Proceedings Paper:
Zhang, Jiawei, Wang, Xiang, Xiao, Bai et al. (7 more authors) (Accepted: 2022) Revisiting Domain Generalized Stereo Matching Networks from a Feature Consistency Perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. . (In Press)

Revisiting Domain Generalized Stereo Matching Networks from a Feature Consistency Perspective

Jiawei Zhang[1], Xiang Wang[1], Xiao Bai[1]*, Chen Wang[1], Lei Huang[2],
Yimin Chen[1], Lin Gu[3,4], Jun Zhou[5], Tatsuya Harada[3,4], Edwin R. Hancock[6,1]
[1]School of Computer Science and Engineering, State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University, China,
[2]SKLSDE, Institute of Artificial Intelligence, Beihang University, Beijing, China,
[3]RIKEN AIP, Tokyo, Japan, [4]The University of Tokyo, [5]Griffith University, [6]University of York

Abstract

Despite recent stereo matching networks achieving impressive performance given sufficient training data, they suffer from domain shifts and generalize poorly to unseen domains. We argue that maintaining feature consistency between matching pixels is a vital factor for promoting the generalization capability of stereo matching networks, which has not been adequately considered. Here we address this issue by proposing a simple pixel-wise contrastive learning across the viewpoints. The stereo contrastive feature loss function explicitly constrains the consistency between learned features of matching pixel pairs which are observations of the same 3D points. A stereo selective whitening loss is further introduced to better preserve the stereo feature consistency across domains, which decorrelates stereo features from stereo viewpoint-specific style information. Counter-intuitively, the generalization of feature consistency between two viewpoints in the same scene translates to the generalization of stereo matching performance to unseen domains. Our method is generic in nature as it can be easily embedded into existing stereo networks and does not require access to the samples in the target domain. When trained on synthetic data and generalized to four real-world testing sets, our method achieves superior performance over several state-of-the-art networks. The code is available online[1].

1. Introduction

Estimating depth from images is a fundamental problem in many computer vision applications such as autonomous driving [42] and robot navigation [1]. Stereo matching is a solution to this task, which finds the matching correspondences between stereo image pairs and recovers the depth through triangulation.

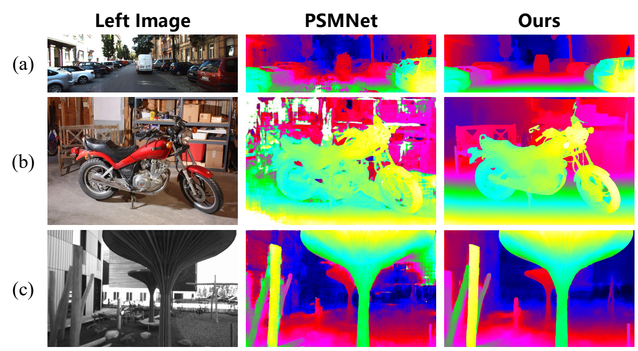


Figure 1. Domain generalization performance of PSMNet with and without our method on samples from (a) KITTI, (b) Middlebury, and (c) ETH3D training sets. All models are trained on the synthetic SceneFlow dataset.

Stereo matching is traditionally solved by a matching cost computation process, which usually consists of four steps [38]: matching cost computation, cost aggregation, disparity regression, and disparity refinement. Recently, end-to-end stereo matching networks [4, 15, 21, 30, 57] have been developed based on the cost computation process of traditional methods and achieved state-of-the-art accuracy. However, the poor generalization performance on unseen domains has been a major challenge for their real-world applications (see Figure 1 for an example).

A common approach to achieve generalization capability is to learn domain-invariant representations [23, 24, 26, 32]. Some stereo matching networks [3, 40, 58] have made attempts to tackle this issue by conducting feature-level alignment to obtain domain-invariant features. These works project the inputs into a domain-invariant feature space, reducing the reliance on domain-specific appearance properties and showing more robustness to domain shifts.

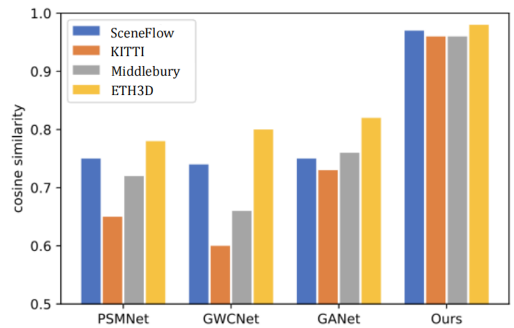
Here, we present a weaker constraint, *stereo feature con-*

*Corresponding author: Xiao Bai (baixiao@buaa.edu.cn).
[1]https://github.com/jiaw-z/FCStereo

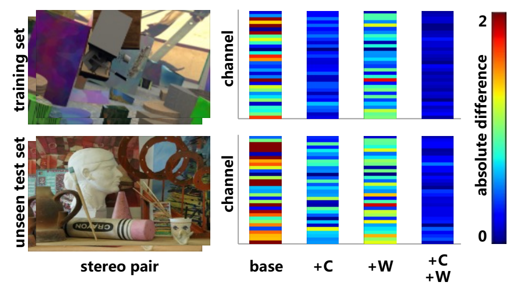
sistency, for domain generalized stereo networks. For each point in the left image, stereo matching looks for its matching one in the right view, which naturally requires *robustness to viewpoint changes*. A domain generalized stereo network is expected to generalize this matching ability to unseen domains, which means, in a nutshell, the *generalization of "robustness to viewpoint changes"*. From this perspective, we believe what a stereo network needs to generalize is the matching relationship, behaving as the feature consistency of paired points. For example, traditional methods, which are largely domain-agnostic [36, 43], compute the matching cost directly on RGB images [17]. Although image contents differ considerably across different domains, the matching pixels have consistent expressions between the stereo viewpoints in most cases, guaranteeing stable matching cost computation to produce reliable disparity maps. We further verified this intuition to a toy pipeline that combines a cost volume constructed directly from RGB images with the common PSMNet cost aggregation module (*cf.* Appendix A). Such a simple pipeline with consistent stereo representations also shows a significant improvement in domain generalization performance.

