcal{L}_{adv}(G_{y2x}, D_x, \mathcal{Y}, \mathcal{X}) + \lambda_{cyc}\mathcal{L}_{cyc}(G_{x2y}, G_{y2x}).\end{aligned} \tag{3}$$

Stereo matching loss. Since our goal is to learn a mapping from real-domain stereo image to disparity map with only annotated synthetic stereo images and unlabeled real ones, it is straight-forward to take advantage of the results of synthetic-to-real translation. Given a paired synthetic tuple (x_l, x_r, x_d), we argue that the translated stereo pair $(G_{x2y}(x_l), G_{x2y}(x_r))$ could be regarded as real-domain images and such translated stereo pair should match its

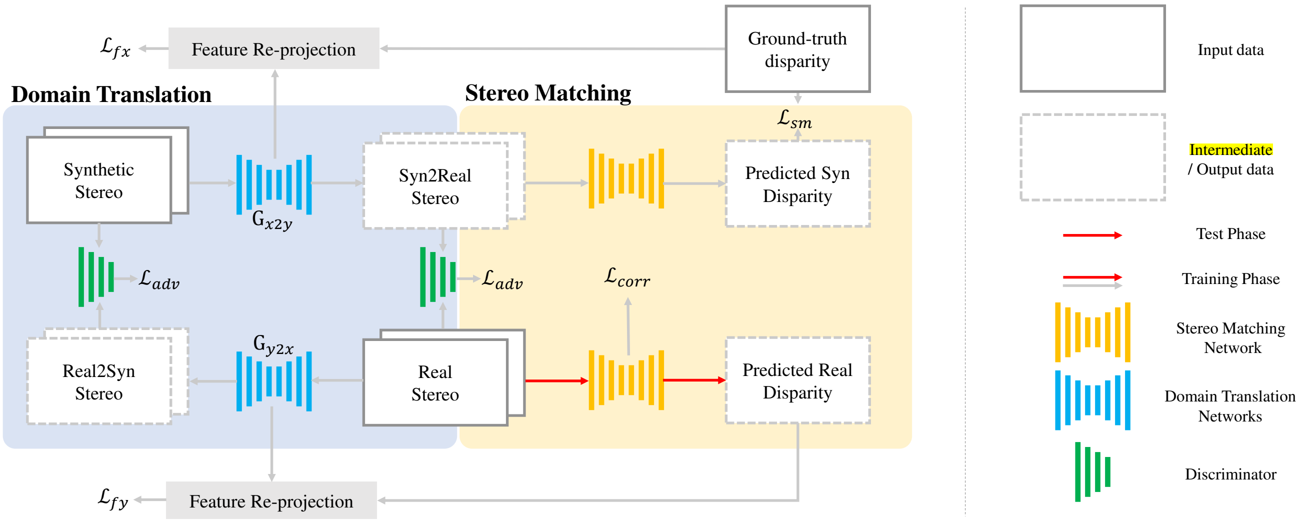

Figure 3. The joint framework of our proposed method. Blue-background block shows our domain translation component and orange-background block shows our stereo matching component. Different blocks, lines and nets are labeled in the rightmost of this figure. F denotes the stereo matching network. Note that we omit cycle consistency loss due to the limited space.

ground-truth disparity x_d. Therefore, we formulate the stereo matching loss as:

$$\mathcal{L}_{sm}(F) = \mathbb{E}_{(x_l,x_r,x_d)\sim\mathcal{X}}\left[\|F(G_{x2y}(x_l), G_{x2y}(x_r)) - x_d\|_1\right], \tag{4}$$

where $F(\cdot,\cdot)$ is the stereo matching network for estimating disparities from real-domain stereo images.

These two losses construct a simple framework that optimizes stereo matching network with the assistance of domain translation networks. However, it may introduce the problem of pixel distortion and stereo mismatch during translation.

3.2. Joint Domain Translation and Stereo Matching

To tackle the above mentioned challenges, we should ensure that domain translation networks only transfer global domain style while maintain the epipolar consistency, which contributes to the improvement of stereo matching. To achieve this, we propose a joint optimization scheme between domain translation and stereo matching with novel constraints.

