CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching

Zhelun Shen[1,2], Yuchao Dai[1]*, Zhibo Rao[1]
[1]Northwestern Polytechnical University, Xi'an, China [2]Peking University, Beijing, China
shenzhelun@pku.edu.cn, daiyuchao@nwpu.edu.cn, raoxi36@foxmail.com

Abstract

Recently, the ever-increasing capacity of large-scale annotated datasets has led to profound progress in stereo matching. However, most of these successes are limited to a specific dataset and cannot generalize well to other datasets. The main difficulties lie in the large domain differences and unbalanced disparity distribution across a variety of datasets, which greatly limit the real-world applicability of current deep stereo matching models. In this paper, we propose CFNet, a Cascade and Fused cost volume based network to improve the robustness of the stereo matching network. First, we propose a fused cost volume representation to deal with the large domain difference. By fusing multiple low-resolution dense cost volumes to enlarge the receptive field, we can extract robust structural representations for initial disparity estimation. Second, we propose a cascade cost volume representation to alleviate the unbalanced disparity distribution. Specifically, we employ a variance-based uncertainty estimation to adaptively adjust the next stage disparity search space, in this way driving the network progressively prune out the space of unlikely correspondences. By iteratively narrowing down the disparity search space and improving the cost volume resolution, the disparity estimation is gradually refined in a coarse-to-fine manner. When trained on the same training images and evaluated on KITTI, ETH3D, and Middlebury datasets with the fixed model parameters and hyperparameters, our proposed method achieves the state-of-the-art overall performance and obtains the 1st place on the stereo task of Robust Vision Challenge 2020. The code will be available at https://github.com/gallenszl/CFNet.

1. Introduction

Stereo matching, *i.e*. estimating a disparity/depth map from a pair of stereo images, is fundamental to various applications such as autonomous driving [3], robot navigation [1], SLAM [8, 10], *etc*. Recently, many deep learning-based stereo methods have been developed and achieved impressive performance on most of the standard benchmarks.

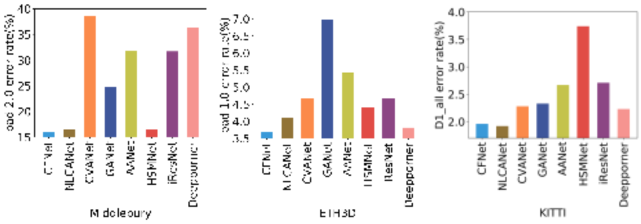
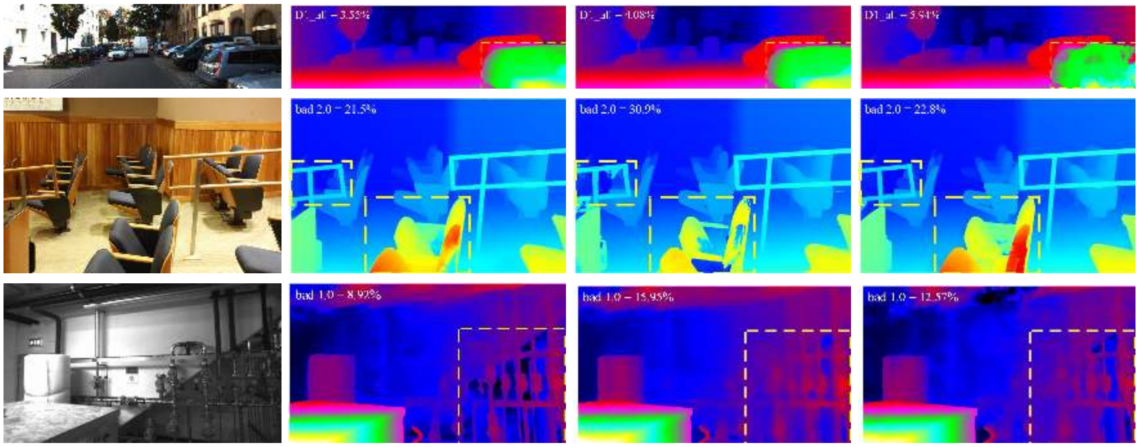


Figure 1. Performance comparison in terms of generalization ability on the Middlebury, ETH3D, and KITTI 2015 datasets. Bad 2.0, bad1.0, and D1_all (the lower the better) are used for evaluation. All methods are trained on the same training images and tested on three datasets with single model parameters and hyper-parameters. Our CFNet achieves state-of-the-art generalization and performs well on all three real-world datasets.

However, current state-of-the-art methods are generally limited to a specific dataset due to the significant domain shifts across different datasets. For example, the KITTI dataset [9, 19] focuses on real-world urban driving scenarios while Middlebury [22] concentrates on indoor high-resolution scenes. Consequently, methods that are state-of-the-art on one dataset often cannot achieve comparable performance on a different one without substantial adaptation (visualization comparison can be seen in Fig. 2). However, real-world applications require the approaches to generalize well to different scenarios without adaptation. Thus, we need to push methods to be robust and perform well across different datasets with the fixed model parameters and hyperparameters.

The difficulties in designing a robust stereo matching system come from the large domain differences and unbalanced disparity distribution between a variety of datasets. As illustrated in Fig. 2 (a), there are significant domain differences across various datasets, *e.g*., indoors vs outdoors, color vs gray, and real vs synthetic, which leads to the learned features distorted and noisy [37]. In addition, as illustrated in Fig. 3, the disparity range of half-resolution images in Middlebury [22] is even more than 6 times larger than full-resolution images in ETH3D [24] (400 vs 64).

