Domain Generalized Stereo Matching via Hierarchical Visual Transformation

Tianyu Chang[1,3], Xun Yang[1*], Tianzhu Zhang[1], Meng Wang[2]
[1]University of Science and Technology of China [2]Hefei University of Technology
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
cty8998@mail.ustc.edu.cn {xyang21, tzzhang}@ustc.edu.cn wangmeng@hfut.edu.cn

Abstract

Recently, deep Stereo Matching (SM) networks have shown impressive performance and attracted increasing attention in computer vision. However, existing deep SM networks are prone to learn dataset-dependent shortcuts, which fail to generalize well on unseen realistic datasets. This paper takes a step towards training robust models for the domain generalized SM task, which mainly focuses on learning shortcut-invariant representation from synthetic data to alleviate the domain shifts. Specifically, we propose a Hierarchical Visual Transformation (HVT) network to 1) first transform the training sample hierarchically into new domains with diverse distributions from three levels: Global, Local, and Pixel, 2) then maximize the visual discrepancy between the source domain and new domains, and minimize the cross-domain feature inconsistency to capture domain-invariant features. In this way, we can prevent the model from exploiting the artifacts of synthetic stereo images as shortcut features, thereby estimating the disparity maps more effectively based on the learned robust and shortcut-invariant representation. We integrate our proposed HVT network with SOTA SM networks and evaluate its effectiveness on several public SM benchmark datasets. Extensive experiments clearly show that the HVT network can substantially enhance the performance of existing SM networks in synthetic-to-realistic domain generalization.

1. Introduction

Stereo Matching (SM) [7, 41, 44] aims to find the matching correspondences between a given stereo image pair and then calculate the disparity for depth sensing in many applications, such as robot navigation and autonomous driving [1, 28]. Recently, it attracts increasing attention in the computer vision community [4, 27, 30, 42].

With the development of deep learning [6, 18, 34–38], Convolutional Neural Network (CNN) based deep SM networks have shown impressive performance benefiting from their strong ability of feature representation. However, due to the scarcity of sufficient labeled realistic training data, existing state-of-the-art (SOTA) SM networks usually are trained on synthetic data, *e.g.* SceneFlow [19], which fail to generalize well to unseen realistic domains as shown in Fig. 1. Generally, the generalizability of cross-domain deep SM networks is mainly hindered by a critical issue: SM networks usually learn superficial shortcut features [5] from synthetic data to estimate the disparity. Specifically, such shortcut features mainly include two types of artifacts: consistent local RGB color statistics and overreliance on local chromaticity features, which are domain-sensitive and non-transferable to unseen domain. The semantic and structural features that are truly desirable are ignored by most existing SM networks. Therefore, the key to addressing the challenging cross-domain SM task is how to effectively learn the domain-invariant representations of the given stereo image pair for synthetic-to-realistic generalization.

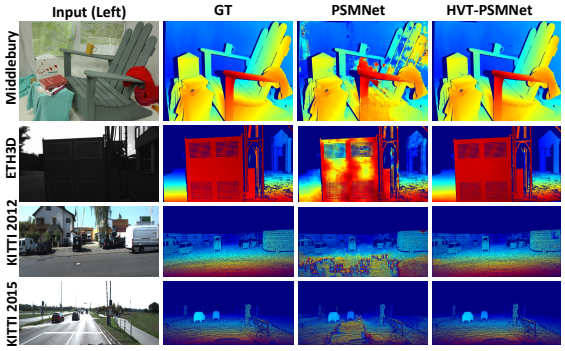

Figure 1. Comparison of the cross-domain SM generalization. Columns from left to right denote a sample image, ground truth disparities, the predicted disparities of the pretrained PSMNet model and our HVT-PSMNet model. Both models are trained on the synthetic SceneFlow [19] dataset and evaluated on the realistic datasets: Middlebury, ETH3D, KITTI 2012 and KITTI 2015.

Several attempts [3, 10, 15, 26, 45] have been made to minimize the synthetic-to-realistic domain gap and learn the domain-invariant representations for the SM task by either 1) exploiting labeled target-domain realistic data to fine-tune the SM network trained with synthetic data [3, 10]

*Corresponding Author

or 2) jointly using the synthetic data and unlabeled target-domain realistic data to train domain adaptive SM networks [15,26,45]. Despite their performance improvement on realistic data, these attempts only work well when the target-domain realistic data is provided during training and thus can not improve the out-of-distribution (OOD) generalization of SM networks, which are less practically useful in real-world scenarios.

In this work, we address the important but less explored challenging problem of single domain generalization in SM, where only the synthetic data is available for training. Considering the fact that most existing SM networks are susceptible to exploiting shortcut cues in synthetic data instead of the semantic and structural correspondences, we propose to learn shortcut-invariant robust representation from synthetic SM image pairs for OOD generalization. Specifically, this paper presents a **H**ierarchical **V**isual **T**ransformation (HVT) network to 1) first transform the synthetic training sample hierarchically into new source domains with diverse distributions from three levels: Global, Local, and Pixel, 2) then maximize the image discrepancy between the synthetic source domain and new domains for significantly altering the original distribution, and minimize the cross-domain feature inconsistency to capture domain-invariant features. In this way, we are able to prevent the model from exploiting the artifacts of synthetic stereo images as shortcut features, thereby estimating the disparity maps more effectively based on the learned shortcut-invariant feature representation. Our basic idea is to diversify the distribution of training data and thus force the network to overlook the artifacts from synthetic domain. Note that our proposed HVT network is simple and can be plug-and-play. We integrate HVT with SOTA SM networks during training and evaluate its effectiveness on several challenging SM benchmark datasets. Extensive experiments clearly show that the HVT network can substantially enhance the performance of existing SM networks in synthetic-to-realistic domain generalization without using any auxiliary data or features [17].

Our contributions can be briefly summarized as follows:

- We devise a simple yet effective domain generalized SM framework. It leverages a hierarchical visual transformation network to effectively diversify the distribution of training data which prevents the model from exploiting the artifacts in synthetic data as shortcuts.
- We formulate novel learning objectives that force the model to effectively optimize three complementary visual transformations by maximizing domain discrepancy and minimizing feature inconsistency between synthetic domain and new domains, thereby facilitating the learning of domain-invariant feature representation.
- Extensive experiments on four realistic SM datasets clearly demonstrate the effectiveness and robustness of our HVT network. The out-of-distribution generalization ability of four SOTA SM methods has been significantly boosted, benefiting from our solution.