Generally, the appearance inconsistency within a stereo pair is limited to a certain range, thus the matching points being very similar. For example, the corresponding points should share the identical incident light as well as albedo and differ in the shading that appeared in left and right cameras. However, when the learned features are used to construct the cost volume, the feature consistency is not preserved, as shown in Figure 2a. And surprisingly, the features are inconsistent even in the training set, which is contrary to the common intuition that the **weight-sharing Siamese** feature extractor has dealt with stereo viewpoint changes and extracted consistent features.

In this paper, we address the domain generalization for stereo matching methods by developing the Feature Consistency Stereo networks (FCStereo). Here comes two challenges: (a) obtaining a high feature consistency on the training set and (b) generalizing this consistency across different domains. We argue that the difficulty of (a) is due to the lack of explicit consistent constraints on features which causes overfitting. We propose the stereo contrastive feature (SCF) loss to encourage the matching points to be close in the representation space. To solve the consistency generalization problem (b), we utilize a proper normalization operation and constrain the feature statistics. A stereo selective whitening (SSW) loss is further introduced to suppress information that is sensitive to stereo viewpoint changes. Figure 2b illustrates the feature differences in a channel-wise manner and shows the role of the two proposed loss terms. SCF loss encourages features to be consistent on the training set. However, we see a degradation of consistency on unseen domains. SSW loss yields a relatively lower consistency compared to the contrastive loss, while the consistency is more robust to domain changes. Jointly using both loss terms enables high feature consistency in various domains. We apply the proposed method to different stereo matching backbones in the experiment and show a significant improvement in generalization performance. It demonstrates that **the generalization of feature consistency between two viewpoints in the same scene translates to the generalization of stereo matching performance to unseen domains** though appears counter-intuitive. A qualitative illustration is shown in Figure 1. The main contributions of this paper are as follows:



(a)



(b)

Figure 2. Analysis about feature consistency of matching points. (a) Evaluation of popular stereo matching backbones on four unseen domains. (b) Visualization of per-channel feature inconsistency. Left-right: PSMNet baseline, with our contrastive loss (**C**), with our whitening loss (**W**), and with both. More details about learned feature are shown in Appendix C.

- We observe that most recent stereo methods learn inconsistent representations for the pairs of matching points, and demonstrate that the generalization performance of stereo networks can be boosted by maintaining a high stereo feature consistency.
- We propose two loss functions, namely the stereo contrastive feature loss and the stereo selective whitening loss, to encourage the stereo feature consistency across domains. These two losses could be easily embedded in the existing stereo networks.
- Our method is applied to several stereo network architectures and shows a significant improvement in their domain generalization performance.

2. Related Work

Deep learning based stereo matching. Since MC-CNN [56] introduced a convolution neural network (CNN) to matching cost calculation, many deep learning based methods have been proposed for stereo matching. Early works simply replace RGB inputs with expressive learned features for higher accuracy, leaving the following traditional steps for cost computation unchanged. For these methods, a crucial attribute of learned features is the consistency between matching pixels [9, 59].

More recently, many methods solve the task in an end-to-end way [4, 22, 30, 46, 60]. Two types of solutions are normally followed by these methods: correlation cost volume based deep neural networks with 2D cost aggregation and concatenation cost volume based stereo networks with 3D cost aggregation. The correlation methods are usually more efficient but cause information loss. DispNetC [30] is the first method that introduces end-to-end regression for stereo matching and builds the cost volume in a correlation way. The correlation based matching strategy is adopted by many works [28,44,52,55] and has achieved impressive and efficient performance. The second category concatenates the stereo features to make full use of information. For example, GCNet [21] stacks two view features to build a 4D cost volume and first utilizes 3D convolution for matching cost aggregation. Methods in this category [4, 48, 57, 60] leverage more complete information of features and have produced higher accuracy on various stereo benchmarks. Our method can be seamlessly integrated into the existing end-to-end stereo networks and improve their generalization performance.

Domain generalized stereo matching. It is important to develop stereo matching networks that are robust to unseen domains. DSMNet [58] uses a domain normalization layer to reduce the shifts of image-level styles and local contrast variations, followed by a trainable non-local graph-based filter to smooth the local sensitive local details. CFNet [40] produces a fused cost volume representation for capturing global and structural information to construct a stereo matching network that is robust to domain changes. Cai *et al*. [3] point out that the poor generalization of stereo networks is caused by the strong dependence of the network on the image appearance, and propose to use a combination of matching functions for feature extraction.