*Yuchao Dai is the corresponding author.



(a) Left image (b) CFNet (c) GANet (d) HSMNet

Figure 2. Visualization of some state-of-the-art methods' performance on three real-world dataset testsets (from top to bottom: KITTI, Middlebury, and ETH3D). All methods are trained with a combination of KITTI, Middlebury, and ETH3D train images. GANet [36] and HSMNet [33] can achieve good performance on one specific dataset but perform poorly on the other two even if they have included targeted domain images in the training process. Our CFNet achieves SOTA or near SOTA performance on all three datasets without any adaptation.

Such unbalanced disparity distribution makes the current approaches trained with a fixed disparity range cannot cover the whole disparity range of another dataset without substantial adaption.

In this paper, we propose a cascade and fused cost volume representation to alleviate the above problems. (**1**) Towards the large domain differences, we propose to fuse multiple low-resolution dense cost volumes to enlarge the receptive field for capturing global and structural representations. Previous work [37] observes that the limited effective receptive field of convolutional neural networks [17] is the major reason the network is domain-sensitive to different datasets and proposes a learnable non-local layer to enlarge the receptive field. Inspired by it, we find that different scale low-resolution cost volumes can cover multi-scale receptive fields and are complementary to each other in promoting the network to look at different scale image regions. Thus, we can fuse multiple low-resolution dense cost volumes to guide the network to learn geometric scene information which is invariant across different datasets. In addition, this operation only adds a slight computation complexity. (**2**) Towards the unbalanced disparity distribution, we propose a cascade cost volume representation and employ a variance-based uncertainty estimation to adaptively adjust the next stage disparity search range. That is, our method only needs to cover the union of all datasets' disparity distribution (disparity range) at the first stage. Then we can employ our uncertainty estimation to evaluate pixel-level confidence of disparity estimation and prune out unlikely correspondences, guiding our network to look at more possible disparity search space at the next stage. In addition, we can save a lot of computational complexity by pruning out unlikely correspondences.

Experimentally, all methods are trained on the same training images and tested on three real-world datasets (KITTI2015, Middlebury, and ETH3D) with fixed model parameters and hyperparameters. As shown in Fig. 1, our method performs well on all three datasets and achieves state-of-the-art overall performance without adaptation.

In summary, our main contributions are:

- We propose a fused cost volume representation to reduce the domain differences across datasets.
- We propose a cascade cost volume representation and develop a variance-based uncertainty estimation to balance different disparity distributions across datasets.
- Our method shows great generalization ability and obtains the 1st place on the stereo task of Robust Vision Challenge 2020.
- Our method has great finetuning performance with low latency and ranks 1st on the popular KITTI 2015 and KITTI 2012 benchmarks among the published methods less than 200ms inference time.

2. Related Work

2.1. Cost Volume based Deep Stereo Matching

Stereo matching has been studied for decades and a well-known four-step pipeline [23] has been established, where cost volume construction is an indispensable step. Typically, a cost volume is a 4D tensor of height, width, disparity, and features. Current state-of-the-art stereo matching methods are all cost volume based methods and they can be categorized into two categories. The first category uses a

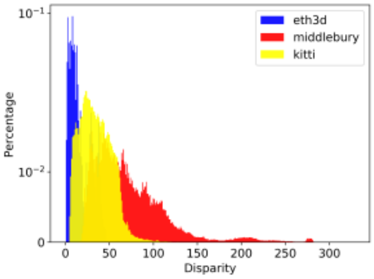


Figure 3. Disparity distribution of KITTI 2015, Middlebury, and ETH3D training sets. We plot disparity distribution of half-resolution images for Middlebury while full-resolution images for other datasets. The disparity distribution across different datasets is unbalanced.

full correlation to generate a single-feature 3D cost volume. Such methods are usually efficient but lose much information due to of the decimation of feature channels. Many previous work, including Dispnet [18], MADNet [29], and AANet [32], belong to this category. The second category usually uses concatenation [14] or group-wise correlation [12] to generate a multi-feature 4D cost volume. Such methods can achieve improved performance while requiring higher computational complexity and memory consumption. Actually, a majority of the top-performing networks in public leaderboards belong to this category, such as GANet [36] and CSPN [5]. These methods generally employ multiple 3D convolution layers to constantly regularize the 4D cost volume and then apply softmax over the disparity dimension to produce a discrete disparity probability distribution. The final predicted disparity is obtained by soft argmin [14], where the output is susceptible to multimodal disparity probability distributions. To address this issue, ACFNet [38] directly supervises the cost volume with unimodal ground truth distributions. In contrast, we define an uncertainty estimation to quantify the degree to which the cost volume tends to be multi-modal distribution, higher implies the higher possibility of estimation error.

2.2. Multi-scale Cost Volume based Deep Stereo Matching

Multi-scale cost volume was firstly applied in the single-feature 3D cost volume based network with the form of two-stage refinement [16] and pyramidal towers [29, 27]. Recently, cascade cost volume representation [34, 11, 4] was proposed in multi-view stereo to alleviate the high computational complexity and memory consumption in employing 4D cost volumes. These methods generally predict an initial disparity at the coarsest resolution. Then, they narrow down the disparity search space and gradually refine the disparity. Recently, Casstereo [11] extends such representation to stereo matching. It uniformly samples a predefined range to generate the next stage's disparity search range. Instead, we employ uncertainty estimation to adaptively adjust the next stage pixel-level disparity search range and push the next stage's cost volume to be predominantly unimodal. Our method also shares similarities with UCSNet [4], which constructs uncertainty-aware cost volume in multi-view stereo. However, it only focuses on dataset-specific performance. In addition, it generates the next stage search range with pre-defined parameters while we use learned parameters, which can better adapt to different datasets.

