PCW-Net: Pyramid Combination and Warping Cost Volume for Stereo Matching

Zhelun Shen[1], Yuchao Dai[2†], Xibin Song[1†], Zhibo Rao[2], Dingfu Zhou[1] and Liangjun Zhang[1]

[1] Robotics and Autonomous Driving Lab, Baidu Research, China
[2] Northwestern Polytechnical University
{shenzhelun, song.sducg, dingfuzhou}@gmail.com; daiyuchao@nwpu.edu.cn; raoxi36@foxmail.com; zhangliangjun@baidu.com;

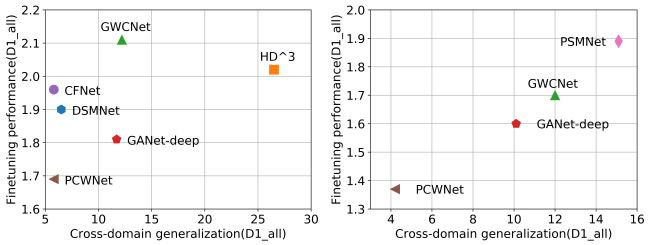
Abstract. Existing deep learning based stereo matching methods either focus on achieving optimal performances on the target dataset while with poor generalization for other datasets or focus on handling the cross-domain generalization by suppressing the domain sensitive features which results in a significant sacrifice on the performance. To tackle these problems, we propose PCW-Net, a **P**yramid **C**ombination and **W**arping cost volume-based network to achieve good performance on both cross-domain generalization and stereo matching accuracy on various benchmarks. In particular, our PCW-Net is designed for two purposes. First, we construct combination volumes on the upper levels of the pyramid and develop a cost volume fusion module to integrate them for initial disparity estimation. Multi-scale receptive fields can be covered by fusing multi-scale combination volumes, thus, domain-invariant features can be extracted. Second, we construct the warping volume at the last level of the pyramid for disparity refinement. The proposed warping volume can narrow down the residue searching range from the initial disparity searching range to a fine-grained one, which can dramatically alleviate the difficulty of the network to find the correct residue in an unconstrained residue searching space. When training on synthetic datasets and generalizing to unseen real datasets, our method shows strong cross-domain generalization and outperforms existing state-of-the-arts with a large margin. After fine-tuning on the real datasets, our method ranks 1^{st} on KITTI 2012, 2^{nd} on KITTI 2015, and 1^{st} on the Argoverse among all published methods as of 7, March 2022.

Keywords: Stereo Matching, Pyramid Cost Volume, Cross-domain generalization

1 Introduction

Stereo matching aims to estimate the disparity map between a rectified image pair, which contributes to various applications, such as autonomous driving [3]

† Corresponding authors



(a) KITTI 2015 (b) KITTI 2012

Fig. 1: **Model generalization ability vs fine-tuning performance on KITTI 2012&2015 datasets.** X-axis: all methods are trained on synthetic datasets and then tested on KITTI training sets to evaluate the cross-domain generalization. Y-axis: all methods are finetuned on the KITTI training sets and then tested on KITTI testing sets to evaluate the fine-tuning performance. D1_all is used for evaluation (the lower the better) and PCWNet is our method, which achieves the best overall performance.

and robotics navigation [1]. Benefiting from the unprecedented development of deep learning technologies, remarkable progress has been achieved in solving the task of stereo matching.

To achieve remarkable stereo matching performance, approaches [21] are usually trained on large-scale synthetic datasets (e.g., SceneFlow [16]) first and then fine-tuned on limited target dataset collected from the real scenarios such as KITTI [8], Middlebury [23], and ETH3D [24]. By extracting representative features [2,19] and constructing powerful cost volume [13,11], these methods achieve state-of-the-art performances on most of the standard stereo matching benchmarks. However, their performance decreases dramatically on unseen real-world scenes due to the large domain gaps across different datasets. Furthermore, these methods even cannot achieve consistent fine-tuning performances on different real-world datasets from similar scenarios. For example, some methods [2,34] perform well on the KITTI datasets [17,8], while having limited performances on the Argoverse benchmark [29] with high image resolutions though both of them are collected by a driving vehicle in the traffic environment.

Meanwhile, many approaches [35,25,28] are also specifically designed to handle domain generalization issues in stereo matching which aims to improve the generalization of the network to unseen scenes. By incorporating geometry priors and extracting domain-invariant features, these methods show strong cross-domain generalization when trained on synthetic datasets and generalized to unseen real datasets. However, such methods [35] normally need a significant sacrifice on accuracy to improve the cross-domain generalization due to the filtration of domain-sensitive features. Thus, a key problem for further research is designing a framework that can achieve excellent performances on the target dataset and also have satisfactory generalization ability to novel scenarios.