2. Related Work

This section briefly introduces the learning based stereo matching (SM) methods from two perspectives: 1) In-Distribution SM and 2) Our-of-Distribution SM.

In-Distribution SM. In the past decade, deep learning has triggered the fast development of the task of SM. Mayer et al. [19] introduced an end-to-end SM network which uses the correlation layer to generate the cost volume and the 2D-CNN to aggregate the cost volume. SegStereo [33] and EdgeStereo [27] exploit the semantic and edge cues to help the disparity prediction. To learn better features, AANet [30] integrates both normal convolution and deformable convolution for feature extraction. Furthermore, GCNet [14] concatenates left and right features for cost volume generation and aggregation with 3D convolutions. GwcNet [10] proposes the group-wise correlation to construct the cost volume with lower memory cost. PSMNet [3] proposes the spatial pyramid pooling module and the stack hourglass network to expand the receptive field. GANet [42] designs the image content guided layers to globally and locally update the cost volume. LEAStereo [4] introduces the neural architecture searching into SM. In recent several years, the Cascade-based SM methods [9, 25, 39] first use the coarsest resolution feature maps to predict an initial disparity, and narrow down the disparity search range to refine the disparity based on the initial disparity. Despite the good performance of existing deep-learning based SM networks, their training set and testing set always follow the independent and identically distributed (IID) assumption. Most existing works usually fail to generalize well on unseen realistic data.

Out-of-Distribution SM. Most recently, the synthetic-to-realistic generalized SM networks [5, 16, 25, 43] have received increasing attention. Different from those fine-tuning or adaptation-based methods [3, 15, 26, 42, 45], they aim to directly learn SM networks that can generalize to unseen domains. Some works propose to improve the generalization ability by devising novel network architectures. For example, DSMNet [43] proposes the domain normalization layer and the trainable non-local graph-based filter to learn domain-invariant structural feature. Besides, both CFNet [25] and RAFT-Stereo [16] introduce a network architecture with multi-scale cost volume fusion and aggregation to learn robust semantic or structural features. Some other efforts attempt to design flexible modules that can facilitate the cross-domain generalization of more existing SM networks. MS-Net [2] points out that learned feature is the core reason for the poor generalization performance, and proposes to use the traditional feature descriptors to construct the cost. FC-Net [46] uses the stereo contrastive

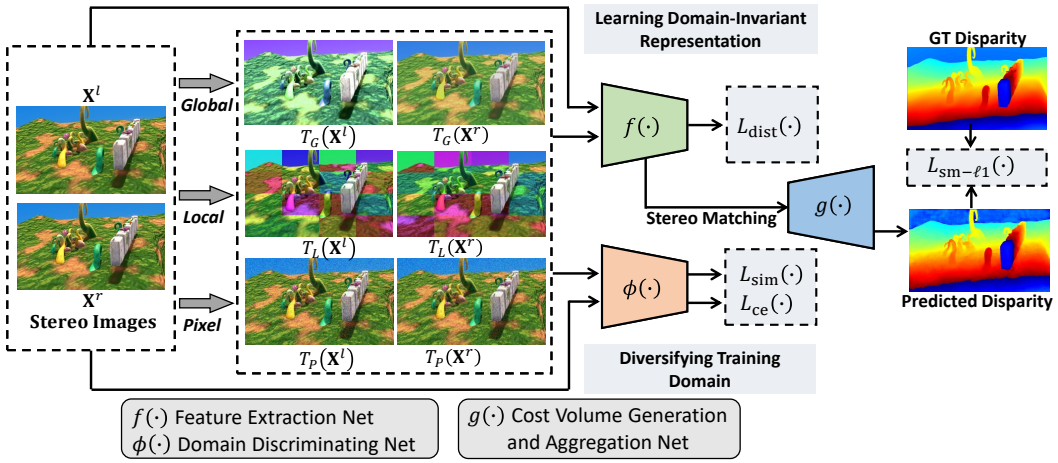

Figure 2. The pipeline of our domain-generalized SM approach. It leverages a hierarchical visual transformation module to improve the diversity of training domain by training a domain discriminator $\phi(\cdot)$. Its main goal is to optimize a feature extractor $f(\cdot)$ that can learn shortcut-robust and domain-invariant representation, thereby facilitating the generalization of SM network in unseen realistic domains.

feature loss and the stereo selective whitening loss to encourage the stereo feature consistency across different domains. GraftNet [17] leverages the robust broad-spectrum feature which is trained on large-scale datasets to replace the original feature to improve the generalization ability. ITSA [5] generates the pixel-perturbed image through the gradient of the features with respect to the input image, and guides the network to learn shortcut-invariant feature.

Different from these methods, we introduce a simple yet effective method to enhance the synthetic-to-realistic generalization. We design a hierarchical visual transformation module to diversify the training domain by maximizing the cross-domain visual discrepancy and propose to learn domain-invariant feature representation by minimizing the cross-domain feature inconsistency. ITSA [5] can be generally seen as a special case of our work that diversifies the distribution in the *Pixel* level. Experiments in the Sec. 4. have clearly validated the effectiveness of our method on improving the synthetic-to-realistic generalization of SM.

3. The Approach

3.1. Problem Overview

This work aims to improve the synthetic-to-realistic domain generalization for SM. Given a synthetic training set $\mathcal{D}_s$ as input, consisting of $|\mathcal{D}_s|$ synthetic stereo image pairs $\{\mathbf{X}_i^l, \mathbf{X}_i^r\}_{i=1}^{|\mathcal{D}_s|}$ and the corresponding ground-truth disparity maps $\{\mathbf{Y}_i^{gt}\}_{i=1}^{|\mathcal{D}_s|}$, the goal is to learn a cross-domain SM network that can effectively predict the disparity between a pair of stereo images from unseen domains $\mathcal{D}_r$. A typical SM network $F_\Theta(\cdot,\cdot)$ can be formally formulated as:

$$\hat{\mathbf{Y}} = F_\Theta(\mathbf{X}^l, \mathbf{X}^r) = s\big(g\big(f(\mathbf{X}^l), f(\mathbf{X}^r)\big)\big), \qquad (1)$$

where Θ denotes the total network parameters and $f(\cdot)$ is a feature extraction module that yields the feature map of stereo images. $g(\cdot)$ is a joint network module that first generates the cost volume by correlation or concatenation strategy, and then performs the cost aggregation and refinement. The final disparity map $\hat{\mathbf{Y}}$ is estimated by converting the refined cost volumes with the soft-argmin [14] operation $s(\cdot)$. The typical SM network $F_\Theta(\cdot,\cdot)$ is optimized by minimizing the smooth-ℓ_1 loss $L_{\text{sm-}\ell_1}\left(F_\Theta(\mathbf{X}^l, \mathbf{X}^r), \mathbf{Y}^{gt}\right)$ [3].