Instance discrimination and contrastive learning. Instance-level discrimination, regarding each instance as a distinct class of its own, plays an important role in representation learning. This paradigm is formulated as a metric learning problem, where features of positive sample pairs are encouraged to be close and those of negative sample pairs are forced to be apart [50]. The following work [41] adopts this idea to specific downstream tasks and shows that the quality of learned representations is heavily affected by the strategy of negative pair selection [41]. Recently, following the idea of instance discrimination, contrastive learning has made remarkable success in self-supervised feature representation learning. MoCo and its variant [5,16] treat contrastive learning as a dictionary look-up process and maintain a momentum updated queue encoder. Some attempts extend this momentum-based contrastive learning framework to pixel-level feature learning [35, 47, 51]. Different from these dense contrastive learning methods, we define positive samples using correspondences given by the ground truth disparity, which is directly tailored to the main task.

Feature covariance. Previous studies have demonstrated that the correlations between feature channels capture the style information of images [11, 12]. This theory is further explored in style transfer [12, 25], image-to-image translation [6], and others [29, 33]. More recently, [7] propose a selective whitening method to remove the style information that is sensitive to domain shifts for robust segmentation, where the style information selection depends on the manually designed photometric transformation. Our approach is inspired by the selective whitening [7], however, we select the information sensitive to stereo viewpoint changes, without relying on photometric transformation.

3. Approach

In this section, we present the details of our method, including a stereo selective whitening loss on the intermediate features and a stereo contrastive feature loss on the final features. Figure 3 depicts the whole framework of our method, where the stereo contrastive feature loss and stereo selective whitening loss are applied to a stereo matching network for encouraging the stereo feature consistency across domains.

3.1. Stereo Contrastive Feature Loss

Stereo features from the last feature extraction layer are used to construct the cost volume, which is the most important internal representation in a deep stereo network. At this stage, we impose a consistency constraint on the features of stereo views. Inspired by the recent success of contrastive learning on unsupervised feature learning via optimizing the pairwise (dis)similarity, we introduce a contrastive learning mechanism on stereo features, namely the Stereo Contrastive Feature (SCF) loss. The proposed contrastive learning mechanism includes a pixel-level contrastive loss applied on stereo features, and a dictionary queue with a momentum updated key encoder, which introduces a rich set of negative samples from different pairs and further improves the feature consistency.

Positive pairs. We consider the pixel vectors in stereo views as a positive pair if their pixel coordinates are the projected locations of the same 3D point. These positive pairs can be collected using the ground truth disparity d of the left

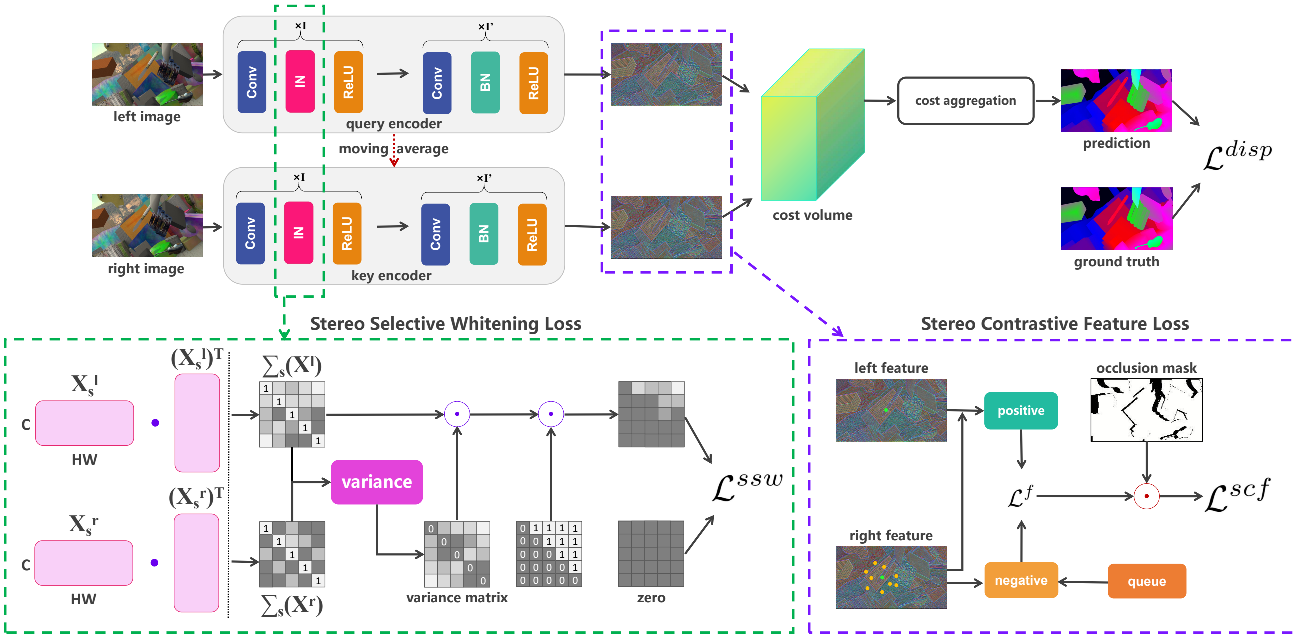


Figure 3. Structure of method. The top part shows the forward pass that the network extracts features from input pairs and regresses the disparity on the feature based cost volume. At the bottom are the proposed two losses for maintaining the feature consistency. As part of the feature contrastive loss, the key encoder for the right image is implemented as a moving average of the query encoder to alleviate the effects of negative sample selections. In the inference process, we use the query encoder to extract features for both left and right images, which is strictly identical to the standard pipeline.

view, *i.e.*, the query feature in the left view $\phi^l_{u,v}$ are paired with the key feature in the right view $\phi^r_{u-d,v}$. Therefore, maintaining *stereo feature consistency* simply becomes promoting the feature consistency of positive pairs.