To relieve the issue mentioned above, we introduce the PCW-Net to construct a **P**yramid **C**ombination and **W**arping cost volume to hit two birds with one stone for achieving both generalization ability and good performance. Specifically, we use the pyramid cost volumes for two purposes. On one hand, we construct multi-scale combination volumes on the upper levels of the pyramid

and develop a cost volume fusion module to integrate them for initial disparity estimation. The pyramid cost volume aims to cover multi-scale receptive fields and boost the network to see different scale regions of the original image. Thus, multi-level information can be fused together, i.e., textures, contours, and areas. Typically, non-local information (such as contours and area) is more robust to domain changes, thus better performance and generalization ability for different resolutions of images can be obtained. On the other hand, we also construct a 3D warping volume at the final level of the pyramid to further refine the initial disparity map. With the constructed 3D warping volume, we can narrow down the residue searching range from an initial disparity searching range to a fine-grained one, which can dramatically alleviate the difficulty of the network to find the correct residue in an unconstrained residue searching space.

To prove the effectiveness of the proposed PCW-Net, we perform extensive experimental evaluations on various benchmarks to verify its fine-tuning performance and generalization ability. When trained on synthetic datasets and generalized to unseen real-world datasets, PCW-Net shows strong cross-domain generalization and outperforms best prior work [25] by a noteworthy margin. After fine-tuning on the real dataset, our method can achieve consistent SOTA performance across diverse datasets. Specifically, it ranked first on KITTI 2012 leaderboard[1], second on KITTI 2015 leaderboard, and first on Argoverse leaderboard[2] [29] among all published methods as of 6 March 2021. As demonstrated in Fig. 1, our method can achieve the best overall performance when considering both the fine-tuning accuracy and cross-domain generalization on the KITTI 2015 benchmark.

Our main contributions can be summarized as:

- An effective framework, *i.e.*, PCW-Net, is proposed which achieves remarkable generalization ability from synthetic dataset to real dataset while also excellent performances on the various target benchmarks after model fine-tuning.
- A novel multi-scale cost volume fusion module is proposed to cover multi-scale receptive fields and extract domain-invariant structural cues, thus better stereo matching performance of different resolutions of images is achieved.
- An efficient warping volume-based disparity refinement module is proposed to narrow down the unconstrained residue searching space to a fine-grained one, which can dramatically alleviate the difficulty of the network to find the correct residue in an unconstrained residue searching space.
- The proposed PCW-Net set new SOTA performance on both KITTI 2012 and Argoverse leaderboards among all the methods with publications, while it also achieves the 2^{nd} on the KITTI 2015 benchmark.

2 Related Work

Cost Volume based Deep Stereo Matching. DispNet [16] first introduces the concept of cost volume (correlation volume) into end-to-end stereo matching

[1] http://www.cvlibs.net/datasets/kitti
[2] https://eval.ai/web/challenges/challenge-page/917/leaderboard/2412

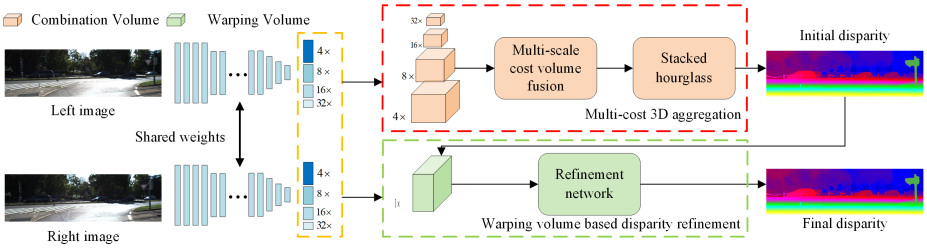


Fig. 2: General Structure of the proposed PCW-Net, which consists of three main modules as multi-scale feature extraction, multi-scale combination volume based cost aggregation, and warping volume based disparity refinement.

methods. Following this work, GCNet [13] proposes to construct concatenation volume and regularize it with 3D convolution layers and GwcNet [11] introduces group-wise correlation to provide better similarity measures. For all these prior works, cost volume construction has been placed in an extremely important position and deserves further exploration.