In the past years, existing efforts primarily focus on how to devise an effective sub-network $g(\cdot)$ for cost volume generation and cost aggregation, while overlook the importance of learning robust feature representations of stereo images that can preserve the semantic and structural cues. Recent studies pointed out that most existing SM networks trained on synthetic data are susceptible to learn superficial shortcut features [5] instead of the desirable semantic and structural features, which leads to poor generalization performance in unseen (realistic) domain due to the significant synthetic-realistic domain gap. Therefore, the main research problem in this work is *how to train an effective feature extraction network $f(\cdot)$ on only synthetic data that can learn generalizable representation of stereo images, so as to estimate reliable disparity map on unseen domain*.

3.2. Our Proposed Method

Our target is to tackle the synthetic-to-realistic generalization in SM. Its main challenge is that we only have synthetic data for training, which has the high risk of overfitting to synthetic data due to its limited diversity. Inspired by [13, 29, 40, 47], we address the domain generalized SM from a new perspective in this work. The rest of this section will introduce our proposed network for this task in detail. The pipeline of our proposed HVT is shown in Fig. 2

3.2.1 Hierarchical Visual Transformation

As we know, the key to addressing domain generalization is learning domain-invariant feature, also called *causal fea-*

ture. Although the explicit form of causal feature is unknown in general, we have the prior that causal feature should remain invariant to certain transformations [29]. For example, the semantic and structural features can be treated as two domain-invariant (causal) features of stereo images for SM, while adjusting the *luminance*, *contrast*, or *saturation* of a stereo image pair will not affect the corresponding disparity map. Our intuitive idea is to leverage the visual transformations that do not change the underlying domain-invariant feature to increase the diversity of training domain, thereby enhancing the generalization performance of SM network. It is a straightforward but effective way. To this end, we design a **Hierarchical Visual Transformation** (HVT) network to diversify the distribution of training domain. Specifically, we learn a set of visual transformations $\mathcal{T} = \{T_1, \cdots, T_M\}$ at different levels to transform the original stereo image pairs as $\big(T(\mathbf{X}^l), T(\mathbf{X}^r)\big)$, where $T \in \mathcal{T}$. We force the visual transformation $T(\cdot)$ to meet the following three requirements:

- $T(\cdot)$ should induce large visual discrepancy between the original stereo image and the transformed images to enlarge the diversity of training domain.
- $T(\cdot)$ should not change the target disparity map of the original stereo image pairs. We should minimize the smooth-ℓ_1 loss $L_{\text{sm-}\ell_1}(F_\Theta(T(\mathbf{X}^l), T(\mathbf{X}^r)), \mathbf{Y}^{gt})$ where the transformed stereo images are fed into.
- The representation of the transformed stereo images $f\left(T(\mathbf{X})\right)$ should be consistent with that of the original images $f\left(\mathbf{X}\right)$, so as to learn domain-invariant features.

Following the above-listed requirements, we describe the implementation of our proposed HVT network as follows.

Implementation of HVT. To effectively learn domain-invariant feature representation for the challenging SM task, we transform the stereo image from three complementary perspectives: *Global*, *Local*, and *Pixel*.

(1) Global: The *Global* visual transformation $T_G(\cdot)$ aims to globally change the distribution of stereo images by sequentially adjusting the basic visual attributes of images: *Brightness*, *Contrast*, *Saturation*, and *Hue* based on the four corresponding sub-transformations $\{T_G^B, T_G^C, T_G^S, T_G^H\}$, following the pipeline of automatic image editing [12]. The three sub-transformations $\{T_G^B, T_G^C, T_G^S\}$ can be formulated as:

$$T_G^I(\mathbf{X}) = \alpha_G^I \mathbf{X} + (1 - \alpha_G^I) o^I(\mathbf{X}), \tag{2}$$

where $I \in \{B, G, S\}$. The α_G^I is a randomly selected positive constant in an adjustable range of $[\tau_{\text{min}}^I, \tau_{\text{max}}^I]$,

$$\begin{cases} \tau_{\text{min}}^I = 1 - \big(\mu\sigma(\varrho_l^I) + \beta\big) \\ \tau_{\text{max}}^I = 1 + \big(\mu\sigma(\varrho_h^I) + \beta\big) \end{cases}, \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\varrho_l^I \in \mathbb{R}^1$ and $\varrho_h^I \in \mathbb{R}^1$ are two learnable parameters, μ and β are two positive hyper-parameters. The operation $o^I(\cdot)$ in Eq. (2) is customized as follows. Specifically, for the *Brightness*, $o^B(\mathbf{X}) = \mathbf{X} \cdot \mathbf{O}$ where $\mathbf{O}$ denotes a zero matrix all of whose entries are zero. For the *Contrast*, $o^C(\mathbf{X}) = \text{Avg}(\text{Gray}(\mathbf{X}))$ where $\text{Gray}(\cdot)$ transforms the RGB image into a gray-scale image and $\text{Avg}(\cdot)$ computes the mean value of all pixels. For the *Saturation*, $o^S(\mathbf{X}) = \text{Gray}(\mathbf{X})$. The *Hue* adjustment is defined as:

$$T_G^H(\mathbf{X}) = \text{Rgb}([\mathbf{h} + \alpha_G^H, \mathbf{s}, \mathbf{v}]), \tag{4}$$

where $[\mathbf{h}, \mathbf{s}, \mathbf{v}] = \text{Hsv}(\mathbf{X})$ denotes the three components of $\mathbf{X}$ in the HSV color space and $\mathbf{h}$ denotes the *Hue* component. $\text{Rgb}(\cdot)$ transforms the image from the HSV space into the RGB space. The $\alpha_G^H \in \mathbb{R}^1$ is randomly selected from an adjustable range of $[\tau_{\text{min}}^H, \tau_{\text{max}}^H]$, where $\tau_{\text{min}}^H = -\mu\sigma(\varrho_l^H) - \beta$ and $\tau_{\text{max}}^H = \mu\sigma(\varrho_h^H) + \beta$.