2.3. Robust Stereo Matching

Recently, researchers have shown an increased interest in robust stereo matching. These methods can be roughly categorized into two types. 1) Cross-domain Generalization: This category aims to improve the generalization of the network to unseen scenes. Towards this end, Jia *et al.* [30] propose to incorporate scene geometry priors into an end-to-end network. Zhang *et al.* [37] introduce a domain normalization and a trainable non-local graph-based filter to construct a domain-invariant stereo matching network. 2) Joint Generalization: This category aims to push the network to perform well on a variety of datasets with the same model parameter. MCV-MFC [16] introduces a two-stage finetuning scheme to achieve a good trade-off between generalization capability and fitting capability in a number of datasets. However, it doesn't touch the inner difference between diverse datasets. To further address this problem, we propose a cascade and fused cost volume representation to alleviate the domain shifts and disparity distribution unbalance between a variety of datasets. Additionally, our method also performs well on cross-domain generalization, which further emphasizes the robustness of our network.

3. Our Approach

3.1. Overview of CFNet

We propose a cascade and fused cost volume representation for robust stereo matching. The overall architecture of our model is shown in Fig. 4, which consists of three parts: pyramid feature extraction, fused cost volume, and cascade cost volume.

Given an image pair, we first employ a siamese unet-like [33, 21] encoder-decoder architecture with skip connections to extract multi-scale image features. The encoder consists of five residual blocks, followed by an SPP module to better incorporate hierarchical context information. Our SPP module is similar to the one used in HSMNet [33] while changing the size of average pooling blocks to $H/s \times W/s$, where $s \in \{32, 64, 96, 128\}$. Compared with the widely used Resnet-like network [11, 12], our method is more efficient and still contains sufficient information for cost aggregation. Experiments show that our pyramid feature extrac-

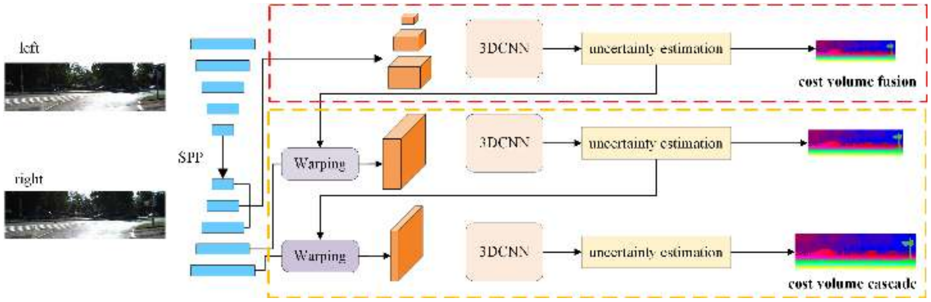


Figure 4. The architecture of our proposed network. Our network consists of 3 parts: pyramid feature extraction, fused cost volume, and cascade cost volume.

tion can achieve comparable performance with lower computational complexity. Then, we divide the multi-scale features into fused and cascade cost volume and predict multi-resolution disparity respectively. The details of cost volume fusion and cost volume cascade will be discussed in the next section.

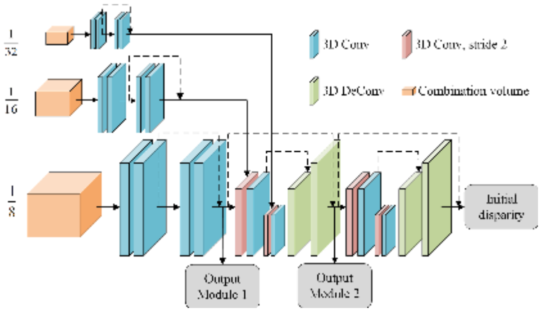


Figure 5. The architecture of our cost volume fusion module. Three low-resolution cost volumes ($i \in (3,4,5)$) are fused to generate the initial disparity map.

3.2. Fused Cost Volume

We propose to fuse multiple low-resolution dense cost volumes (smaller than 1/4 of the original input image resolution in our paper) to reduce the domain shifts between different datasets for initial disparity estimation. Existing approaches [11, 4, 34] have realized the importance of employing multi-scale cost volumes. However, these methods generally abandon small-resolution cost volumes because such cost volumes don't contain sufficient information to generate an accurate disparity map respectively. Instead, we argue that different scale small-resolution cost volumes can cover multi-scale receptive fields, guiding the network to look at different scale image regions. Thus, they can be fused together to extract global and structural representations and generate a more accurate initial disparity map than higher-resolution sparse cost volume. More precisely, we first construct low-resolution cost volumes at each scale respectively, and then design a cost volume fusion module to integrate them in an encoder-decoder process. We provide details about these two steps below.

Cost volume construction: Inspired by [26, 12], we propose to use both feature concatenation and group-wise correlation to generate combination volume. The combination volume is computed as:

$$V^i_{concat}(d^i,x,y,f) = f^i_L(x,y)||f^i_R(x-d^i,y)$$

$$V^i_{gwc}(d^i,x,y,g) = \frac{1}{N^i_c/N_g}\left\langle f^{ig}_l(x,y), f^{ig}_r(x-d^i,y)\right\rangle \qquad (1)$$

$$V^i_{combine} = V^i_{concat}||V^i_{gwc}$$

where $||$ denotes the vector concatenation operation. N_c represents the channels of extracted feature. N_g is the amount of group. $\langle,\rangle$ represents the inner product. f^i denotes the extracted feature at scale (stage) i and $i=0$ represents the original input image resolution.