Deep Stereo Matching with Disparity Refinement. Recently, many researchers [20,27,26,14,7,36,38] attempt to integrate the disparity refinement step into an end-to-end model. [20] introduces a two-stage network called CRL in which the first stage extends DispNet [16] to get an initial disparity map and the second stage refines the initial disparity map in a residual manner. MCV-MFC [14] proposes to calculate reconstruction error in feature space rather than color space and share features between disparity estimation network and refinement network. PWCNet [26] proposes a context network, which is based on dilated convolutions to refine flow. However, existing methods mainly depend on the fitting capabilities of the networks to directly regress a residue with context information. Different from these works[20,14], here we introduce the warping volume to guide the disparity refinement. Specifically, the warping volume is constructed by warped right image features and left image features according to a pre-defined residue range. That is the warping volume narrows down the residue searching space from initial disparity searching space to a fine-grained one, which makes the network easier to find the corresponding pixel-level residue.

Multi-scale-based Deep Stereo Matching. Multi-scale information has been widely employed in deep stereo matching methods. These methods can be roughly categorized into two types: (1) The first category [2,14,19] usually employs a multi-scale feature extraction network to generate feature maps at different scales and then fuse them to construct a single volume at a fixed resolution. That is these methods mainly use multi-scale features rather than multi-scale cost volumes. (2) The second category [32,10,4,15] proposes to construct cascade pyramid cost volume and progressively regress a high-quality disparity map from the coarsest cost volume. That is these methods employ each scale cost volume to estimate disparity maps separately. Different from the former two categories, our work selects to directly fuse multi-scale combination volumes to capture a more robust feature representation for initial disparity estimation. Then, we employ warping volume to further refine the initial disparity. More related to our work is SSPCV [30], which also proposes a cost volume fusion module. How-

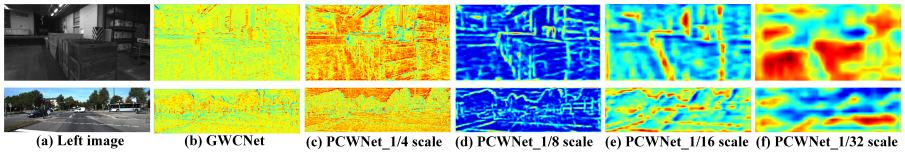


Fig. 3: Visualization of extracted multi-scale feature maps on two real-world datasets (from top to bottom: ETH3D and KITTI). All methods are trained on synthetic data (SceneFlow) and tested on unseen real scenes. Note that GWCNet only extracts feature maps at 1/4 scale for following single-scale cost volume construction while our method extracts multi-scale feature maps for pyramid cost volume construction.

ever, SSPCV just fuses the pyramid cost volume by constantly employing 3D hourglass modules to regularize the upsampled cost volume. Such operation is time-consuming and GPU memory-unfriendly.

3 Proposed Approach

We propose a PCW cost volume to effectively exploit the multi-scale cues for accurate and robust disparity estimation. The architecture of our network is illustrated in Fig. 2, which consists of three parts: multi-scale feature extraction, multi-scale combination volume based cost aggregation, and warping volume based disparity refinement. Specifically, the extracted multi-scale features are first employed to construct a pyramid cost volume. Then, the pyramid volumes have been used for two purposes. Firstly, we construct combination volumes on the upper levels of the pyramid and develop a cost volume fusion module to integrate them for initial disparity estimation. Secondly, we construct the warping volume at the last level of the pyramid for disparity refinement. Details of each module will be introduced as follows.

3.1 Multi-scale Features Extraction

As shown in Fig. 2, given an image pair, following the Resnet-like network proposed in [11,2], we use three convolution layers with 3×3 kernels, four basic residual blocks, and a $\times 2$ dilated block to get the unary feature map at the first level (1/4 of the original input image size). Then three residual blocks with stride 2 are employed to obtain the feature maps at the other three levels with $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ of the original input image size. With the extracted features, a series of pyramid cost volumes can be constructed at different levels.

3.2 Combination volume based 3D Aggregation

We propose to construct multi-scale combination volumes and develop a cost volume fusion module for initial disparity estimation. Previous work [35] observes that the limited effective receptive field of current deep stereo matching methods will drive the network to learn domain-sensitive local features. Instead, our method can cover multi-scale receptive fields and boost the network to see