Before diving into novel constraints, we would first introduce our newly-proposed multi-scale feature re-projection module, which establishes a bidirectional connection between domain translation component and stereo matching component by left-right consistency check, as illustrated in Figure 4. For each intermediate layer of domain translation networks, the inversely warped right feature map should be the same as its corresponding left feature map. This inverse warping operation is completed with properly downsampled disparity map using differentiable bilinear sampling technique [17]. Note that the given disparity could be either ground-truth one for synthetic stereo or estimated one for real stereo, which calculate feature re-projection loss for synthetic or real stereo images respectively. The former endows the domain translation networks with strong epipolar constraints while the latter provides extra supervision for training stereo matching network.

Feature re-projection loss for synthetic images. We argue that the intermediate feature maps for generating the domain-translated left and right images should be the same at 3D physical locations. To model this constraint, we utilize synthetic ground-truth disparity to warp the intermediate feature maps of both G_{x2y} and G_{y2x} along the synthetic-real-synthetic cycle translation. If the stereo image pairs are well translated, the inversely warped right feature map should match the left feature exactly. The feature re-projection loss for synthetic images is formulated as

$$\begin{aligned}
&\mathcal{L}_{fx}(G_{x2y}, G_{y2x}) \\
&= \mathbb{E}_{(x_l,x_r,x_d)\sim\mathcal{X}} \frac{1}{T_1}\sum_{i=1}^{T_1}\left[\left\|W(G^{(i)}_{x2y}(x_r), x_d) - G^{(i)}_{x2y}(x_l)\right\|_1\right. \\
&\left. + \left\|W(G^{(i)}_{y2x}(G_{x2y}(x_r)), x_d) - G^{(i)}_{y2x}(G_{x2y}(x_l))\right\|_1\right],
\end{aligned} \tag{5}$$

where T_1 is the total number of layers of translation networks, $G^{(i)}(x)$ denotes the feature of image x at ith-layer the translation network G, the inverse warping function $W(G^{(i)}(x_r), x_d)$ warps the right feature map $G^{(i)}(x_r)$ with the ground-truth disparity x_d.

Feature re-projection loss for real images. For a general stereo matching network such as DispNet [29], it naturally outputs multi-scale disparities, which can be formed from correlation features at different neural network layers. These multi-scale disparity maps can be used to warp the in-

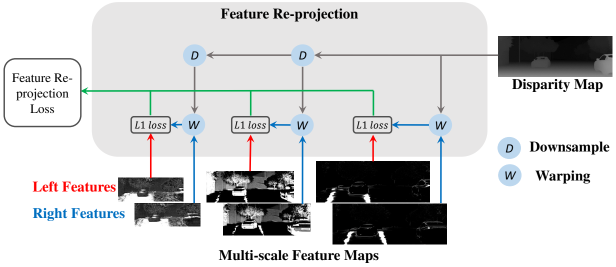

Figure 4. Detailed structure of feature re-projection module. This figure demonstrates the calculating process of feature re-projection loss for synthetic data with ground-truth disparity. Note that stereo matching networks usually output multi-scale disparities, so we remove downsample function when dealing with real data.

termediate feature maps for both G_{x2y} and G_{y2x} along the real-synthetic-real cycle translation. Then the $L1$ distance between the left feature and the inversely warped right feature provides an extra supervision for updating the parameters of disparity estimation network F. This loss could be formulated as

$$\begin{aligned}&\mathcal{L}_{fy}(F)\\ &= \mathbb{E}_{(y_l,y_r)\sim(\mathcal{Y}_L,\mathcal{Y}_R)} \frac{1}{T_1}\sum_{i=1}^{T_1}\Big[\Big\|W(G^{(i)}_{y2x}(y_r),\hat{y}_d) - G^{(i)}_{y2x}(y_l)\Big\|_1\\ &+ \Big\|W(G^{(i)}_{x2y}(G_{y2x}(y_r)),\hat{y}_d) - G^{(i)}_{x2y}(G_{y2x}(y_l))\Big\|_1\Big],\end{aligned} \tag{6}$$

where $\hat{y}_d$ is the estimated disparity of real stereo image pairs by $F(y_l, y_r)$.

Different from previous works which directly warp images at the origin scale [6, 46], our warping operation is based on multi-scale feature maps. Since features at different layers model image structures of different scales, this constraint could help supervise the training of stereo matching network from multiple scales (from global to local regions), leading to impressive improvement on disparity estimation accuracy. In addition, it leaves some space for fine-grained noise modeling upon pixel level (see Figure 2), which would be introduced in the mode seeking regularization term, described later in this section.