Note that the disparity searching index d^i is defined as $d^i \in \{0,1,2\ldots\frac{D_{\max}}{2^i}-1\}$ and the hypothesis plane interval equals to 1 in the fused cost volume representation. That is, these cost volumes are all dense cost volumes with the size of $\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D_{\max}}{2^i} \times F$. By densely sampling the whole disparity range in small resolution, we can efficiently generate a coarsest disparity map. Then we can employ variance-based uncertainty estimation to narrow down the disparity searching space at higher resolution and refine the disparity estimation in a coarse-to-fine manner. Details will be introduced in Section 3.3.

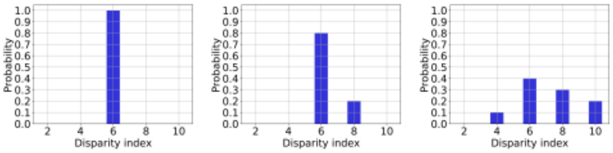
Cost Volume fusion: Following the method proposed in [26], we use an improved encoder-decoder architecture to fuse low-resolution cost volumes. The architecture is shown in Fig. 5. Specifically, we first employ four 3D convolution layers with skip connections to regularize each cost volume and use a 3D convolution layer (stride of two) to down-sample the combination volume of scale 3 from 1/8 to 1/16 of the input image resolution. Next, we concatenate them (the down-sampled cost volume and the next stage combination volumes) at the feature dimension and then decrease the feature channel to a fixed size via one additional 3D convolution layer. Then, we apply a similar operation to progressively down-sample the cost volume to 1/32 of the original input image resolution and adopt 3-D transposed convolution to up-sample the volume in the decoder. In addition, we utilize one 3-D hourglass network to further regularize and refine the volume. Finally, an output module is applied to predict the disparity. The output module contains two more 3D convolution layers, aiming at obtaining a 1-channel 4D volume. To transform volume into disparity, we apply soft argmin [14] operation to generate initial disparity map D^3. The soft argmin operation is defined as:

$$\widehat{d^i} = \sum_{d=0}^{\frac{D_{\max}}{2^i}-1} d \times \sigma(-c_d^i), \tag{2}$$

where σ denotes the softmax operation and c represents the predicted 1-channel 4D volume. $\sigma(-c_d)$ denotes the discrete disparity probability distribution and the estimated disparity map is susceptible to all disparity indexes.

3.3. Cascade Cost Volume

Given the initial disparity estimation, the next step is to construct a fine-grained cost volume and refine disparity maps in a coarse-to-fine manner. Considering the next stage disparity search range, uniform sampling a pre-defined range is the most straightforward way [11]. However, such an approach assumes that all pixels are the same and cannot make adaptive pixel-level adjustments. For example, we should expand the searching range of ill-posed and occluded pixels. Furthermore, the disparity distribution of datasets with different characteristics is usually unbalanced. Thus, a question arises, can we make networks avoid being affected by invalid disparity indexes in a large initial disparity range and capture more possible pixel-level disparity search space with the prior knowledge of the last stage's disparity estimation.



$\hat{d} = 6.0, U = 0.0$ (a)Unimodal

$\hat{d} = 6.4, U = 0.64$ (b)Predominantly unimodal

$\hat{d} = 7.2, U = 3.36$ (c)Multi-modal

Figure 6. Some samples of uncertainty estimation. Expected value (ground truth) is 6px. The disparity searching range is from 2 to 10 with 5 hypothesis planes.

To tackle this problem, we propose an adaptive variance-based disparity range uncertainty estimation. As introduced in related work, discrete disparity probability distribution reflects the similarities between candidate matching pixel pairs and the final predicted disparity is a weighted sum of all disparity indexes according to their probability. Thus, the ideal disparity probability distribution should be unimodal peaked at true disparities. However, the actual probability distribution is predominantly unimodal or even multimodal at some pixels. Previous work [38, 14] has discovered that the degree of multimodal distribution is highly correlated with the probability of prediction error. In addition, ill-posed areas, texture-less regions, and occlusions tend to be multimodal distribution as well as high estimation error rate. Therefore, we propose to define an uncertainty estimation to quantify the degree to which the cost volume tends to be multi-modal distribution and employ it to evaluate the pixel-level confidence of the current estimation. The uncertainty is defined as:

$$\begin{aligned} \mathrm{U}^i &= \sum_{\forall d^i} \left(d - \hat{d}^i\right)^2 \times \sigma(-c_d^i), \\ \widehat{d^i} &= \sum_{\forall d^i} d \times \sigma(-c_d^i), \end{aligned} \tag{3}$$

where σ denotes the softmax operation and c represents the predicted 1-channel 4D volume. As shown in Fig. 6, the uncertainty of unimodal distribution equals to 0 and the more the probability distribution is toward the multi-modal distribution, the higher the uncertainty and error. Therefore, it's reasonable to employ uncertainty to evaluate the confidence of disparity estimation, higher uncertainty implies a higher probability of prediction error and a wider disparity searching space to correct the wrong estimation (visualization can be seen in Fig. 8). Thus, the next stage's disparity searching range is defined as:

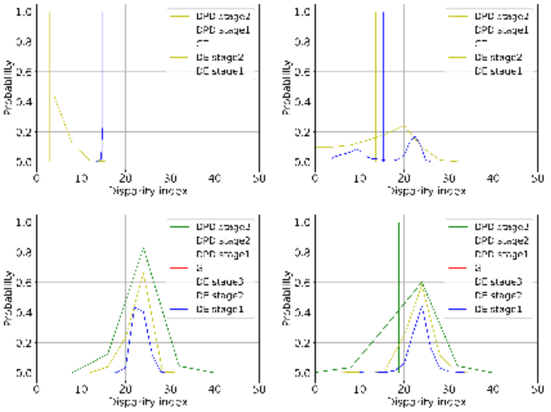
$$\begin{aligned} d_{\max}^{i-1} &= \delta(\widehat{d^i} + \left(\alpha^i + 1\right)\sqrt{U^i} + \beta^i), \\ d_{\min}^{i-1} &= \delta(\widehat{d^i} - \left(\alpha^i + 1\right)\sqrt{U^i} - \beta^i), \end{aligned} \tag{4}$$

where δ denotes bilinear interpolation. α and β are normalization factors, which is initialized as 0 and gradually learns a weight. α and β can also be set as hyper parameters, while experiment shows that the learned parameters are more robust than human-selected parameters [4]. Then we can employ uniform sampling to get next stage discrete hypothesis depth planes d^{i-1}:

$$\begin{aligned} d^{i-1} &= d_{\min}^{i-1} + n(d_{\max}^{i-1} - d_{\min}^{i-1}) / \left(N^{i-1} - 1\right), \\ n &\in \{0, 1, 2 \ldots N^{i-1} - 1\}, \end{aligned} \tag{5}$$

where N^{i-1} is the number of hypothesis planes at stage $i-1$. Then, a fine-grained cost volume at stage $i-1$ is similarly defined based on Eq.1, which leads to a sparse cost volume with the size of $\frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}} \times N^{i-1} \times F$. After getting the next stage cost volume, similar cost aggregation network (omitting the solid line in Figure 5) can be employed to predict this stage's disparity map. By iteratively narrow down the disparity range and higher the cost volume resolution, we can refine the disparity in a coarser to fine manner.

While it is noteworthy that the insight of cascade cost volume has also been investigated in [11], our work differs from theirs in the following three main aspects: First, casstereo constructs higher-resolution sparse cost volume to predict an initial disparity estimation. Besides, previous work [34] argues that small resolution cost volume is too small to generate a reasonable initial disparity map to be refined. In contrast, we prove such cost volumes can even generate a more accurate initial disparity map than higher-resolution sparse cost volume with cost volume fusion. (As shown by the comparison between the estimation of casstereo at stage 2 and estimation of our CFNet at stage 3 in Fig. 7 (a) and (b)). Second, casstereo only uniform sampling a pre-defined range to generate the next stage disparity search range. Instead, we develop a variance-based uncertainty estimation to adaptively adjust the next stage disparity search range which can push disparity distribution to be more predominantly unimodal (Fig. 7(b)). Third, by the cooperation of cascade and fused cost volume representation, our method can better cover the corresponding ground truth value in the final stage disparity search range and corrects some biased results in casstereo (Fig. 7(a)).



(a) (b)

Figure 7. Some examples of disparity probability distributions of cascade stereo (first row) and CFNet (second row). DPD: disparity probability distribution, DE: disparity estimation, GT: ground truth. Final disparity estimation is the estimation of stage 1.

4. Experiments

4.1. Dataset

SceneFlow: This is a large synthetic dataset including 35,454 training and 4,370 test images with a resolution of 960×540 for optical flow and stereo matching. We use it to pre-train our network.

Middlebury: Middlebury [22] is an indoor dataset with 28 training image pairs (13 of them are additional training images) and 15 testing image pairs with full, half, and quarter resolutions. It has the highest resolution among the three datasets and the disparity range of half-resolution image pairs is 0-400. We select half-resolution image pairs to train and test our model.

KITTI 2012&2015: They are both real-world datasets collected from a driving car. KITTI 2015 [19] contains 200 training and another 200 testing image pairs while KITTI 2012 [9] contains 194 training and another 195 testing image pairs. Both training image pairs provide sparse ground-truth disparity and the disparity range of them is 0-230.

ETH3D: ETH3D [24] is the only grayscale image dataset with both indoor and outdoor scenes. It contains 27 training and 20 testing image pairs with sparsely labeled ground-truth. It has the smallest disparity range among three datasets, which is just in the range of 0-64.

4.2. Implementation Details

We use PyTorch to implement our 3-stage network and employ Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) to train the whole network in an end-to-end way. The batch size is set to 8 for training on 2 Tesla V100 GPUs and the whole disparity search range is fixed to 256 during the training and testing process. We employ the smooth L1 loss function to train our network and include all intermediate outputs in the loss weight. N^1 and N^2 are set as 12 and 16, respectively. Asymmetric chromatic augmentation and asymmetric occlusion [33] are employed for data augmentation.