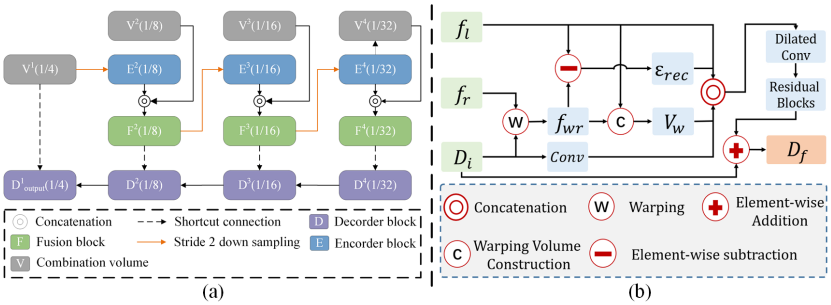
different scale regions of the original image by fusing multi-scale cost volumes. As shown in the figure 3, we visualize the extracted multi-scale feature map on various real datasets. It can be seen from figure 3. (b) that GWCNet [11] only extracts 1/4 scale features of the input image, which only contains local information such as textures, thus the performance is limited. On the contrary, our method extracts features with multi-scales, which contain much more high-level information (sub-figs (c)-(f)), i.e., textures (c), contours (d,e), and areas (f). Typically, non-local information (such as contours and area) is more robust to domain changes and that is why our method achieves better generalization ability. Moreover, sub-figs (a) shows that the used two real datasets have significant domain shifts, e.g., indoors vs outdoors and color vs gray. However, our method can still extract domain-invariant contours (sub-figs (d)-(e)) and areas (sub-figs (f)) information from two real datasets, which further verifies the effectiveness of the proposed method. In addition, high-level information, i.e., contours and area can drive the network to better learn the affiliation between an object and its sub-region, e.g., textureless regions and repeated patterns such as car window is a part of the car, thus, better performance and generalization ability for different resolutions (high and low resolutions) of images can be obtained.

Multi-scale Combination Volume Construction The combination volume is constructed at 4 pyramid levels and for each level i, the combination volume V^i_{comb} is a 4D volume with the size of $H^i \times W^i \times D^i \times C$ which includes concatenation volume V^i_{concat} and group-wise correlation volume V^i_{corr} [11], where (H^i, W^i) is the spatial size. Assuming the extracted feature at level i is f^i, the combination volume V^i_{comb} can be computed as:

$$\begin{aligned} V^i_{comb} &= V^i_{concat} \,||\, V^i_{corr}, \\ V^i_{concat}(d,x,y) &= \delta_1(f^i_L(x,\ y)) \,||\, \delta_1(f^i_R(x-d,\ y)) \\ V^i_{corr}(d,x,y,g) &= \frac{1}{N^i_c/N_g} \left\langle \delta_2(f^{ig}_L(x,\ y)), \delta_2(\ f^{ig}_R(x-d,\ y)) \right\rangle \end{aligned} \tag{1}$$

where $||$ denotes concatenation operation at the feature axis and f^i_L, f^i_R are extracted features at left and right images respectively. f^{ig} are grouped features, which are evenly divided from the extracted feature f^i according to the number of group N_g. d denotes all disparity levels in $(0, D^i_{\max})$, N_c is the channels of f^i and $\langle , \rangle$ represents the inner product. Different with gwcnet[11], during the construction of combination volume, we add one more convolution layer without activation function and batch normalization (named as normalization layer δ) to make the two terms of feature (f^i and f^{ig}) share the same data distribution. Experimental results show that this simple while efficient operation can optimize the two terms of cost volume complementary to each other and thus promote the final performance. Then the multi-scale combination volume will be fused together to predict the initial disparity map.

Multi-scale Cost Volumes Fusion The multi-scale cost volume fusion module is shown in Fig. 4 (a), where the combination volumes, encoder blocks, fusion blocks, and decoder blocks are denoted as V^i, E^i, F^i, D^i, respectively,



(a) (b)

Fig. 4: (a) Structure of multi-scale cost volume fusion module. (b) Structure of warping volume-based refinement network. D_f denotes the final disparity estimation.

($i \in \{1, 2, 3, 4\}$ denotes different levels). The final output fused cost volume is D^1_{output}. Then we use three stacked 3D hourglass networks to further process the fused cost volume and generate the initial disparity map d_i.

Fusion blocks. The proposed fusion blocks have two main inputs. i) The encoder blocks, which characterize the information of higher resolution cost volume. ii) The combination volume, which directly measures the similarity between the left feature and the corresponding right feature according to a coarser disparity index. By employing the fusion blocks, we can integrate multi-scale cost volume and boost the network to evaluate the similarity of the left feature and candidate matching right feature at different scale disparity plane intervals, e.g., each disparity index represents 4 pixels interval at scale one while 32 pixels interval at scale four. Specifically, the fusion process can be formulated as:

$$F^i = \text{Conv}(V^i || E^i), \tag{2}$$

where $||$ denotes the concatenation operation at the feature axis and Conv() refers to the 3D convolution layer.

Encoder block. Encoder block is downsampled from the previous scale fusion block by a 3D convolution with stride 2, except for E^1, which is directly downsampled from the first scale combination volume.