Correlation consistency loss. Feature re-projection losses may not totally address the stereo-mismatch issue yet. Since there is no ground-truth disparity for real-domain stereo images, warping features with estimated disparity may introduce some bias into the joint framework. For example, the value of $\mathcal{L}_{fy}$ for a certain *left-right-disparity* tuple may be 0, but it still makes a limited effect on stereo matching, even makes a negative effect. This is because the phenomenon of pixel distortion during domain translation and inaccurate estimation during stereo matching occur simultaneously.

To reduce such impact, stereo matching network is utilized to supervise both G_{y2x} and G_{x2y} along the real-synthetic-real cycle translation. We denote the reconstructed real image by such cycle translation as $y' = G_{x2y}(G_{y2x}(y))$ for ease of presentation. Given a pair of real stereo images (y_l, y_r), we could obtain their reconstructed pair (y'_l, y'_r). The correlation features of (y'_l, y'_r) from each layer of stereo matching network should match those of (y_l, y_r). In addition, we make a cross-pair for constructing a tighter loss, which is calculated by pushing correlation features of both (y'_l, y_r) and (y_l, y'_r) to be close to those of (y_l, y_r). Therefore, we formulate this constraint for real-domain images as the correlation consistency loss between multi-layer correlation features:

$$\begin{aligned}&\mathcal{L}_{corr}(G_{x2y}, G_{y2x})\\ &= \mathbb{E}_{(y_l,y_r)\sim(\mathcal{Y}_L,\mathcal{Y}_R)} \frac{1}{T_2}\sum_{i=1}^{T_2}\Big[\Big\|F^{(i)}(y'_l, y_r) - F^{(i)}(y_l, y_r)\Big\|_1\\ &+ \Big\|F^{(i)}(y_l, y'_r) - F^{(i)}(y_l, y_r)\Big\|_1\\ &+ \Big\|F^{(i)}(y'_l, y'_r) - F^{(i)}(y_l, y_r)\Big\|_1\Big],\end{aligned} \tag{7}$$

where T_2 is the total number of correlation aggregation layers which are after the individual image feature encoding layers and $F^{(i)}(y_l, y_r)$ denotes the correlation aggregation feature of the stereo pair (y_l, y_r) at ith-layer of the stereo matching network F.

Mode seeking loss. The above losses could well maintain the stereo consistency of the domain-translated images. However, in practice, the stereo images also show slight variations between the left and right images, because of sensor noise, different camera configurations, etc. To model such left-right image variations, we propose a mode seeking regularization term following [28] to make the generators create small but realistic variations between the generated left and right images, as demonstrated in Figure 2. A Gaussian random map z is introduced into the synthetic-to-real translation networks $G_{x2y}(x, z)$ to model the variations of the generated images. When training domain translation networks, we attempt to maximize the $L1$ distance between two generated outputs from the same original image x with two different random maps z_1 and $z_2 \sim p(z)$, where $p(z)$ denotes a prior Gaussian distribution with zero mean and unity variance. Since this term has no optimal point, we linearly decay its weight to zero during training. This loss is formulated as

$$\begin{aligned}&\mathcal{L}_{ms}(G_{x2y})\\ &= \mathbb{E}_{x\sim\{\mathcal{X}_L,\mathcal{X}_R\},z_1,z_2\sim p(z)}\left[\frac{\|z_1 - z_2\|_1}{\|G_{x2y}(x, z_1) - G_{x2y}(x, z_2)\|_1}\right].\end{aligned} \tag{8}$$

Dataset	Method	D1-all (%)		EPE		>2px (%)		>4px (%)		>5px (%)		Time
		Noc	All	Noc	All	Noc	All	Noc	All	Noc	All	(s)
Synthia to KITTI2015	Inference	10.75	11.14	1.817	1.961	20.52	20.86	8.40	8.85	5.68	6.06	0.06
	SL+Ad [40]	10.02	10.58	1.596	1.724	19.86	20.16	7.98	8.42	5.53	5.82	0.19
	L2A+Wad [39]	9.88	10.48	1.569	1.697	17.32	17.70	6.78	7.12	5.01	5.54	0.23
	CycleGAN	10.20	10.69	1.653	1.890	17.83	18.15	6.83	7.39	5.10	5.65	0.06
	Proposed	**8.78**	**9.26**	**1.488**	**1.631**	**15.74**	**16.09**	**5.73**	**6.17**	**4.55**	**5.08**	0.06
Driving to KITTI2015	Inference	52.65	53.07	9.351	9.513	63.95	64.30	45.07	45.52	39.33	39.79	0.06
	SL+Ad [40]	39.16	39.49	4.698	4.775	53.33	53.61	30.22	30.56	24.18	24.52	0.19
	L2A+Wad [39]	26.33	26.90	2.878	3.017	40.59	41.57	17.31	18.01	12.55	13.27	0.23
	CycleGAN	31.23	31.74	3.272	3.444	44.34	45.29	19.76	20.34	15.08	15.68	0.06
	Proposed	**25.18**	**25.71**	**2.584**	**2.752**	**39.16**	**40.24**	**15.83**	**16.55**	**11.04**	**11.60**	0.06