Negative pairs. While each of the pixels in ϕ^l has $HW - 1$ potential negative pairs in ϕ^r, involving all potential negative pairs in the contrastive loss would lead to a huge computational cost. To overcome this issue, we use a naive method that randomly samples N non-matching points from the right feature ϕ^r in a local window with a size of 50×50 to form N negative pairs.

Momentum encoder. The selection strategy of negative samples is particularly important as it heavily affects the property of learned representations [41, 51]. In our experiment, the feature consistency is not high enough when the negative samples are limited to the same stereo pair. Using pixels from other images as negative samples can be more natural [35] and it fits well with a dictionary queue for negative samples.

We follow [16] to maintain a dynamic dictionary queue that stores preceding negative samples and change the architectural design of weight-sharing feature extractors into an asymmetric pair of query and key encoders (with weights $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$). The capacity of the queue is fixed as K and the oldest samples are progressively replaced after each iteration. The key encoder is modeled as a momentum-based moving average of the query encoder:

$$\boldsymbol{\eta}_t = m\boldsymbol{\eta}_{t-1} + (1-m)\boldsymbol{\theta}_t, \tag{1}$$

where t is the number of iterations and $m \in [0, 1]$ is a momentum value. Such a design can provide content from different images for negative samples, which reduces the opportunity for features to focus too much on the contents of the current image. As the momentum value m, plays a core role in making use of a queue [16], we evaluate the feature consistency in the experiment. It indicates that a relatively large momentum value (*e.g.*, $m = 0.9999$, our default) plays a core role in extracting stereo consistent representations.

Pixel-wise contrastive loss. We measure the similarity of feature pairs with dot product, and adopt a pixel-level InfoNCE [34] to our problem:

$$\mathcal{L}^f(u,v) = -\log \frac{\exp\left(\phi^l_{u,v} \cdot \phi^r_{u-d,v}/\tau\right)}{\sum_{\phi^n \in \mathcal{F}(u,v)} \exp\left(\phi^l_{u,v} \cdot \phi^n/\tau\right)}, \tag{2}$$

where $\mathcal{F}(u,v)$ denotes the negative sample set of sample $\phi^l_{u,v}$, consisting N samples from the right feature ϕ^r and K samples from the dictionary queue, and τ is a temperature hyper-parameter [50]. We set $N = 60$, $K = 6000$, and $\tau = 0.07$.

Non-matching region removal. We leverage ground truth disparity to collect pixel feature pairs as positive pairs in the contrastive framework. However, some pairs collected in this way don't originate from the same point in the 3D world, due to factors like occlusions. Hence these *non-matching* pairs should be detected and eliminated from the positive sample set. A widely used matching confidence criterion, left-right geometric consistency check, can be leveraged to detect and remove those non-matching sample pairs. The reprojection error $\mathbf{R}$ is computed as the difference of ground truth disparity values at paired pixel locations in stereo images and could be served as a criterion for matching validity check. Then the mask $\mathbf{M}$ denoting the remaining matching regions is defined as:

$$\mathbf{M}_{u,v} = \begin{cases} 1, & \mathbf{R}_{u,v} < \delta \\ 0, & otherwise \end{cases} \tag{3}$$

where δ is set to 3 as a threshold. And our SCF loss is defined as the weighted average of $\mathcal{L}^f$ over the pixel coordinate space $\mathcal{C}$:

$$\mathcal{L}^{scf} = \frac{1}{\sum_{(u,v)\in\mathcal{C}} \mathbf{M}_{u,v}} \sum_{(u,v)\in\mathcal{C}} L^f(u,v) \odot \mathbf{M}_{u,v}. \tag{4}$$

3.2. Stereo Selective Whitening Loss

With the contrastive loss, stereo network extracts consistent representation on the training set. However, the degradation of feature consistency across different domains has become the main obstacle to the further improvement of generalization performance. We build the stereo selective whitening (SSW) loss based on [7] to address this problem.

Generally, stereo networks use batch normalization (BN) [20] as their default feature normalization operation. During training, BN regularizes the feature with the mini-batch statistics and uses population statistics of the training set during inference [19], which makes the statistics of networks data-dependent [58] and is sensitive to the change of domains. To generalize the feature consistency to different domains, we change some default BN layers into instance normalization (IN) [45] layers, which regularizes each sample separately, therefore, is independent of training set statistics. For each sample $\mathbf{X} \in \mathbb{R}^{C\times HW}$, IN transforms it into $\hat{\mathbf{X}} \in \mathbb{R}^{C\times HW}$:

$$\hat{\mathbf{X}}_i = \frac{1}{\sigma_i}(\mathbf{X}_i - \mu_i), \tag{5}$$

where μ_i and σ_i are mean and the standard deviation of $\hat{\mathbf{X}}$ along the channel index i.