Decoder blocks. Decoder blocks comprise two main components. i) The main data flow, which continually upsamples different scale decoder blocks from D^4 to D^1. ii) The shortcut connection, which combines scale-matching fusion (encoder blocks at scale one) and decoder blocks by element-wise addition. By employing the shortcut connection, we can control the contribution of the last scale decoder block and thus balance the information flowing between upsampled decoder blocks and corresponding fusion blocks. Specifically, the decoder process can be formulated as:

$$D^i = \begin{cases} \text{Conv}^T(D^{i+1}) + S(F^i) & if \quad i = 2, 3, 4 \\ \text{Conv}^T(D^{i+1}) + S(V^i) & if \quad i = 1 \end{cases} \tag{3}$$

Where $\text{Conv}^T()$ denotes the 3D transposed convolution. $S()$ refers to the shortcut connection, which is implemented by $1 \times 1 \times 1$ 3D convolution.

3.3 Warping Volume-based Disparity Refinement

As an essential step in typical stereo matching algorithms, disparity refinement has been widely used in deep learning-based methods. Different from previous

stereo matching methods [20,14] which learn the residual disparity value directly by the network, a multi-modal input is introduced to help our network more purposefully learn the residue. Specifically, our multi-modal input consists of the 3D warping volume, initial disparity map, left features, and reconstructed error, where the 3D warping volume is at the core. By employing the 3D warping volume, we can define a fine-grained residue searching range and alleviate the difficulty of the network to find the correct residue in an unconstrained residue searching space. Below we will describe each input in more detail.

3D Warping Volume. We employ the left feature and warped right feature to construct the warping volume at the last level of the pyramid. Other than the right features we used in the combination volume, we select to warp the right features according to the estimated initial disparity D_i. In this case, we can narrow down the residue searching range from initial disparity searching range $(0, D^i_{\max})$ to a fine-grained one $(D_i - d_{res}, D_i + d_{res})$. Intuitively, the residual disparity is small. Hence, a small residue searching range d_{res} is enough to correct the wrong correspondences. Specifically, the warping volume is computed as:

$$\begin{aligned} V_w(d_{res}, x, y) &= \tfrac{1}{N_c} \left\langle f_l(x, y), f_{wr}(x - d_{res}, y) \right\rangle, \\ f_{wr} &= \text{warping}(f_r, D_i), \end{aligned} \tag{4}$$

where f_l and f_r are upsampled from the first level feature to the original image size, d denotes all residue levels in $(D_i - d_{res}, D_i + d_{res})$ and $\langle\ ,\ \rangle$ represents the inner product.

Besides, the warping operation is implemented differentially by bilinear sampling [12]. Note that the proposed warping volume measures the similarities between left features and warped right features at each residue level which guides the network to output the optimal residual disparity with the most similarity. Moreover, we construct 3D warping volume ($H \times W \times D \times 1$) by inner product to avoid 3D convolutions which can significantly decrease the computational complexity and memory consumption.

Reconstructed Error. We introduce the reconstructed error to identify inaccurate regions of initial disparity estimation, which can be computed as:

$$\mathcal{E}_{rec} = f_l(x, y) - f_{wr}(x, y). \tag{5}$$

The definition of our reconstruction error is inspired by the typical left-right consistency check, while we select to construct it at the feature level rather than the image level. By employing the reconstructed error to indicate the incorrect regions of initial disparity, our refinement network can better identify the pixels that should be further optimized.

Left Image Feature and Initial Disparity. Left image features and initial disparity map are the other two inputs of our refinement network. The initial disparity map provides the network a base estimation for further optimization and the left image feature contains the context informing for residual learning. To balance the weight of multi-model input, the one-channel initial disparity map is regularized by a convolution layer to generate a 32-channel feature map.

Warping Volume-based Refinement Network. In summary, the warping volume, initial disparity map, left image features, and reconstructed error are the input of our refinement network. The detailed architecture of the refinement network is given in Fig. 4 (b). A dilated convolution [33] based network is employed to enlarge the receptive field which can enhance the network to give a better estimation in low-texture and occluded regions. Specifically, it has 5 convolution layers and three basic residual blocks with different dilation constants. The dilation constants are 1, 1, 2, 4, 8, 16, 1, and 1 from top to bottom.