Inspired by the two-stage finetuning strategy [16], we propose a three-stage finetune strategy to train our network. First, following the method proposed in [26], we use switch training strategy to pre-train our model in the SceneFlow dataset. Specifically, we first use ReLU to train our network from scratch for 20 epochs, then we switch the activation function to Mish and prolong the pre-training process in the SceneFlow dataset for another 15 epochs. Second, we jointly finetune our pre-train model on four datasets, i.e., KITTI 2015, KITTI2012, ETH3D, and Middlebury for 400 epochs. The initial learning rate is 0.001 and is down-scaled by 10 after epoch 300. Third, we augment Middlebury and ETH3D to the same size as KITTI 2015 and finetune our model for 50 epochs with a learning rate of 0.0001. The core idea of our three-stage finetune strategy is to prevent the small datasets from being overwhelmed by large datasets. By augmenting small datasets at stage three and train our model with a small learning rate, our strategy makes a better trade-off between generalization capability and fitting capability on three datasets.

4.3. Ablation Study

We perform various ablation studies to show the effectiveness of each component in our network. We divide 20% of the smallest dataset (5 images) from each real dataset (KITTI 2015, Middlebury, and ETH3D) as a validation set and use the rest of them as a training set to finetune our

Experiment	Method	KITTI D1_all	Middlebury bad 2.0	ETH3D bad 1.0	time (s)
Feature Extraction	Resnet-like-network	1.76	22.81	**3.49**	0.270
	Pyramid Feature Extraction	**1.71**	**22.27**	3.57	**0.225**
Cost Volume Fusion	Not Fuse	1.79	22.65	3.67	**0.220**
	Fuse	**1.71**	**22.27**	**3.57**	0.225
Cost Volume Cascade	Uniform Sample	1.92	23.8	3.97	0.225
	UE + Hyperparameters	1.78	23.13	3.83	0.225
	UE + Learned Parameters	**1.71**	**22.27**	**3.57**	**0.225**
Loss weight	Loss1:1,Loss2:1,Loss3:1	1.84	23.87	**3.47**	0.225
	Loss1:1,Loss2:0.7,Loss3:0.5	1.79	22.97	3.62	0.225
	Loss1:2,Loss2:1,Loss3:0.5	**1.71**	**22.27**	3.57	**0.225**
Fine-tuning strategy	two stages	**1.70**	22.77	3.99	0.234
	three stages_no augment	**1.70**	22.57	3.92	0.234
	three stages	1.71	**22.27**	**3.57**	**0.234**

Table 1. Ablation study results of the proposed network on KITTI 2015, Middlebury, and ETH3D validation set. UE: uncertainty estimation. Loss i: the loss weight at stage i. j stages: j-stage finetune strategy. Three stages_no augment: three stages without small dataset augment.We test a component of our method individually in each section of the table and the approach which is used in our final model is underlined. Time is measured on the KITTI dataset by a single Tesla V100 GPU.

pretrain model. Results are shown in Table 1. Below we describe each component in more detail.

Feature extraction: We compare our pyramid feature extraction with the most widely used Resnet-like-network [11, 12]. As shown, our pyramid feature extraction can achieve similar performance with a faster speed, likely because the employing of small scale features is also helpful in feature extraction.

Cost volume fusion: We fuse three small-resolution cost volumes to generate the initial disparity map. Here, we test the impact when only a single volume is used. Cost volume fusion can achieve better performance with a slight additional computational cost.

Cost volume cascade: We test three ways of generating the next stage's disparity searching space in cascade cost volume representation. As shown, learned parameters based uncertainty estimation achieves the best performance with tiny additional computation complexity. Furthermore, comparing with using hyperparameters, it is adaptive and eliminates the need to adjust hyperparameters according to different datasets. To further emphasize the effectiveness of our uncertainty estimation, we visualize the error map and uncertainty map in Fig. 8. As shown, the error map is highly correlated with the uncertainty map. Experimentally, by removing 1% of uncertain pixels ($\sqrt{U} >= 2.5$), we decrease the D1_all error rate by 29.68% (from 1.55% to 1.09% in KITTI 2015 validation set). Similar situations can be observed on the other two datasets. Thus, it is reasonable to employ uncertainty estimation to evaluate the pixel-level confidence of disparity estimation.

Loss weight: The N-stages model outputs N disparity maps. We test three different settings of loss weight in our three-stage model. Experiments show that later stages shall set a larger loss weight which matches our intuition.

Finetuning strategy: We test three terms of finetuning strategy. As shown, only extending the number of iterations can not improve the accuracy of predictions on small datasets. Instead, our strategy can greatly alleviate the problem of small datasets being overwhelmed by large ones.

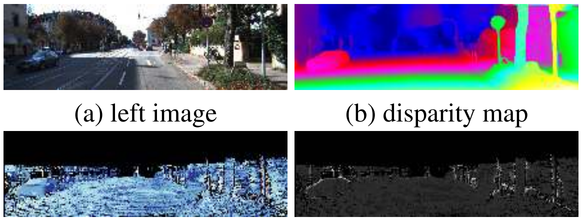

(a) left image
(b) disparity map
(c) error map
(d) uncertainty map

Figure 8. Comparison between error map and uncertainty map. Red and white denote large errors and high uncertainty, respectively. As shown, the error map is highly correlated with the uncertainty map. See examples of the other two datasets in the supplementary material.

4.4. Robustness Evaluation

As mentioned before, we define robustness as joint generalization. Such generalization is essential for current methods, which are limited to specific domains and cannot get comparable results on other datasets. This is also the goal of Robust Vision Challenge 2020[†]. Towards this end, we evaluate methods' robustness by their performance on three real datasets without finetuning.