We further consider the information stored in the feature covariance, which is not dealt with by IN. The proposed SSW seeks for learning viewpoint-invariant representation by suppressing the feature covariance components that are sensitive to stereo view changes. In particular, we firstly compute the variance matrix $\mathbf{\Sigma}(\hat{\mathbf{X}}) \in \mathbb{R}^{C\times C}$ of the IN regularized representation $\hat{\mathbf{X}}$:

$$\mathbf{\Sigma}(\hat{\mathbf{X}}) = \frac{1}{HW}(\hat{\mathbf{X}})(\hat{\mathbf{X}})^{\mathrm{T}}. \tag{6}$$

We then compute the covariance matrices $\mathbf{V} \in \mathbb{R}^{C\times C}$ between each left view feature covariance $\mathbf{\Sigma}(\hat{\mathbf{X}}^l)$ and its corresponding right feature $\mathbf{\Sigma}(\hat{\mathbf{X}}^r)$, where n indexes the samples:

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{\Sigma}_n} &= \frac{1}{2}(\mathbf{\Sigma}_n(\hat{\mathbf{X}}^l) + \mathbf{\Sigma}_n(\hat{\mathbf{X}}^r)) \\ \mathbf{V} &= \frac{1}{2N}\sum_{n=1}^{N}((\mathbf{\Sigma}_n(\hat{\mathbf{X}}^l) - \boldsymbol{\mu}_{\mathbf{\Sigma}_n})^2 + (\mathbf{\Sigma}_n(\hat{\mathbf{X}}^r) - \boldsymbol{\mu}_{\mathbf{\Sigma}_n})^2), \end{aligned} \tag{7}$$

The element $\mathbf{V}_{i,j}$ from the variance matrix measures how sensitive the correspondence between the i-th and the j-th channels to stereo viewpoint changes. Covariance elements with high variances between left and right features, which are considered as sensitive components to stereo view changes, should be considered in the whitening loss. In practice, all covariance elements are grouped into 3 clusters by the magnitudes of variance [7], and we choose the one with the highest variance value, termed $\mathcal{G}_p$. Then a selective mask $\tilde{\mathbf{M}} \in \mathbb{R}^{C\times C}$ is computed:

$$\tilde{\mathbf{M}}_{i,j} = \begin{cases} 1, & \mathbf{V}_{i,j} \in \mathcal{G}_p \\ 0, & otherwise \end{cases} \tag{8}$$

The SSW loss is imposed on the left regularized features:

$$\mathcal{L}^{ssw} = \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} ||\mathbf{\Sigma}_\gamma(\hat{\mathbf{X}}^l) \odot \tilde{\mathbf{M}} \odot \hat{\mathbf{M}}||_1, \tag{9}$$

where $\hat{\mathbf{M}}$ is a strict upper triangular matrix as the covariance matrix is symmetric, Γ is the number of layers to which the SSW loss is applied, and γ indexes the corresponding layer (*i.e.* conv1, conv2_x in PSMNet). With the SSW loss, the stereo network learns to rely less on stereo irrelevant information to form its feature representation. The differences within a stereo image pair are mostly limited to specific physical characteristics, such as diffuse reflection of light, which gives the stereo model the possibility to learn some general knowledge from limited training data.

3.3. Training Objectives

Our final training loss is a weighted sum of the disparity loss and above-mentioned losses:

$$\mathcal{L} = \mathcal{L}^{disp} + \lambda^{scf}\mathcal{L}^{scf} + \lambda^{ssw}\mathcal{L}^{ssw}, \tag{10}$$

where $\mathcal{L}^{disp}$ is a commonly used per-pixel smooth-$\mathcal{L}_1$ loss for disparity regression. λ^{scf} and λ^{ssw} are the balancing weights. During back-propagation, all other modules are updated by classical gradient descent methods, except that the right feature extractor is implemented as a moving average of the left extractor.

Backbone	Contrastive Loss	Momentum Encoder	Stereo Whitening	KITTI 2012	KITTI 2015	Middlebury half	Middlebury quarter	ETH3D
PSMNet [4]	✗	✗	✗	26.5	27.9	26.9	20.0	23.8
	✓	✗	✗	18.4	19.0	24.1	15.4	17.6
	✓	✓	✗	10.5	12.7	22.2	15.0	17.1
	✗	✗	✓	13.2	15.5	20.5	13.8	14.1
	✓	✓	✓	**7.0**	**7.5**	**18.3**	**12.1**	**12.8**
GWCNet [15]	✗	✗	✗	20.2	22.7	34.2	18.1	30.1
	✓	✗	✗	12.3	16.5	25.8	15.5	13.3
	✓	✓	✗	11.2	12.1	24.8	15.2	12.8
	✗	✗	✓	12.0	13.5	24.6	14.9	12.5
	✓	✓	✓	**7.4**	**8.0**	**21.0**	**11.8**	**11.7**
GANet [57]	✗	✗	✗	10.1	11.7	20.3	11.2	14.1
	✓	✗	✗	9.1	9.5	18.1	10.5	12.1
	✓	✓	✗	7.2	7.5	16.3	10.1	11.3
	✗	✗	✓	8.4	9.0	16.8	10.2	10.5
	✓	✓	✓	**5.7**	**6.4**	**16.0**	**9.8**	**9.2**

Table 1. Ablation study of each key component with various backbones on the KITTI, Middlebury, and ETH3D training sets. Threshold error rates (%) are adopted for evaluation.

4. Experiments

In this section, we make a detailed analysis of some commonly used stereo methods to illustrate that the existing framework lacks explicit constraints on features. We also perform ablation studies on different datasets, including KITTI [13, 31], Middlebury [37], ETH3D [39], DrivingStereo [54], to verify the role of different components. We compare our method with existing domain generalized stereo networks to show the effectiveness of our method.