3.4 Loss Function

Inspired by previous work [2,11], we employ smooth L_1 loss function [9] to train our network in an end-to-end way. For each cost volume fusion module and stacked hourglass network in cost aggregation, the same output module and soft argmin operation are used to get intermediate disparity map [11]. In total, we get six disparity maps d_0, d_1, d_2, d_3, d_4, d_5 and the loss function is described as:

$$\mathcal{L} = \sum_{j=0}^{j=5} w_j \cdot \mathcal{L}_{\text{smooth-L1}}(d_j - \widehat{d}), \tag{6}$$

$\mathcal{L}_{\text{smooth-L1}}$ represents the $smooth - L1$ loss and $\widehat{d}$ represents the ground-truth disparity and w_j is the weight of the j^{th} estimation of disparity map.

4 Experimental Results

We evaluate our PCW-Net on various of benchmarks, including: Scene Flow [16], ETH3D [24], KITTI 2012&2015 [8,17], and Argoverse [29].

4.1 Datasets

(1). **SceneFlow:** is a large synthetic dataset with 35,454 training and 4,370 test images of size 960×540. It includes "Flyingthings3D", "Driving", and "Monkaa" with dense and accurate ground-truth for training. Here, we use the Finalpass of the Scene Flow datasets for pre-training. (2). **ETH3D:** is a grayscale image dataset with both indoor and outdoor scenes. The 27 training image pairs of ETH3D are employed to verify the generalization of different approaches. (3). **Middlebury:** is an indoor dataset with 15 training image pairs and 15 testing image pairs with full, half, and quarter resolutions. We select half-resolution training image pairs to evaluate the generalization of different approaches. (4). **KITTI 2015 & KITTI 2012:** are collected from the real world with a driving car. KITTI 2015 contains 200 training and 200 testing image pairs while KITTI 2012 provides 194 training and 195 testing image pairs, respectively. For each dataset, we select 180 image pairs from the training split for training and the rest image pairs are taken as the validation set. (5). **Argoverse:** is a high-resolution real-world dataset collected from a driving car. It provides 5530 training images and 1094 testing images of size 2056×2464. We use it to evaluate the performance of our method on high-resolution datasets, e.g., 10 times higher than KITTI.

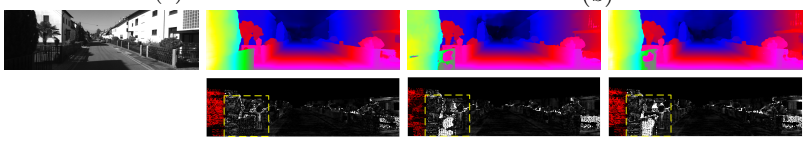
Table 1: (a) Evaluation Results on the KITTI 2012&2015 benchmark and all pixels in occluded and non-occluded areas are evaluated. (b) Evaluation Results on the Argoverse stereo benchmark. For a clear comparison, we highlight the best result in **bold** and the second-best result in blue for each column. All metrics are the lower the better.

Methods	KITTI 2015			KITTI 2012			
	D1-bg	D1-fg	D1-all	2 px	3 px	4 px	5 px
CSPN [5]	1.51	**2.88**	1.74	2.27	1.53	1.19	0.98
GANet-deep [34]	1.48	3.46	1.81	2.50	1.60	1.23	1.02
ACFNet [37]	1.51	3.80	1.89	2.35	1.54	1.21	1.01
GWCNet [11]	1.74	3.93	2.11	2.71	1.70	1.27	1.03
SSPCVNet [30]	1.75	3.89	2.11	3.09	1.90	1.41	1.14
PSMNet [2]	1.86	4.62	2.32	3.01	1.89	1.42	1.15
LEAStereo [6]	1.40	2.91	**1.65**	2.39	1.45	1.08	0.88
Our PCW-Net	**1.37**	3.16	1.67	**2.18**	**1.37**	**1.01**	**0.81**

(a)

Mehtod	10 px(%)			5 px(%)			3 px(%)		
	all	fg	bg	all	fg	bg	all	fg	bg
4Fun	1.79	2.20	1.62	3.39	3.07	3.52	6.92	4.41	7.95
SMD-Stereo	1.90	2.26	1.75	3.62	3.15	3.81	7.32	4.48	8.49
Cicero-stereo	1.99	2.29	1.87	3.68	3.13	3.90	**6.37**	**4.13**	**7.29**
NLCANet [22]	2.00	2.38	1.85	3.69	3.31	3.84	7.44	4.60	8.59
GANet-refine [34]	2.17	2.23	2.15	3.73	3.09	3.99	7.35	4.43	8.55
CFNet [25]	2.38	3.79	1.80	4.05	4.72	3.78	7.60	6.18	8.18
PSMNet [2]	3.05	3.81	2.75	4.85	4.98	4.79	8.51	6.20	9.45
Our PCW-Net	**1.64**	**1.98**	**1.49**	**3.17**	**2.89**	**3.28**	7.05	4.29	8.18

(b)



(a) left image (b) PCW-Net (c) GANet-deep (d) GWCNet

Fig. 5: Visualization results on KITTI 2012 testset. The left panel shows the left input image of the stereo image pair, and for each example, the first row shows the predicted colorized disparity map and the second row shows the error map.