Note that the order of the sequential sub-transformations $\{T_G^B, T_G^C, T_G^S, T_G^H\}$ is random to improve the diversity.

(2) Local: The *Local* visual transformation $T_L(\cdot)$ aims to locally modify the distribution of stereo training images. The basic pipeline is that we first slice the given stereo image into $N' \times N'$ non-overlapping patches $\{\mathbf{x}_1^p, \cdots, \mathbf{x}_{N' \times N'}^p\}$, then transform each image patch $\mathbf{x}_i^p$ by a global transformation $T_L^p(\cdot)$, and finally merge these transformed image patches into the stereo image based on the original sequential order. The output can be described by

$$T_L(\mathbf{X}) = \text{Merge}\left([T_L^p(\mathbf{x}_1^p), \cdots, T_L^p(\mathbf{x}_{N' \times N'}^p)]\right), \tag{5}$$

where $\text{Merge}(\cdot)$ denotes the *Merge* operation.

Note that patch-level transformation $T_L^p(\cdot)$ for each patch doesn't share parameters with each other to improve the diversity. $T_L^p(\cdot)$ can be implemented by existing style transfer networks, *e.g.* Adain [22] or Fourier-based method [31]. To better complementing our *Global* transformation, patch-level transformation $T_L^p(\cdot)$ is implemented by $T_G(\cdot)$ in Eq.(2) and Eq.(4) for its simplicity and effectiveness.

(3) Pixel: The *Pixel* visual transformation $T_P(\cdot)$ aims to perturb the given stereo images with a pixel-level perturbation matrix. It can be formulated as:

$$T_P(\mathbf{X}) = \mathbf{X} + \big(\mu\sigma(\mathbf{W}) + \beta\big)\mathbf{P} \tag{6}$$

where $\mathbf{P} \in \mathbb{R}^{H \times W \times 3}$ is a randomly generated Gaussian matrix with the mean of 0 and the standard deviation of 1, and $\mathbf{W} \in \mathbb{R}^{H \times W \times 3}$ is a learnable matrix. Different from the *Global* and *Local* transformations, $T_P(\cdot)$ can alter the distribution of synthetic data in a more granular fashion.

Remark: The implementation of HVT is mainly motivated by the empirically found evidence [5] that existing SM networks are susceptible to exploit common artifacts (*e.g.* consistent local RGB color statistics and overreliance on local chromaticity features) of synthetic stereo images as shortcuts, which can be clearly illustrated by Fig. 3. We observe from Fig. 3 that the PSMNet is quite sensitive to the different levels of perturbation. It reflects that PSM-

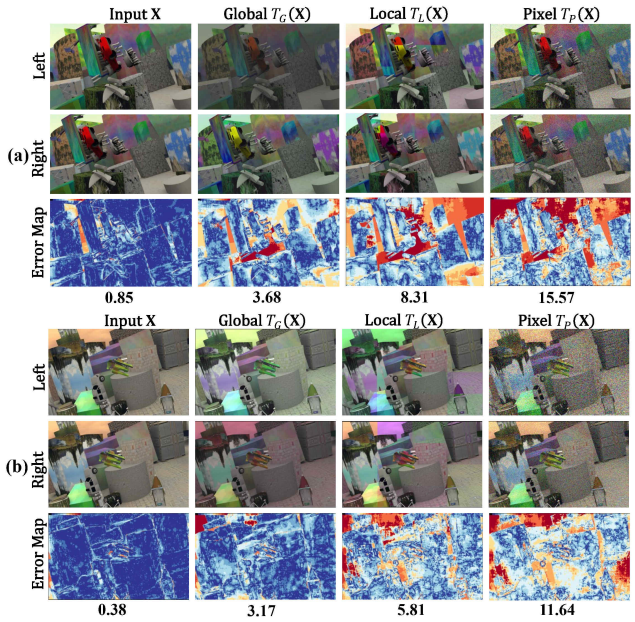

Figure 3. Examples of visualized output of three visual transformations (top-2 rows of (a) and (b)) and the corresponding disparity error maps $|\hat{\mathbf{Y}} - \mathbf{Y}^{gt}|$ (the 3-rd row of (a) and (b)) of the PSMNet trained on the synthetic SceneFlow dataset. The EPE values are marked in the bottom of the 3-rd row for comparison.

Net doesn't capture the robust semantic and structural features of stereo images. Our method is specially devised to improve the robustness and generalization ability of existing SM method. The design of the three different transformations $\{T_G(\mathbf{X}), T_L(\mathbf{X}), T_P(\mathbf{X})\}$ can effectively diversify the training domain and prevent the shortcuts from being encoded into the stereo image representation $f(\mathbf{X})$.

3.2.2 Learning Objectives

To effectively train our proposed HVT network for the domain generalized SM task, we introduce the following loss terms for network optimization.

Maximizing Cross-Domain Visual Discrepancy: In this work, without having access to the target domain, we propose to improve the SM network's generalization by transforming the synthetic source domain data into several new source domains $\{T_G(\mathbf{X}), T_L(\mathbf{X}), T_P(\mathbf{X})\}$ to diversify the training domain. Our first objective is to force the distributions of new source domains to be dissimilar as possible to the original distribution of synthetic data. Then we should maximize the following cross-domain visual discrepancy as

$$\max L_{\text{disc}}(\mathbf{X}) = \frac{1}{3}\sum_{J} d(T_J(\mathbf{X}), \mathbf{X}) \tag{7}$$

where $J \in \{G, L, P\}$, and $d(\cdot)$ is a domain discrepancy measure. In this work, we introduce a specific feature extraction network $\phi(\cdot)$ for domain discriminating. Then we implement Eq. (7) by minimizing the domain similarity as

$$\min L_{\text{sim}}(\mathbf{X}) = \frac{1}{3}\sum_{J} \text{Cos}\left(\phi(T_J(\mathbf{X})), \phi(\mathbf{X})\right), \tag{8}$$

where $\text{Cos}(\cdot,\cdot)$ denotes the cosine similarity function. $\phi(\mathbf{X})$ denotes the pooled feature vector from the domain discriminating network $\phi(\cdot)$. Besides, to further improve the domain discrepancy, we also minimize the following cross-entropy loss for domain classification:

$$\min L_{\text{ce}}(\mathbf{X}) = \text{CE}\left(\{\phi(T_J(\mathbf{X})), \phi(\mathbf{X})\}, \mathcal{Y}_d\right), \tag{9}$$

where $\text{CE}(\cdot,\cdot)$ denotes the standard cross-entropy loss and $\mathcal{Y}_d$ denotes the domain labels of four source domains.