4.1. Datasets

SceneFlow [30] is a large synthetic dataset containing three subsets: Driving, Monkaa, and FlyingThings3D. The training set includes $35k$ pairs of synthetic stereo images and dense ground-truth disparities with a resolution of 960×540, which is used to train networks from scratch in our experiments.
KITTI2012 [13] and **KITTI2015** [31] both collect outdoor driving scenes with a full resolution of 1242×375. They provide 394 stereo pairs with sparse ground-truth disparities for training and 395 pairs for testing. We use the training sets to evaluate the generalization performance of networks.
Middlebury 2014 [37] is an indoor dataset, providing 28 training (including 13 additional stereo pairs) and 15 testing stereo pairs with full, half, and quarter resolutions. We use half and quarter resolution training sets to evaluate the generalization ability of networks.
ETH3D [39] consists of 27 grayscale image pairs for training and 20 for testing. It includes both indoor and outdoor scenes. We use the training set for generalization performance evaluation.
DrivingStereo [54] is a large-scale real dataset. A subset of it contains 2,000 stereo pairs collected under different weathers (sunny, cloudy, foggy, and rainy). We evaluate the generalization performance on these challenging scenes.

4.2. Implementation Details

We implement our method in PyTorch and train it with Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The batch size is set to 12 on GPUs. We train the models from scratch with the learning rate of 0.001 for 15 epochs and 0.0001 for further 5 epochs. We randomly crop the raw images to 512×256 as input. For all datasets, color normalization is used with the mean ([0.485, 0.456, 0.406]) and variation ([0.229, 0.224, 0.225]) of the ImageNet [8] for data preprocessing. We set the maximum disparity as $D = 192$, and all ground-truth disparities larger than D are excluded from the loss calculation. During training, we use asymmetric query and key encoders to extract features from left and right images, respectively. In the test phase, the query encoder is used as the feature extractor for both left and right images, which is the symmetric design strictly identical to the standard stereo pipeline.

4.3. Ablation Study

In this section, We present detailed ablation studies to evaluate and analyze the effectiveness of our method.
Key components: We evaluate the effectiveness of each key component of our pipeline. Here, three networks are selected as baseline models. PSMNet [4] is a widely-adopted backbone. It constructs a concatenation based cost volume and hopes that the cost aggregation network can learn a similarity measurement function from scratch. GWC-

m	-	0.9	0.99	0.999	0.9999	0.99999
SceneFlow	0.86	0.88	0.92	0.96	0.97	0.98
KITTI	0.78	0.82	0.85	0.91	0.92	0.92

Table 2. Feature consistency with different momentum values in both seen (SceneFlow) and unseen (KITTI) domains. Cosine similarity is adopted as the vector-wise similarity metric for evaluation. '-' denotes the contrastive learning setting without queue and momentum encoder. We select PSMNet as the baseline model.

Methods	KITTI	Middlebury
PSMNet Baseline	12.7	22.2
+ Instance Norm [45]	8.5	19.1
+ Domain Norm [58]	8.1	18.8
+ Instance Whitening [7]	8.0	18.6
+ Our Stereo Selective Whitening	**7.5**	**18.3**

Table 3. Comparisons with existing normalization layers on the KITTI 2015 and the half resolution Middlebury training sets. Threshold error rates (%) are adopted. We select PSMNet as the baseline model.

Net [15] is selected as it constructs the cost volume with a group-wise correlation, which provides better similarity measures than learning from scratch. GANet [57] is one of the top-performing networks, guiding the cost aggregation with low-level features. As shown in Table 1, applying the contrastive loss on the final features during training significantly improves the domain generalization performance. Benefiting from the negative samples from different stereo pairs and the momentum encoder, it generalizes better to various domains, *e.g*. 15.2% on KITTI, 4.7% on Middlebury, and 6.7% on ETH3D for PSMNet. In addition, the error rates on unseen domains are reduced by the whitening loss. It shows that maintaining the feature consistency across different domains can effectively improve the generalization performance. Furthermore, with the combination of these key components, the models significantly outperform their corresponding baseline models in unseen domains.

Momentum value: Table 2 shows the consistency of learned stereo features with different momentum values (m in Equation (1)). Compared to the standard Siamese encoder without the dictionary queue, our momentum encoder is shown beneficial for the feature consistency of positive pairs, and this behavior holds for both seen and unseen domains. We also adopt different momentum values m and show that a relatively large m is vital to achieve higher stereo feature consistency.

Normalization layer: We evaluate the effectiveness of our proposed stereo whitening with batch normalization [20], instance normalization [45], domain normalization [58] and instance whitening [7]. All other settings except for the normalization method are kept the same in the experiment. Compared with the general methods, our whitening loss is specifically designed for stereo matching and helps the model to generalize better to unseen domains, as shown in Table 3.