4.2 Implementation Details

The proposed framework is implemented using Pytorch and trained in an end-to-end manner with Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$). Inspired by HSM-Net [31], we employ asymmetric chromatic augmentation and asymmetric occlusion for data augmentation. Moreover, we proposed a *switch training strategy* to train our model for better network parameters. Specifically, it can be realized in three steps. First, the *Relu* activation function is employed to train our network from scratch on the SceneFlow dataset for the first 20 epochs. We set the initial learning rate as 0.001 and down-scale it by 2 times after epoch 12, 16, and 18, respectively. Then, *Mish* [18] is used to prolong the pre-training process on the SceneFlow dataset for another 15 epochs. Finally, the pre-trained models are fine-tuned on KITTI 2015 and KITTI 2012 for another 400 epochs. The learning rate of this process begins at 0.001 and decreases to 0.0001 after epoch 200. Similar to other approaches, we only use the training images of KITTI 2012 for the fine-tuning process on KITTI 2012 benchmark while we merge the training images of both datasets for the training of KITTI 2015 benchmark. For all the experiments, the batch size is set to 4 for training on 2 NVIDIA V100 GPUs and the weights of six outputs are 0.5, 0.5, 0.5, 0.7, 1.0, and 1.3. The inherent principle of the proposed switch training strategy will be discussed in the supplementary materials.

4.3 Fine-tuning Performance Evaluation

In this section, we conduct experiments on various benchmarks to verify our claim in Sec. 1 that the proposed method can achieve consistent SOTA fine-tuning performance on diverse real-world datasets with different proprieties.

Specifically, Argoverse [29] and KITTI 2012&2015 [8,17] are used for evaluation. Below we describe each dataset's result in more detail.

Results on KITTI 2012&2015. We train our model on the SceneFlow dataset first and then fine-tune it on the KITTI dataset. Here, we compare our fine-tuned model with other existing state-of-the-art methods. All results are obtained from the official KITTI evaluation website. Tab. 1(a) illustrates the comparison of the proposed method with others on the KITTI-2012. It can be shown that the proposed method achieves the best performances across all the pixels error thresholds. For the ranking criterion e.g., three-pixel-error rate, our model achieves a 1.37% overall error rate which outperforms our base model GWCNet [11] by 19.4%. Furthermore, compared to the current best-published method LEAStereo [6], our method can also achieve a 5.5% error reduction on the overall three-pixel-error rate.

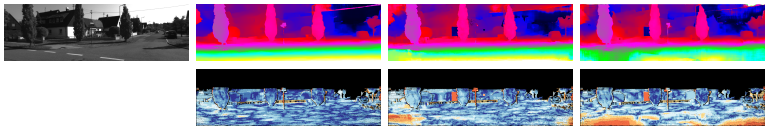
The comparison with other state-of-the-art approaches on the KITTI-2015 benchmark is given in Tab. 1(a). From this table, generally, we can easily find that the proposed method achieves 1 first-place and 1 second-places among all the three categories. Specifically, our method achieves a 1.67% overall three-pixel-error rate, which surpasses the base model GWCNet by 20.85%. Compared to LEAStereo [6], we can obtain very similar results, especially for the ranking criterion "D1-all" category (1.65 vs 1.67).

Qualitative comparison results on the KITTI 2012 benchmark are shown in Fig. 5, and we can see that our method shows significant improvement in ill-posed regions and fence regions (see dash boxes in the picture). The visualization results further support our claim that employing multi-scale cost volumes can guide the network to learn the affiliation between an object and its sub-region, thus promoting the estimation of the textureless region and repeated pattern. More qualitative results are given in the supplementary materials.

Results on Argoverse. Argoverse is a high-resolution real-world dataset collected from a driving car. In comparison to KITTI, it has 10 times the resolution and 16 times as many training frames, making it a more robust and challenging dataset. Similar to the KITTI, we train our model on the SceneFlow dataset first and then fine-tuning it on the Argoverse dataset. Here, we compare our fine-tuned model with other existing state-of-the-art methods in Tab. 1(b). All results are obtained from the official Argoverse evaluation website. To be clear, *the 10-pixel error* is taken as the *official evaluation metric* in this benchmark due to its high image resolution. From this table, we can easily find that existing state-of-the-art stereo matching methods [34,2,22,25] cannot achieve consistent finetuning performance on the Argoverse dataset. This is likely caused by the different proprieties between KITTI and Argoverse, e.g., high-resolution vs low-resolution and large-scale dataset vs small-scale dataset. Instead, as shown in Tab. 1(b), we can easily find that the proposed method achieves 6 first places among all the nine categories, which further verifies our claim that the proposed method can achieve consistent performance on diverse datasets. We attribute this result to the proposed multi-scale cost volume fusion module, which can cover multi-scale receptive fields and boost the network to see different scale regions