Table 1. Evaluation results of the proposed method compared to different methods on Synthia-to-KITTI2015 and Driving-to-KITTI2015. Lower value means better performance.

3.3. Full Objective and Optimization

Putting all the losses introduced above into an overall objective function, we obtain

$$\begin{aligned}&\mathcal{L}(F, G_{x2y}, G_{y2x}, D_x, D_y)\\=&\mathcal{L}_{cdt}(G_{x2y}, G_{y2x}, D_x, D_y) + \lambda_{sm}\mathcal{L}_{sm}(F)\\+&\lambda_{fx}\mathcal{L}_{fx}(G_{x2y}, G_{y2x}) + \lambda_{fy}\mathcal{L}_{fy}(F)\\+&\lambda_{corr}\mathcal{L}_{corr}(G_{x2y}, G_{y2x}) + \lambda_{ms}\mathcal{L}_{ms}(G_{x2y}),\end{aligned} \tag{9}$$

where $\lambda_s, s \in \{sm, fx, fy, corr, ms\}$ weigh the relative importance among different objectives. We would discuss the effectiveness of each objective in Section 4 by ablation study. Our final goal is to solve the following optimization problem:

$$\max_{D_x, D_y} \min_{F, G_{x2y}, G_{y2x}} \mathcal{L}(F, G_{x2y}, G_{y2x}, D_x, D_y). \tag{10}$$

4. Experiment

4.1. Implementation Detials

Network and training. We adopt the architecture for our generator and dicriminator networks from CycleGAN [47] with patch discriminator [16] and take DispNet [29] as our stereo matching network. We implement this method on *Pytorch*. For training our proposed joint domain translation and stereo matching framework, we partition the training into two stages. In the warm-up stage, we first train the domain translation networks with only $\mathcal{L}_{cdt}$ and $\mathcal{L}_{fx}$ for 10 epochs, using Adam optimizer [20] with the momentum $\beta_1 = 0.5, \beta_2 = 0.999$ and learning rate $\alpha = 0.0002$. Then we train the stereo matching network with only $\mathcal{L}_{sm}$ for 50 epochs, using Adam optimizer with the momentum $\beta_1 = 0.9, \beta_2 = 0.999$ and learning rate $\alpha = 0.0001$. In the second stage, we train these two components together in an end-to-end manner and maintain the hyper-parameters unchanged. We alternatively optimize domain translation nets and stereo matching net with the full objective. We empirically set the trade-off factors as $\lambda_{cyc} = 10$, $\lambda_{sm} = 1$, $\lambda_{fx} = 5$, $\lambda_{fy} = 5$, $\lambda_{corr} = 1$ and $\lambda_{ms} = 0.1$.

Datasets. We take three datasets to testify the effectiveness of our proposed method. Two of them are synthetic datasets and the last one is real dataset. The first is Driving, a subset of a large synthetic dataset Sceneflow [29], which describes a virtual-world car driving scene. It contains fast sequences and slow sequences with both forward driving and backward driving scenes, the number of images summing up to $4,400$ totally. The image size in this dataset is 540×960 and the range of disparity value is $0 - 300$. The second is Synthia-SF [31], which contains 6 sequences featuring different scenarios and traffic conditions. There are $2,224$ images with associated ground-truth disparity maps. The image size is 1080×1920 and its range of disparity is similar to Driving dataset. The last real dataset is KITTI2015 [7], containing 200 training images collected in real scenarios. Its image size is around 385×1242 with disparity ranging from 0 to around 180. Due to the inconsistency of object size between Synthia-SF and KITTI2015, we resize all images in Synthia-SF to half and the corresponding disparity value is divided by 2.