Methods	KITTI		Middlebury	
	cosine	>3px	cosine	>2px
AANet [52]	0.76	12.3	0.77	28.1
Cas-PSMNet [14]	0.58	16.5	0.63	27.8
AcfNet [60]	0.58	27.4	0.61	27.1
CDN-PSMNet [10]	0.57	40.0	0.61	35.0
PSMNet [4]	0.65	27.9	0.71	26.9
GWCNet [15]	0.60	22.7	0.67	34.2
GANet [57]	0.73	11.7	0.76	20.3
DSMNet [58]	0.83	6.5	0.85	13.8
FC-PSMNet (ours)	0.98	7.5	0.95	18.3
FC-GWCNet (ours)	0.97	8.0	0.95	21.0
FC-GANet (ours)	0.98	6.4	0.97	16.0
FC-DSMNet (ours)	**0.99**	**6.2**	**0.98**	**12.0**

Table 4. Evaluation of feature consistency and generalization performance on the KITTI 2015 and the half resolution Middlebury training sets. Cosine similarity and threshold error rates (%) are adopted.

Feature consistency of various architectures: We present the feature consistency and the generalization performance of various networks. We use the cosine similarity as a vector-wise metric to evaluate the consistency between matching features. The aforementioned baseline methods, PSMNet [4], GWCNet [15], GANet [57], are included in this experiment. AANet [52] is selected as the representation for full correlation cost volume based methods [30,52]. We also evaluate the CasPSMNet [14] as the representation for coarse-to-fine methods. DSMNet [58] is designed for domain generalization and extracts the non-local features. In addition to the common disparity regression loss, AcfNet [60] imposes additional constraints on the probability distribution derived from the filtered cost volume. CDN [10] replaces the commonly used softargmin-based regression loss with a Wasserstein distance based loss. As shown in Table 4, these popular methods all lack explicit constraints on features and extract inconsistent representations.

4.4. Cross-domain Evaluation

We compare our methods with several other stereo matching methods, including traditional methods, well-researched end-to-end methods, and domain generalized methods by training on synthetic SceneFlow training set and evaluating on four real-world datasets. The asymmetric augmentation [49, 53] is used to prevent the model from overfitting. Table 5 summarizes the comparisons. Our method achieves superior generalization performance than others.

4.5. Evaluation on Challenging Weathers

In this section, we evaluate the generalization performance of our method on some challenging domains. We train the baseline models and our models under the asym-

Methods	KITTI 2012	KITTI 2015	Middlebury half	Middlebury quarter	ETH3D
CostFilter [18]	21.7	18.9	40.5	17.6	31.1
PatchMatch [2]	20.1	17.2	38.6	16.1	24.1
SGM [17]	7.1	7.6	25.2	10.7	12.9
Training set	SceneFlow				
PSMNet [4]	26.5	27.9	26.9	20.0	23.8
GWCNet [15]	20.2	22.7	34.2	18.1	30.1
GANet [57]	10.1	11.7	20.3	11.2	14.1
MS-PSMNet [3]	13.9	7.8	19.9	10.8	16.8
MS-GCNet [3]	**5.5**	6.2	18.5	10.3	8.8
DSMNet [58]	6.2	6.5	13.8	8.1	6.2
FC-PSMNet (ours)	7.0	7.5	18.3	12.1	12.8
FC-GWCNet (ours)	7.4	8.0	21.0	11.8	11.7
FC-GANet (ours)	5.7	6.4	16.0	9.8	9.2
FC-DSMNet (ours)	**5.5**	**6.2**	**12.0**	**7.8**	**6.0**
Training data	SceneFlow + Asymmetric Augmentation				
PSMNet [4]	6.0	6.3	15.8	9.8	10.2
GANet [57]	5.5	6.0	13.5	8.5	6.5
STTR [27]	8.7	6.7	15.5	9.7	17.2
CFNet [40]	4.7	5.8	15.3	9.8	**5.8**
FC-PSMNet (ours)	5.3	5.8	15.1	9.3	9.5
FC-GANet (ours)	**4.6**	**5.3**	**10.2**	**7.8**	**5.8**

Table 5. Cross-domain generalization evaluation on the KITTI, Middlebury, ETH3D training sets. Threshold error rates (%) are adopted.

Methods	KITTI	Cloudy	Foggy	Rainy	Sunny
PSMNet [4]	6.3	7.9	10.8	12.2	7.4
FC-PSMNet (ours)	**5.8**	**4.3**	**6.2**	**7.2**	**4.9**
GANet [57]	6.0	5.7	8.2	10.0	5.4
FC-GANet (ours)	**5.3**	**3.3**	**4.0**	**7.0**	**3.3**

Table 6. Generalization evaluation on the KITTI and the half resolution DrivingStereo data sets of different weather conditions. Threshold error rates (%) are adopted.

metric augmentation. The trained models are tested on stereo pairs collected in four challenging weather conditions provided by DrivingStereo. The KITTI 2015 training set is also evaluated as it collects similar outdoor driving scenes under ideal weather conditions. The results are summarized in Table 6. Compared with baseline models, our models generalize better to images under ideal weather conditions, and the improvement is more obvious in challenging weather conditions. Figure 4 shows the qualitative results.