Minimizing Cross-Domain Feature Inconsistency: To enhance the model's generalization ability, we need to learn domain-invariant representation of stereo images. We expect that our visual transformation $T(\cdot)$ doesn't change the core features (*e.g.*, semantic and structural features) of images for stereo matching. Therefore, our second objective is to force the feature representation of transformed stereo images to be consistent with that of original synthetic images. Then we minimize the following pairwise distance term:

$$\min L_{\text{dist}}(\mathbf{X}) = \frac{1}{3}\sum_{J} \left\| f\left(T_J(\mathbf{X})\right) - f\left(\mathbf{X}\right)\right\|_2, \tag{10}$$

which facilitates the learning of shortcut-robust and domain invariant features for domain-generalized stereo matching. In summary, our model is trained by minimizing the linear combination of the above-mentioned learning objectives:

$$\begin{aligned}\min \mathcal{L} =& L_{\text{sm-}\ell_1}(\hat{\mathbf{Y}}, \mathbf{Y}^{gt}) + \\ & \frac{1}{2}\left(\lambda_1 L_{\text{dist}}(\mathbf{X}) + \lambda_2 L_{\text{sim}}(\mathbf{X}) + \lambda_3 L_{\text{ce}}(\mathbf{X})\right),\end{aligned} \tag{11}$$

The full loss $\mathcal{L}$ is computed as the average over a training batch. The smooth-ℓ_1 loss $L_{\text{sm-}\ell_1}$ is computed based on not only the original synthetic image pairs $\{\mathbf{X}^l, \mathbf{X}^r\}$ but also the transformed image pairs $\{T_J(\mathbf{X}^l), T_J(\mathbf{X}^r)\}$. λ_1, λ_2, and λ_3 are three trade-off hyper-parameters.

4. Experiments

In this section, we quantitatively and qualitatively conduct extensive experiments to answer the following research questions: **R1:** Can the HVT effectively improve the generalization performance of existing SM networks? Can our performance reach the SOTA level? **R2:** Do the three transformations in the HVT effectively complement to each other? **R3:** Can our approach learn the domain-invariant features? **R4:** How is the robustness of our approach when tested in complex realistic scenarios?

4.1. Dataset and Experimental Setting

Datasets. We use the SceneFlow [19] dataset for training and four realistic SM datasets (KITTI 2012 [8], KITTI 2015 [20], Middlebury [23] and ETH3D [24]) for e-

Baselines	Methods	KITTI 2015		KITTI 2012		Middlebury(H)		ETH3D		References
		EPE	D1(3px)	EPE	D1(3px)	EPE	D1(2px)	EPE	D1(1px)	
-	GANet [42]	2.31	11.7	1.93	10.1	5.41	20.3	1.33	14.1	CVPR 2019
	CasStereo [9]	2.42	11.9	2.12	11.8	3.71	17.2	0.87	7.8	CVPR 2020
	DSMNet [43]	1.46	6.5	1.26	6.2	2.62	13.8	0.69	6.2	ECCV 2020
PSMNet [3]	PSMNet [3]	3.17	16.3	2.69	15.1	7.65	34.2	2.33	23.8	CVPR 2018
	MS-PSMNet [2]	1.64*	7.8	2.33*	14.0	4.72*	19.8	1.42*	16.8	3DV 2020
	FC-PSMNet [46]	1.58*	7.5	1.42*	7.0	4.14*	18.3	1.25*	12.8	CVPR 2022
	ITSA-PSMNet [5]	1.39*	5.8	1.09*	5.2	3.25*	12.7	0.94*	9.8	CVPR 2022
	Graft-PSMNet [17]	1.32	5.3	1.09	5.0	2.34	10.9	1.16	10.7	CVPR 2022
	HVT-PSMNet	**1.14±0.02**	**4.9±0.12**	**0.93±0.02**	**4.3±0.06**	**1.46±0.13**	**10.2±0.16**	**0.47±0.03**	**6.9±0.23**	Ours
GwcNet [10]	GwcNet [10]	3.43	22.7	2.77	20.2	7.23	37.9	2.78	54.2	CVPR 2019
	FC-GwcNet [46]	1.72*	8.0	1.45*	7.4	5.14*	21.1	1.13*	11.7	CVPR 2022
	ITSA-GwcNet [5]	1.33*	5.4	1.02*	4.9	2.73*	11.4	0.62*	7.1	CVPR 2022
	HVT-GwcNet	**1.15±0.02**	**5.0±0.11**	**0.88±0.02**	**3.9±0.13**	**1.29±0.13**	**10.3±0.21**	**0.46±0.08**	**5.9±0.26**	Ours
CFNet [25]	CFNet [25]	1.71	6.0	1.04	5.2	3.24	15.4	0.48	5.72	CVPR 2021
	ITSA-CFNet [5]	**1.09**	**4.7**	0.87	4.2	1.87	10.4	0.45	5.1	CVPR 2022
	HVT-CFNet	1.10±0.04	4.9±0.16	**0.85±0.02**	**4.0±0.14**	**1.79±0.22**	**10.2±0.16**	**0.39±0.02**	**4.5±0.24**	Ours
RAFT [16]	RAFT [16]	1.26	5.7	1.01	5.1	1.92	12.6	0.36	3.3	3DV 2021
	HVT-RAFT	**1.12±0.02**	**5.2±0.09**	**0.87±0.02**	**3.7±0.08**	**1.37±0.11**	**10.4±0.14**	**0.29±0.01**	**3.0±0.09**	Ours

Table 1. Performance comparison with SOTA domain generalized SM networks. The * denotes our reproduced EPE results since the authors only use the D1 metric. As we didn't have a validation set for model selection, we report the average result over last 5 epochs.

valuation. SceneFlow [19] is a large-scale synthetic SM dataset containing 35,454 training stereo image pairs and 4,370 testing image pairs with a resolution of 960×540. All the synthetic image pairs have dense annotations of ground-truth disparities. KITTI 2012 [8] and KITTI 2015 [20] consists of 194 and 200 training stereo image pairs collected in outdoor driving scenes with sparse annotations of disparities, respectively. Middlebury [23] includes 15 high-resolution stereo image pairs in indoor scenes. In our experiments, we use the half-resolution version (Middlebury(H)) for evaluation. ETH3D [24] has 27 grayscale stereo image pairs collected from both indoor and outdoor scenes.