Evaluation metrics. We testify the effectiveness of our proposed method by the following evaluation metrics. Endpoint error (EPE) is the mean average disparity error in pixels. D1-all means the percentage of pixels whose absolute disparity error is larger than 3 pixels or 5% of ground-truth disparity value. Percentages of erroneous pixels larger than $2, 4, 5$ are reported. All these evaluation metrics are calculated for both non-occluded (Noc) and all (All) pixels. The inference time on single TITAN-X GPU is also recorded.

4.2. Comparison with Other Methods

We first investigate whether the proposed method is superior to other related methods or not, whose results are summarized in Table 1. We take two synthetic data - Syn-

Dataset	Ablation objective	D1-all (%)		EPE		>2px (%)		>4px (%)		>5px (%)	
		Noc	All	Noc	All	Noc	All	Noc	All	Noc	All
Synthia to KITTI2015	w/o $\mathcal{L}_{corr}$	8.95	9.45	1.532	1.675	16.00	16.36	5.88	6.32	4.65	5.20
	w/o $\mathcal{L}_{fx}$	9.46	10.02	1.570	1.706	16.89	17.13	6.34	6.79	4.98	5.43
	w/o $\mathcal{L}_{fy}$	9.32	9.89	1.552	1.690	16.73	16.95	6.20	6.62	4.84	5.31
	w/o $\mathcal{L}_{ms}$	9.04	9.53	1.538	1.668	16.13	16.48	5.94	6.43	4.72	5.26
	full obj.	**8.78**	**9.26**	**1.488**	**1.631**	**15.74**	**16.09**	**5.73**	**6.17**	**4.55**	**5.08**
Driving to KITTI2015	w/o $\mathcal{L}_{corr}$	25.64	26.16	2.633	2.804	39.88	40.96	16.46	17.12	11.57	12.11
	w/o $\mathcal{L}_{fx}$	26.38	26.95	2.883	3.029	40.61	41.64	17.28	17.96	12.64	13.28
	w/o $\mathcal{L}_{fy}$	26.22	26.79	2.843	2.998	40.42	41.28	17.06	17.85	12.09	12.66
	w/o $\mathcal{L}_{ms}$	25.45	25.98	2.601	2.782	39.76	40.85	16.30	16.95	11.34	11.93
	full obj.	**25.18**	**25.71**	**2.584**	**2.752**	**39.16**	**40.24**	**15.83**	**16.55**	**11.04**	**11.60**

Table 2. Evaluation results of the proposed method with different objectives by ablation study. Lower value means better performance.

thia and Driving as our source-domain dataset, and one real dataset - KITTI2015 as our target-domain dataset. A dubbed method without domain translation, which is called Inference, is to train the stereo matching network on synthetic data and then directly predict disparity map on real data. Two state-of-the-art unsupervised adaptation methods for stereo matching are compared. Particularly, we use SL+Ad to denote unsupervised online adaptation method described in [40] and use L2A+Wad to denote unsupervised adaptation via meta learning framework described in [39]. Moreover, since there is no stereo matching-specific domain adaptation technique developed, we choose CycleGAN [47] as our baseline for comparison. For the sake of fair comparison, we set the stereo matching network of all methods to DispNet [29].

As could be seen from Table 1, all of the methods perform better on Synthia-to-KITTI2015 than Driving-to-KITTI2015 because there is a larger gap between Driving and KITTI2015. Among these methods, Inference perform worst due to the natural gap between synthetic and real domain. SL+Ad updates the stereo matching network by calculating the error between the inversely-warped left image and real left image when accessing new stereo images. L2A+Wad proposes a novel weight confidence-guided adaptation technique and updates the network in a meta-learning manner. These two methods mitigated the domain gap to a little bit extent but meanwhile brought some extra calculation burden to inference process. Their inference time increase from 0.06 seconds to 0.19 and 0.23 seconds respectively. The translation results of CycleGAN have the problem of pixel distortion, as introduced in Section 1, so it performed not well enough. The proposed joint domain translation and stereo matching framework, with novel stereo constraints, beat all the above methods by reducing the number of erroneous pixels considerably. The significant improvements in all evaluation metrics demonstrate the superiority of our method. In addition, the inference time of our method is same as that of original DispNet because all the extra domain translation and auxiliary training is completed in the procedure of offline training.