4.6. Fine-tuning on KITTI

We evaluate the fine-tuned accuracy on the KITTI benchmark. The models are first trained on SceneFlow data and fine-tuned on the KITTI 2015 training set for a further 1000 epochs. During fine-tuning, we use the query encoder to extract two view features, which is the symmetrical feature extraction identical to the standard pipeline. We set the learning rate at 0.001 for the first 600 epochs and decrease it to 0.0001 for the rest 400 epochs. Table 7 shows results on the benchmark. We see that our models can obtain comparable performance to their counterparts. In addition, we explore the performance of our method with the more limited fine-

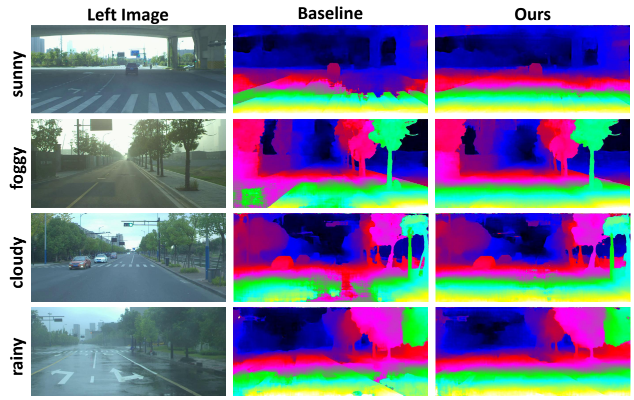


Figure 4. Qualitative results on different weather conditions of DrivingStereo. PSMNet is selected as our baseline model.

Methods	All-D1(%) bg	All-D1(%) fg	All-D1(%) all	Noc-D1(%) bg	Noc-D1(%) fg	Noc-D1(%) all
Fine-tuning set	KITTI (full)					
PSMNet [4]	1.86	4.62	**2.32**	1.71	4.31	2.14
FC-PSMNet (ours)	1.86	4.61	**2.32**	1.73	4.19	**2.13**
Fine-tuning set	KITTI (40)					
PSMNet [4]	4.15	7.03	4.63	3.92	6.36	4.32
FC-PSMNet (ours)	3.10	6.94	**3.74**	2.88	6.27	**3.44**
Fine-tuning set	KITTI (1)					
PSMNet [4]	4.83	14.26	6.40	4.57	13.38	6.02
FC-PSMNet (ours)	3.34	12.56	**4.87**	3.05	11.56	**4.45**

Table 7. After fine-tuning evaluation on the KITTI 2015 benchmark. Different subsets are used for fine-tuning.

tuning set, which is meant for practical applications because the data available for many real-world scenes is very limited. The KITTI (40) is a popular validation set [4], which collects 40 images of representative scenes from the KITTI 2015 training set. And KITTI (1) only consists of the first training image. As shown in Table 7, our models perform better than their counterparts with limited fine-tuning data.

5. Conclusion

We have introduced a feature consistency idea to improve the domain generalization performance of end-to-end stereo networks. We propose to explicitly impose a contrastive loss on learned features during training to maintain the consistency between stereo views. Then we restrict the intermediate feature representations with a selective whitening loss, which helps to maintain the feature consistency on unseen domains. Experimental results show that our approach significantly improves the generalization performance of end-to-end stereo matching networks.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No.61772057 and 62106012), Beijing Natural Science Foundation (4202039).

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] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011. 8

[3] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization. In *2020 International Conference on 3D Vision (3DV)*, pages 364–373. IEEE, 2020. 1, 3, 8

[4] 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, 6, 7, 8, 12, 13

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[6] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10639–10647, 2019. 3

[7] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 3, 5, 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[9] 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. 3

[10] Divyansh Garg, Yan Wang, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Wasserstein distances for stereo disparity estimation. *arXiv preprint arXiv:2007.03085*, 2020. 7

[11] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28:262–270, 2015. 3

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3

[13] 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. 6

[14] 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. 7

[15] 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, 6, 7, 8

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3, 4

[17] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 2, 8

[18] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2012. 8

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 5

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 5, 7

[21] 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. 1, 3

[22] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 573–590, 2018. 3

[23] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot domain adaptation with a physics prior. *arXiv preprint arXiv:2108.05137*, 2021. 1

[24] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 1

[25] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086*, 2017. 3

[26] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 1

[27] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Un-

berath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6197–6206, 2021. 8

[28] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2811–2820, 2018. 3

[29] Ping Luo. Learning deep architectures via generalized whitened neural networks. In *International Conference on Machine Learning*, pages 2238–2246. PMLR, 2017. 3

[30] 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, 3, 6, 7

[31] 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. 6

[32] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 1

[33] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1863–1871, 2019. 3

[34] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017. 4

[35] Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. *arXiv preprint arXiv:2011.05499*, 2020. 3, 4

[36] Matteo Poggi, Alessio Tonioni, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Continual adaptation for deep stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[37] 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. 6

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

[39] 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. 6

[40] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 1, 3, 8

[41] Jaime Spencer, Richard Bowden, and Simon Hadfield. Scale-adaptive neural dense features: Learning via hierarchical context aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6200–6209, 2019. 3, 4, 12

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

[43] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised domain adaptation for depth prediction from images. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2396–2409, 2019. 2

[44] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–204, 2019. 3

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

[46] Chen Wang, Xiang Wang, Jiawei Zhang, Liang Zhang, Xiao Bai, Xin Ning, Jun Zhou, and Edwin Hancock. Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition*, 124:108498, 2022. 3

[47] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3

[48] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime stereo image depth estimation on mobile devices. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5893–5900. IEEE, 2019. 3

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

[50] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3, 4

[51] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 3, 4

[52] 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. 3, 7

[53] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 7

[54] 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. 6

[55] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6044–6053, 2019. 3

[56] 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. 3

[57] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 1, 3, 6, 7, 8, 13

[58] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision*, pages 420–439. Springer, 2020. 1, 3, 5, 7, 8

[59] 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. 3

[60] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12926–12934, 2020. 3, 7