Metrics. Following the setting of [17], we use both **EPE** rate (End-Point Error) and **D1** error rate (%) with different pixel threshold ρ as the metrics to get a more comprehensive evaluation. The EPE measures the average disparity error over all pixels, which reflects the average disparity estimate. The D1 error rate computes the percentage of error pixels with absolute error larger than a specific threshold ρ. The pixel threshold is set as ρ=3px for KITTI 2012 [8] and KITTI 2015 [20], ρ=2px for Middlebury [23] and ρ=1px for ETH3D [24], respectively, following recent works [17, 46].

Baselines. We select four SM networks as baselines, including two well-studied and commonly-used methods (PSMNet [3] and GwcNet [10]) and two SOTA methods (CFNet [25] and RAFT [16]), in our experiments. We will integrate our HVT networks with these baseline networks to validate the effectiveness of HVT on domain generalization.

Implementation Details. Following [43], we use the domain normalization technique in the feature extraction module for all baselines. Except the RAFT [16], all the baselines are implemented by PyTorch [21] with Adam (β_1 = 0.9, β_2 = 0.999) optimizer. We train the model for 45 epochs with the batch size of 8. Besides, we introduce an incremental training scheme to stably optimize the network in this work. Specifically, we train the model with $T_G(\cdot)$ only for 15 epochs at Stage I, then we further train the model with $\{T_G(\cdot), T_L(\cdot)\}$ for another 15 epochs at Stage II, and finally we continually train the model with $\{T_G(\cdot), T_L(\cdot), T_P(\cdot)\}$ for the remaining 15 epochs at Stage III. The learning rate is set to 0.001 and decreased by half after epoch 10, 20, 30 and 40. For the RAFT [16] baseline, we train the model for total 30K steps and train each stage 10K steps. The optimizer and learning rate follow its original setting. Besides, the asymmetric chromatic augmentation is not used in any models to avoid unfair experimental comparison. We select ResNet18 [11] as the domain discriminating network $\phi(\cdot)$ for its simplicity. For different hyper-parameters, the μ and β in Eq. (3) and (6) are set as 0.1 and 0.15 respectively. The λ_1, λ_2 and λ_3 in Eq. (11) are set as 1, 0.5 and 0.5 respectively. The $N' = 4$ in $T_L(\cdot)$. *Note that all the SM networks are only trained on the synthetic dataset, SceneFlow. The target data is strictly inaccessible during training.*

4.2. Overall Performance Comparison

R1: Comparison with Baselines and SOTAs. In Tab. 1, we report the comparison of generalization performance between the baselines (PSMNet, GwcNet, CFNet, and RAFT) and our HVT-integrated methods, as well as the comparison between ours and the reported SOTA results, across the four realistic datasets. We have the following observations:

- The synthetic-to-realistic generalization performances of all the baselines are consistently improved by our HVT in all settings. Specifically, compared with the widely-used PSMNet [3], the performance improvements *w.r.t.* the D1 metric are 11.4%, 10.8%, 24%, and 16.9%, respectively on the four datasets. Similar improvement can also be observed on GwcNet [10]: 17.7%, 16.3%, 27.6%, and

Baselines	*Global*	*Local*	*Pixel*	KITTI 2015 EPE	KITTI 2015 3px	Middlebury EPE	Middlebury 2px	ETH3D EPE	ETH3D 1px
PSMNet [3]	✓	✗	✗	1.31	6.07	2.04	13.7	0.48	9.4
	✗	✓	✗	1.28	6.03	1.69	12.5	0.59	10.7
	✗	✗	✓	1.27	6.29	2.39	14.6	0.57	9.9
	✓	✓	✗	1.23	5.47	1.57	11.1	0.53	9.1
	✓	✗	✓	1.19	5.57	**1.44**	11.2	0.51	7.7
	✗	✓	✓	1.27	5.72	1.58	10.9	0.62	7.9
	✓	✓	✓	**1.14**	**4.93**	1.46	**10.2**	**0.47**	**6.9**
GwcNet [10]	✓	✗	✗	1.28	5.89	1.88	12.6	0.74	8.2
	✗	✓	✗	1.26	5.88	1.64	11.9	0.68	8.5
	✗	✗	✓	1.29	6.18	1.83	12.4	0.71	8.3
	✓	✓	✗	1.23	5.38	1.58	10.4	0.58	7.8
	✓	✗	✓	1.19	5.52	1.49	10.7	0.55	7.1
	✗	✓	✓	1.18	5.47	1.57	10.6	0.52	6.9
	✓	✓	✓	**1.15**	**5.02**	**1.29**	**10.3**	**0.46**	**5.9**

Table 2. Ablation studies on the effect of three visual transformations (*Global*, *Local*, and *Pixel*) on three datasets using the two well-studied baselines, PSMNet [3] and GwcNet [10].

48.3% *w.r.t.* the D1 metric. Even compared with the SOTA CFNet [25] and RAFT [16], the D1 error rates are still substantially decreased: 1.1%, 1.2%, 5.2%, and 1.2% for CFNet, and 0.5%, 1.4%, 2.2%, and 0.3% for RAFT. The quantitative comparison clearly shows the effectiveness of HVT on enhancing the robustness of SM models. The improvement can be mainly attributed to the fact that our HVT significantly diversify the visual distribution of original synthetic images, thus preventing the model from building the spurious relationship between the pairwise input and the target disparity. The model will not easily find the matching correspondences across the stereo images by just leveraging the shortcut cues, *e.g.*, local RGB color statistics and chromaticity features.

- The improvements of generalization performance brought by HVT on the Middlebury and ETH3D datasets seem to be much larger that those on the KITTI 2012 and 2015 datasets. The reason is that Middlebury has higher resolution images in realistic scenarios than those in KITTI 2012 and 2015, and the images in ETH3D are all grayscale which are very different from the colorful synthetic training images. It reflects that our HVT can perform well in diverse realistic scenarios that differ from training domain in resolution and color distribution.
- Our HVT-enhanced methods almost outperform all the SOTA methods except ITSA-CFNet on KITTI 2015. It shows the strong potential of our HVT in improving the cross-domain generalization. Note that the three transformations devised in this work are all simple and straightforward. The performance can be further improved if we introduce more delicate transformations into our HVT. Besides, the performance of ITSA-CFNet on KITTI 2015 is already very high *w.r.t.* the EPE and D1 metrics, 1.09 and 4.7%. Our HVT-CFNet's performance is on a par with that of the ITSA-CFNet.
- Benefiting from using HVT, the performances of weak baselines, *i.e.*, PSMNet and GwcNet, have been increased to the SOTA level, especially on the top-three datasets. It reflects that the critical step for the domain-generalized SM task is to learn domain-invariant representation from diverse training domain instead of devising different complex network modules for cost volume generation or aggregation. It further validates the rationale of our HVT.