4.3. Ablation Study

We then investigate how each objective term influence the performance of unsupervised stereo matching quantitatively by ablation study. Besides cycle domain translation loss and stereo matching loss, we propose four novel objectives for regularizing the basic problem formulation including correlation consistency loss, mode seeking loss, feature re-projection loss for real stereo and for synthetic stereo. We would train our joint framework by removing one of them and then record the corresponding D1-all, EPE, and bad pixel percentage with threshold 2, 4 and 5, as summarized in Table 2. The results of ablation study on both Synthia and Driving source dataset show similar trend. In general, feature re-projection loss for synthetic stereo and real stereo is more effective than that of correlation consistency loss and mode seeking loss. We try to analyze the reasons in the following.

First of all, among all four proposed objectives, feature re-projection loss for synthetic stereo $\mathcal{L}_{fx}$ is most effective on our joint framework. The reasons are as follows: 1) it ensures that translated outputs be stereo-consistent with inputs, which is vital to stereo matching loss in the presence of a large amount accurate disparities; 2) it benefits the training of stereo matching network with feature re-projection loss for real stereo by well-learned translation networks.

The effect of feature re-projection loss for real stereo $\mathcal{L}_{fy}$ is runner-up, because it actually provides extra training signals for training stereo matching network. However, such supervision signals are obtained from the warping of features in domain translation networks, so its performance is highly dependent on how well domain translation networks are trained by $\mathcal{L}_{fx}$ to a large degree.

Thirdly, correlation consistency loss $\mathcal{L}_{corr}$ may contribute to this framework marginally in the presence of feature re-projection losses. It serves as a complement to $\mathcal{L}_{fx}$.

Synthia-to-KITTI2015					
Models	Inference		Proposed		Time
	D1-all	EPE	D1-all	EPE	(s)
DispNet	11.14	1.961	9.26	1.631	0.06
GwcNet [11]	7.46	1.576	5.74	1.424	0.32
Driving-to-KITTI2015					
Models	Inference		Proposed		Time
	D1-all	EPE	D1-all	EPE	(s)
DispNet	53.07	9.513	25.71	2.752	0.06
GwcNet [11]	28.21	3.275	12.17	1.980	0.32

Table 3. The effect of different stereo matching network. Lower value means better performance.

As analyzed above, feature re-projection loss for real stereo images usually benefits from the well-trained translation networks by $\mathcal{L}_{fx}$. However, sometimes the value of feature re-projection loss for real stereo images may be low, but contrarily, both pixel distortion in translation and inaccurate estimation in stereo matching occur simultaneously. This correlation consistency loss could help only at this time.

Finally, D1-all results would drop a little bit without mode seeking loss. Because mode seeking loss actually provides fine-grained diversity to translated results and essentially helps stereo matching network learn a more robust disparity estimation network. In other words, stereo matching networks would learn to reduce the influence of various noise and lighting conditions during training.

Thanks to the integration of all the above four objectives described in Equation 9, we have obtained great improvement on filling the synthetic-to-real gap in stereo matching.

4.4. The Effect of Stereo Matching Network

In this part, we show how the structure of stereo matching network influences the performance of our proposed joint domain translation and stereo matching framework. We compare DispNet with one of the recently-proposed state-of-the-art stereo matching model GwcNet [11]. Their D1-all and EPE scores and inference time are reported in Table 3. As can be seen, GwcNet [11] performs far better than DispNet on both datasets and evaluation metrics. When using Synthia as our synthetic training data, our proposed model could help DispNet reduce D1-all and EPE by around 16.8%. It also makes GwcNet reduce D1-all by 23% and reduce EPE by 9.6%. For Driving training data whose domain gap to KITTI2015 is larger, our method could also help stereo matching network obtain very competitive performance. After trained with our proposed framework, D1-all is reduced by 51.5% and EPE 71% for DispNet respectively and D1-all is reduced by 56.8% and EPE by 39.6% for GwcNet respectively.

D1-all				
Test Train	KITTI2012		Cityscapes	
	Inference	Proposed	Inference	Proposed
Synthia	13.34	11.56	31.69	22.93
Driving	56.31	25.57	60.50	32.14
EPE				
Test Train	KITTI2012		Cityscapes	
	Inference	Proposed	Inference	Proposed
Synthia	2.121	1.936	11.805	6.701
Driving	11.669	2.832	15.468	8.506

Table 4. Generalization capability of our proposed method. We test our performance on two other real dataset: KITTI2012 and Cityscapes. Models are trained with only synthetic dataset and KITTI2015 dataset.