We list some state-of-the-art methods' performance in Table 2. It can be seen from this table that HSMNet_RVC [33] ranks first on the Middlebury dataset. But it can't get comparable results on the other two datasets (3rd on ETH3D and 6th on KITTI 2015). In particular, its performance on KITTI 2015 dataset is far worse than the other five. This is because this method is specially designed for high-resolution datasets and can't generalize well on other datasets. GANet [36] is the top-performing method in the KITTI dataset. However, the error rate of D1_all increased by 28.73% (from 1.81% to 2.33%) after adding training images of the other two datasets and only ranks 4th on KITTI 2015. In addition, it still cannot get a good result on the other two datasets (6th on ETH3D and 7th on Middlebury). The similar situation also appeared on other methods. In contrast, our method shows great generalization ability and performs well on all three datasets (2nd on KITTI 2015, 1st on ETH3D, and 2nd on Middlebury) and achieves the best overall performance. We also compare our methods with the top two methods in the previous Robust Vision Challenge. As shown in Table 2, our approach outperforms

[†] http://www.robustvision.net/

Method	KITTI				Middlebury				ETH3D				Overall
	D1_bg	D1_fg	D1_all	Rank	bad 1.0	bad 2.0	avg error	Rank	bad 1.0	bad 2.0	avg error	Rank	Rank
NLCANet_V2_RVC [20]	**1.51**	3.97	**1.92**	**1**	29.4	16.4	5.60	3	4.11	1.2	0.29	2	2
HSMNet_RVC [33]	2.74	8.73	3.74	6	31.2	16.5	**3.44**	**1**	4.40	1.51	0.28	3	3
CVANet_RVC	1.74	4.98	2.28	3	58.5	38.5	8.64	5	4.68	1.37	0.34	4	4
AANet_RVC [32]	2.23	4.89	2.67	5	42.9	31.8	12.8	6	5.41	1.95	0.33	5	5
GANet_RVC [36]	1.88	4.58	2.33	4	43.1	24.9	15.8	7	6.97	1.25	0.45	6	6
CFNet_RVC(ours)	1.65	**3.53**	1.96	2	**26.2**	**16.1**	5.07	2	**3.7**	**0.97**	**0.26**	**1**	**1**
iResNet_ROB [15, 16]	2.27	4.89	2.71	3	45.9	31.7	6.56	2	4.67	1.22	0.27	3	3
Deeppruner_ROB [7]	-	-	2.23	2	57.1	36.4	6.56	3	3.82	1.04	0.28	2	2
CFNet_RVC(ours)	**1.65**	**3.53**	**1.96**	**1**	**26.2**	**16.1**	**5.07**	**1**	**3.7**	**0.97**	**0.26**	**1**	**1**

Table 2. Joint generalization comparison on ETH3D, Middlebury, and KITTI2015 datasets. **Top:** Generalization comparison with methods who participated in the Robust Vision challenge 2020. **Bottom:** Generalization comparison with the top 2 methods in the past two years. All methods are tested on three datasets without adaptation. The overall rank is obtained by Schulze Proportional Ranking [25] to joining multiple rankings into one. As shown, our method achieves the best overall performance.

Method	KITTI2012 D1_all(%)	KITTI2015 D1_all(%)	Middlebury bad 2.0(%)	ETH3D bad 1.0(%)
PSMNet [2]	15.1	16.3	39.5	23.8
GWCNet [12]	12.0	12.2	37.4	11.0
CasStereo [11]	11.8	11.9	40.6	7.8
GANet [36]	10.1	11.7	32.2	14.1
DSMNet [37]	6.2	6.5	**21.8**	6.2
CFNet	**4.7**	**5.8**	28.2	**5.8**

Table 3. Cross-domain generalization evaluation on ETH3D, Middlebury, and KITTI training sets. All methods are only trained on the Scene Flow datatest and tested on full-resolution training images of three real datasets.

Deeppruner_ROB and iResNet_ROB on all three datasets with a remarkable margin.

We also notice that some previous work defines robustness as cross-domain generalization. To further emphasize the effectiveness of our method, we compare our method with some state-of-the-art methods by training on synthetic images and testing on real images. As shown in Table 3, our method far outperforms domain-specific methods [2, 12, 11, 36] on all four datasets. DSMnet [37] is specially designed for cross-domain generalization. Our method can surpass it on three datasets, which further shows our cascade and fused cost volume representation is an efficient approach for robust stereo matching.

4.5. Results on KITTI Benchmark

Although our focus is not on domain-specific performance, we still fine-tune our model on KITTI 2012 and KITTI 2015 benchmark to show the efficiency of our method. Specifically, we adjust each stage's (except stage 3) stack hourglass number to be one and pretrain our model on Scene Flow datasets with the same training strategy. Then the pre-train model is finetuned on KITTI 2015 and KITTI 2012 datasets for 300 epochs, respectively. The learning rate starts at 0.001 and decreases to 0.0001 after 200 epochs. Following [7], we combine KITTI 2012 and KITTI 2015 image pairs for the evaluation of KITTI 2015 while only use KITTI 2012 image pairs for the evaluation of KITTI 2012.

Some SOTA real-time methods and best-performing approaches are listed in Table 4. We find that our method achieves 1.88% D1_all error rate, a 6% error reduction from our base model casstereo [11] with 3 times faster speed and gets similar performance with other best-performing approaches such as GANet-deep [36] and ACFNet [38]. Furthermore, our method outperforms all published methods faster than 200ms with a noteworthy margin on both datasets which implies the efficiency of our method.