Baselines	Obj-1	Obj-2	KITTI 2015 EPE	KITTI 2015 3px	Middlebury EPE	Middlebury 2px	ETH3D EPE	ETH3D 2px
PSMNet [3]	✗	✗	1.67	7.74	2.88	15.1	0.89	13.2
	✗	✓	1.31	6.23	2.04	11.7	0.63	10.1
	✓	✓	**1.14**	**4.93**	**1.46**	**10.2**	**0.47**	**6.9**
GwcNet [10]	✗	✗	1.64	7.58	2.65	14.6	0.86	12.4
	✗	✓	1.28	6.07	1.93	11.3	0.57	9.2
	✓	✓	**1.15**	**5.02**	**1.29**	**10.3**	**0.46**	**5.9**

Table 3. Ablation studies of our two main optimization objectives: Obj-1 (*i.e.*, Eq. (8) and Eq. (9)) and Obj-2 (*i.e.*, Eq. (10)).

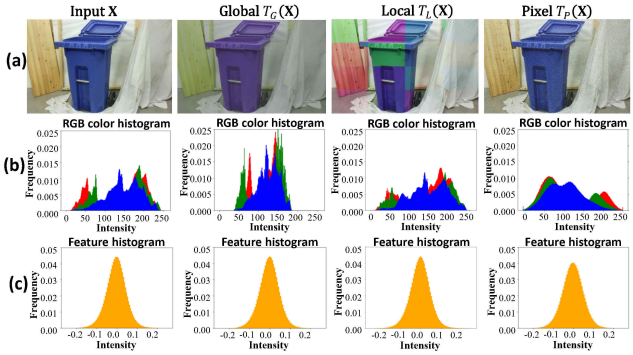

Figure 4. Visualization of RGB color histograms (see (b)) and visual feature histograms (see (c)) of original image and transformed images (see (a)) based on the HVT-PSMNet method. The original image is randomly selected from the realistic Middlebury dataset.

4.3. Ablation Studies

R2: The Complementation of HVTs. As shown in Tab. 2, we investigate the effect of the HVT module by gradually adding the visual transformations in two baselines, PSMNet [3] and GwcNet [10]. We observe that the three visual transformations can complement well to each other. Specifically, only using one of the three transformations can also facilitate the performance improvement of baselines. When using more visual transformations, we observe consistent performance gains across the three realistic datasets. For example, the D1 error rates of $T_G(\cdot)$-enhanced PSMNet have been decreased from 6.07%, 13.7%, and 9.4% to 5.57%, 11.2%, and 7.7%, respectively, when jointly using $T_G(\cdot)$ and $T_P(\cdot)$. The best performances are observed when all the three transformations are jointly used. Similar performance improvements can also be observed when Gwc-

Methods	Sunny	Cloudy	Rainy	Foggy	Avg.
PSMNet [3]	62.5	60.1	60.5	68.6	63.9
FT-PSMNet [5]	**4.0**	**2.9**	11.5	6.5	6.3
FC-PSMNet [46]	4.9	4.3	**7.2**	6.2	5.7
ITSA-PSMNet [5]	4.8	3.2	9.4	6.3	5.9
HVT-PSMNet	4.2	3.1	8.7	**5.6**	**5.4**
GwcNet [10]	18.1	24.7	28.2	28.3	24.8
FT-GwcNet [5]	**3.1**	**2.5**	12.3	6.0	6.0
ITSA-GwcNet [5]	4.4	3.3	9.8	5.9	5.9
HVT-GwcNet	3.4	3.5	**8.6**	**5.6**	**5.3**

Table 4. Robustness comparison of different methods on the DrivingStereo [32] dataset collected from complex realistic scenarios: *Sunny*, *Cloudy*, *Rainy*, and *Foggy*. The D1 (3px) metric is used.

Net is used as the SM baseline.

R3: Learning Domain-Invariant features. The core of our approach is to learn shortcut-robust features that can generalize well to unseen domains. As illustrated in Fig. 4 (b), we observe that the hierarchical visual transformations significantly alter the RGB color histograms of original stereo image in the levels of *Global*, *Local*, and *Pixel*. Benefiting from the minimization of cross-domain feature inconsistency in Eq. (10), the image representation $f(\mathbf{X})$ seems to be invariant across different domains as shown in Fig. 4 (c). Tab. 3 quantitatively investigate the effect of learning objective in Eq. (10). We can observe a clear performance improvement by just minimizing the cross-domain feature inconsistency (Obj-2 in Tab. 3). Obviously, by jointly optimizing the two objectives, we obtain the best generalization performance. The quantitative and qualitative results clearly reflects the importance and efficacy of the two learning objectives in Sec. 3.2.2.

R4: Robustness to Complex Realistic Scenarios. In this section, we evaluate the generalization of our HVT-enhanced methods on the DrivingStereo [32] dataset which is collected from complex realistic scenarios: *Sunny*, *Cloudy*, *Rainy*, and *Foggy*, as shown in Fig. 5. We conduct experiments using the PSMNet [3] and GwcNet [10] as the baseline frameworks. The results of officially released fine-tuned (FT) networks of PSMNet and GwcNet on the realistic KITTI 2015 dataset are included for comparison. Besides, we also include the results of SOTA domain generalized SM methods [5, 46] in the Tab. 4. We have the following observations from Tab. 4 and Fig. 5:

- Our methods (HVT-PSMNet and HVT-GwcNet) obtain the best overall performance (5.4% and 5.3%) *w.r.t.* the average D1 error rate over the four groups of weather conditions, which demonstrates the efficacy of HVT and the strong robustness of HVT-based methods.
- HVT-PSMNet and HVT-GwcNet not only outperform the baselines PSMNet and GwcNet, respectively, by a large margin, but also decrease the error rates of the FT-PSMNet and FT-GwcNet on the *Rainy* and *Foggy* groups.