4.5. Generalization to Other Real Datasets

To demonstrate the generalization capability of stereo matching network trained in our joint optimization framework, we test their performance on other two real datasets - KITTI2012 [8] and Cityscapes [2], whose results are summarized in Table 4. Images in KITTI2012 have very similar domain style to those in KITTI2015 due to their similar camera setting. Therefore, the performance gain with the help of domain translation on KITTI2012 is similar to that on KITTI2015. For Cityscapes real dataset, both D1-all and EPE scores almost reduce by half. These significant improvements demonstrate great generalization capability of our proposed joint framework.

5. Conclusion and Future Work

In this paper we propose a novel end-to-end framework that trains domain translation networks and stereo matching network jointly. The newly-introduced stereo constraints including correlation consistency loss, bi-directional multi-scale feature re-projection loss and mode seeking loss regularize this joint framework to achieve better performance on stereo matching without ground-truth. The experimental results testify the effectiveness of our proposed framework in bridging the synthetic-to-real domain gap.

Our proposed framework successfully mitigated the gap between synthetic and real domain, yet there usually exist other gaps on intrinsics and disparity distribution between real-domain stereo images and translated-real stereo images, which is not explicit in our experimental datasets. Further study is also required to facilitate the generalization capability of our framework when meeting such datasets.

Acknowledgement. This work is supported in part by SenseTime Group Limited, and in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14213616, CUHK14207319, CUHK14208619, and in part by Research Impact Fund R5001-18.

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1, 3

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8

[3] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[4] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1180–1189, 2015. 3

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, pages 1–35, 2016. 3

[6] Ravi Garg, BG Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756, 2016. 5

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 6

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 2, 8

[9] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, pages 723–773, 2012. 3

[10] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, pages 723–773, 2012. 3

[11] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 1, 3, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 2

[13] Daniel Hernandez-Juarez, Lukas Schneider, Antonio Espinosa, David Vazquez, Antonio M. Lopez, Uwe Franke, Marc Pollefeys, and Juan Carlos Moure. Slanted stixels: Representing san franciscos steepest streets. In *British Machine Vision Conference (BMVC), 2017*, 2017. 1

[14] Heiko Hirschmller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 807–814, 2005. 2

[15] 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 *Proceedings of the 35th International Conference on Machine Learning*, pages 1989–1998, 2018. 2, 3

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017. 6

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28*, pages 2017–2025, 2015. 2, 4

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[19] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 66–75, 2017. 1, 3

[20] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[21] Andreas Klaus, Mario Sormann, and Konrad Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 15–18, 2006. 2

[22] Vladimir Kolmogorov and Ramin Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 508–515, 2001. 2

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1, 2

[24] Hsin-Ying Lee, , Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 2, 3

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems 30*, pages 700–708. 2017. 2, 3

[26] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jiaguang Sun, Yuchen Guo, and Philip S. Yu. Transfer sparse coding for robust image representation. In *Proceedings of the 2013*

IEEE Conference on Computer Vision and Pattern Recognition, pages 407–414, 2013. 3

[27] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5695–5703, 2016. 3

[28] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5

[29] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3, 4, 6, 7

[30] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *ICCV Workshops*, 2017. 1, 3

[31] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 6

[32] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[33] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, pages 7–42, 2002. 2

[34] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *Asian Conference on Computer Vision*, 2018. 3

[35] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. 3

[36] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450, 2016. 3

[37] Jian Sun, Nan-Ning Zheng, Heung-Yeung Shum, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2

[38] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised adaptation for deep stereo. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 3

[39] Alessio Tonioni, Oscar Rahnama, Tom Joy, Luigi Di Stefano, Ajanthan Thalaiyasingam, and Philip Torr. Learning to adapt for stereo. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 6, 7

[40] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 6, 7

[41] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[42] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 3

[43] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. SegStereo: Exploiting semantic information for disparity estimation. In *ECCV*, 2018. 3

[44] Lidong Yu, Yucheng Wang, Yuwei Wu, and Yunde Jia. Deep stereo matching with explicit cost aggregation subarchitecture. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3

[45] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1592–1599, 2015. 3

[46] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9788–9798, 2019. 5

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 3, 6, 7