Method	KITTI2012 3px(%)		KITTI2015 D1_all(%)		time (s)
	Noc	All	Noc	All	
LEAStereo [6]	**1.13**	**1.45**	**1.51**	**1.65**	0.3
GANet-deep [36]	1.19	1.60	1.63	1.81	1.8
AcfNet [38]	1.17	1.54	1.72	1.89	0.48
Casstereo [11]	-	-	1.78	2.0	0.6
HITNet [28]	1.41	1.89	1.74	1.98	**0.015**
HDˆ3 [35]	1.40	1.80	1.87	2.02	0.14
AANet+ [32]	1.55	2.04	1.85	2.03	0.06
HSMNet [33]	1.53	1.99	1.92	2.14	0.14
Deeppruner [7]	-	-	1.95	2.15	0.18
CFNet(ours)	**1.23**	**1.58**	**1.73**	**1.88**	0.18

Table 4. Results on KITTI benchmark. **Top:** Comparison with best-performing methods. **Bottom:** Comparison with real-time methods. All methods are finetuned on specific datasets.

5. Conclusion

We have proposed a cascade and fused cost volume representation for robust stereo matching. We first introduce a fused cost volume to alleviate the domain shifts across different datasets for initial disparity estimation. Then we construct cascade cost volume to balance the different disparity distribution across datasets, where the variance-based uncertainty estimation is at the core. We use it to adaptively narrow down the next stage's pixel-level disparity searching space. Experiment results show that our approach performs well on a variety of datasets with high efficiency. In the future, we plan to extend our cost volume representation to semi-supervised or self-supervised setup [31, 13, 39].

Acknowledgements This research was supported in part by National Key Research and Development Program of China (2018AAA0102803) and National Natural Science Foundation of China (61871325). We would like to thank the anonymous area chairs and reviewers for their useful feedback.

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, page 21, 2011. 1

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 8

[3] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2722–2730, 2015. 1

[4] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2524–2534, 2020. 3, 4, 5

[5] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. In *IEEE Transactions on pattern analysis and machine intelligence (TPAMI)*, 2019. 3

[6] 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 (NIPS)*, 33, 2020. 8

[7] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4384–4393, 2019. 8

[8] Jakob Engel, Jörg Stückler, and Daniel Cremers. Large-scale direct slam with stereo cameras. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 1935–1942, 2015. 1

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 1, 6

[10] Ruben Gomez-Ojeda, Francisco-Angel Moreno, David Zuñiga-Noël, Davide Scaramuzza, and Javier Gonzalez-Jimenez. Pl-slam: A stereo slam system through the combination of points and line segments. *IEEE Transactions on Robotics*, 35(3):734–746, 2019. 1

[11] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, 2020. 3, 4, 5, 6, 7, 8

[12] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3273–3282, 2019. 3, 4, 7, 8

[13] Rongrong Ji, Ke Li, Yan Wang, Xiaoshuai Sun, Feng Guo, Xiaowei Guo, Yongjian Wu, Feiyue Huang, and Jiebo Luo. Semi-supervised adversarial monocular depth estimation. *IEEE transactions on pattern analysis and machine intelligence(TPAMI)*, 42(10):2410–2422, 2019. 8

[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 *IEEE International Conference on Computer Vision (ICCV)*, pages 66–75, 2017. 3, 5

[15] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2811–2820, 2018. 8

[16] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. Stereo matching using multi-level cost volume and multi-scale feature constancy. In *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2019. 3, 6, 8

[17] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pages 4898–4906, 2016. 2

[18] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 3

[19] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 1, 6

[20] Zhibo Rao, Mingyi He, Yuchao Dai, Zhidong Zhu, Bo Li, and Renjie He. Nlca-net: a non-local context attention network for stereo matching. *APSIPA Transactions on Signal and Information Processing (ATSIP)*, 9:1–13, 2020. 8

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015. 3

[22] 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 (GCPR)*, pages 31–42, 2014. 1, 6

[23] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision (IJCV)*, 47(1-3):7–42, 2002. 2

[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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3260–3269, 2017. 1, 6

[25] Markus Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner

election method. *Social Choice and Welfare*, 36(2):267–303, 2011. 8

[26] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Msmd-net: Deep stereo matching with multi-scale and multi-dimension cost volume. *arXiv preprint arXiv:2006.12797*, 2020. 4, 6

[27] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, 2018. 3

[28] Vladimir Tankovich, Christian Häne, Sean Fanello, Yinda Zhang, Shahram Izadi, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. *arXiv preprint arXiv:2007.12140*, 2020. 8

[29] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 195–204, 2019. 3

[30] Jialiang Wang, Varun Jampani, Deqing Sun, Charles Loop, Stan Birchfield, and Jan Kautz. Improving deep stereo network generalization with geometric priors. *arXiv preprint arXiv:2008.11098*, 2020. 3

[31] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J. Brostow, and Michael Firman. Learning stereo from single images. In *European Conference on Computer Vision (ECCV)*, pages 722–740, 2020. 8

[32] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020. 3, 8

[33] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, 2019. 2, 3, 6, 7, 8

[34] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4877–4886, 2020. 3, 4, 6

[35] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6044–6053, 2019. 8

[36] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 185–194, 2019. 2, 3, 7, 8

[37] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Europe Conference on Computer Vision (ECCV)*, pages 420–439, 2020. 1, 2, 3, 8

[38] 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 (AAAI)*, pages 12926–12934, 2020. 3, 5, 8

[39] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *CoRR*, abs/1709.00930, 2017. 8