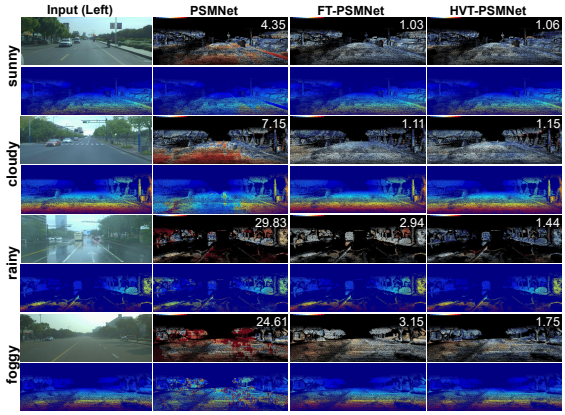

Figure 5. Qualitative results on the DrivingStereo [32] dataset. For each group, top row shows the left image and the EPE error maps of PSMNet, fine-tuned (FT) PSMNet, and our HVT-PSMNet, respectively. Bottom row shows the GT disparity map and the predicted disparity maps of different methods. The EPE values are marked in the upper right corner of error maps.

On another two groups, both fine-tuned (FT) models perform slightly better than our HVT, since images in the two groups are similar to the KITTI training data. The results reflect that our HVT can effectively deal with the domain shift between the synthetic data and realistic data.

- Fig. 5 shows the qualitative robustness comparison in different weather conditions. The observation from Fig. 5 is consistent with that from Tab. 4, which further validates the potential of HVT in enhancing the model robustness.

5. Conclusion

In this work, we have devised an effective HVT module for the problem of domain generalized SM. To prevent the network from exploiting the shortcut features for disparity estimation, we propose to 1) first diversify the training domain by leveraging the three complementary visual transformations, where the cross-domain visual discrepancy is maximized, and 2) minimize the cross-domain feature inconsistency to effectively capture domain-invariant features. Our proposed method is simple and can be flexibly integrated with most existing SM networks. Extensive experimental results show that our HVT consistently advances the learning of shortcut-robust features and substantially improves the generalization performance of SM networks in unseen realistic scenarios. In the future, we will attempt to design more delicate visual transformations to enhance HVT and extend HVT for addressing other 3D vision tasks.

6. Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62272435, Grant U22A2094 and Grant 72188101.

References

[1] Joydeep Biswas and Manuela Veloso. Depth camera based localization and navigation for indoor mobile robots. In *RGB-D Workshop at RSS*, volume 2011, 2011. 1

[2] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization. In *2020 International Conference on 3D Vision (3DV)*, pages 364–373. IEEE, 2020. 2, 6

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 1, 2, 3, 6, 7, 8

[4] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33:22158–22169, 2020. 1, 2

[5] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13022–13032, 2022. 1, 2, 3, 4, 6, 8

[6] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4065–4080, 2021. 1

[7] Mohammed E Fathy, Quoc-Huy Tran, M Zeeshan Zia, Paul Vernaza, and Manmohan Chandraker. Hierarchical metric learning and matching for 2d and 3d geometric correspondences. In *Proceedings of the european conference on computer vision (ECCV)*, pages 803–819, 2018. 1

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 5, 6

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 2, 6

[10] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 1, 2, 6, 7, 8

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

[12] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):1–17, 2018. 4

[13] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7130–7140, 2022. 3

[14] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 2, 3

[15] Xing Li, Yangyu Fan, Zhibo Rao, Guoyun Lv, and Shiya Liu. Synthetic-to-real domain adaptation joint spatial feature transform for stereo matching. *IEEE Signal Processing Letters*, 29:60–64, 2021. 1, 2

[16] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2, 6, 7

[17] Biyang Liu, Huimin Yu, and Guodong Qi. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13012–13021, 2022. 2, 3, 6

[18] Xueliang Liu, Xun Yang, Meng Wang, and Richang Hong. Deep neighborhood component analysis for visual similarity modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11:1 – 15, 2020. 1

[19] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1, 2, 5, 6

[20] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 5, 6

[21] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[22] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liu. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, 30:6594–6608, 2021. 4

[23] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014. 5, 6

[24] Thomas Schops, Johannes L Schonberger, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017. 5, 6

[25] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13906–13915, 2021. 2, 6, 7

[26] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Yuexin Ma, Zhe Wang, and Jianping Shi. Adastereo: An efficient domain-adaptive stereo matching approach. *International Journal of Computer Vision*, 130(2):226–245, 2022. 1, 2

[27] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, 128(4):910–930, 2020. 1, 2

[28] Chen Sun, Jean M Uwabeza Vianney, and Dongpu Cao. Affordance learning in direct perception for autonomous driving. *arXiv preprint arXiv:1903.08746*, 2019. 1

[29] Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–385, 2022. 3, 4

[30] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 1, 2

[31] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 4

[32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 8

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

[34] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. 1

[35] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua. Interpretable fashion matching with rich attributes. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019. 1

[36] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1939–1947, 2020. 1

[37] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31:1204–1216, 2022. 1

[38] Xun Yang, Peicheng Zhou, and Meng Wang. Person reidentification via structural deep metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30:2987–2998, 2019. 1

[39] Chengtang Yao, Yunde Jia, Huijun Di, Pengxiang Li, and Yuwei Wu. A decomposition model for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6091–6100, 2021. 2

[40] Yu Yu, Shahram Khadivi, and Jia Xu. Can data diversity enhance learning generalization? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, 2022. 3

[41] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1592–1599, 2015. 1

[42] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H-S Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 1, 2, 6

[43] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision*, pages 420–439. Springer, 2020. 2, 6

[44] Feihu Zhang and Benjamin W Wah. Fundamental principles on learning new features for effective dense matching. *IEEE Transactions on Image Processing*, 27(2):822–836, 2017. 1

[45] Haoyuan Zhang, Lap-Pui Chau, and Danwei Wang. Soft warping based unsupervised domain adaptation for stereo matching. *IEEE Transactions on Multimedia*, 2021. 1, 2

[46] Jiawei Zhang, Xiang Wang, Xiao Bai, Chen Wang, Lei Huang, Yimin Chen, Lin Gu, Jun Zhou, Tatsuya Harada, and Edwin R Hancock. Revisiting domain generalized stereo matching networks from a feature consistency perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13001–13011, 2022. 2, 6, 8

[47] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020. 3