Table 2: Cross-domain generalization evaluation on four real datasets. For a fair comparison, all methods are only trained on the SceneFlow training set and tested on four real datasets. We highlight the best result in **bold** and the second-best result in blue for each column. All the metrics are the lower the better. Half resolution training sets of Middlebury is employed for evaluation.

Method	KITTI2012 D1_all(%)	KITTI2015 D1_all(%)	Middlebury(half) bad 2.0(%)	ETH3D bad 1.0(%)	time (s)
HD^3	23.6	26.5	37.9	54.2	**0.14**
PSMNet	15.1	16.3	25.1	23.8	0.41
GWCNet	12.0	12.2	24.1	11.0	0.32
GANet	10.1	11.7	20.3	14.1	1.8
DSMNet	6.2	6.5	**13.8**	6.2	1.5
CFNet	**4.7**	**5.8**	19.5	**5.8**	0.22
Our PC-Net	4.5	5.8	19.00	5.4	0.33
Our PCW-Net	**4.2**	**5.6**	**15.77**	**5.2**	0.44



(a) left image (b) PCW-Net (c) GANet-deep (d) GWCNet

Fig. 6: Cross-domain generalization comparison on KITTI2012 trainset. All methods are trained on the synthetic dataset and tested on KITTI2012 trainsets. The left panel shows the left input image of stereo image pairs, and for each example, the first row shows the predicted colorized disparity map and the second row shows the error map.

of the original image. Such an operation is well suited for both low-resolution and high-resolution images. Specifically, our method outperforms state-of-the-art approaches on overall ten-pixel-error and five-pixel-error rates with 1.64% and 3.17%. Cicero-stereo is the best method on the three-pixel-error rate and our method can achieve comparable results with it, especially for the ranking criterion "fg" category (4.29% vs 4.13%). Note that the evaluation images in Argoverse Dataset are with high resolution (2056×2464). Thus, ten-pixel-error and five-pixel-error are the main evaluation metrics. All in all, our method ranks 1^{st} on the Argoverse leaderboard and sets a new SOTA performance.

4.4 Cross-Domain Generalization Evaluation

In this section, we conduct experiments to verify our claim in Sec. 1 that the proposed PCWNet can achieve strong cross-domain generalization. Specifically, we design an experiment by training the model on the synthetic data only and testing it on four real datasets such as KITTI 2012, KITTI 2015, ETH3D, and Middlebury. To make a fair comparison, all the methods are trained only on the Scene Flow dataset (without any other synthetic or real data will be used, e.g., Carla [35]). The comparison with other approaches is given in Tab. 2. From this table, we can find that our method outperforms the baseline model *gwcnet* on all four datasets with a large margin. Compared to the second-best method CFNet[25], our proposed PCNet (refers to the network without the warping volume based disparity refinement) has achieved comparable performance and the proposed PCWNet can further surpass it on all four datasets. Specifically, the error rate on KITTI 2012, KITTI 2015, ETH3D, and Middlebury has been de-

creased by 10.64%, 3.45%, 10.34%, and 19.13%, respectively compared to CFNet. Most importantly, both CFNet and DSMNet are specially designed for cross-domain generalization and will make a significant compromise on finetuning performance, e.g., the D1_all error rate of CFNet [25] on the KITTI 2015 benchmark is 1.88%, which is 10.11% higher than ours. In summary, the comparison between these domain-generalization methods further shows that our PCW-Net can make a good balance between performance and generalization.

In addition, we compare the generalization results of our method with some state-of-the-art methods in Fig. 6. From this figure, we can clearly see that most existing dataset-specific methods [11,34] generalize poorly to unseen real scenes while our method can correct most errors and generate a reasonable result. More qualitative results on other datasets will be given in the supplementary materials.

4.5 Ablation Studies

To verify the effectiveness of different modules, we set a series of experiments in this section. For efficient evaluation, only the KITTI 2015 dataset (without pretraining from Scene Flow) has been used for training and evaluation. Generally, four types of experiments have been executed here.

Multi-scale cost volume fusion. The proposed multi-scale cost volume fusion module consists of the combination volumes, encoder blocks, fusion blocks, and decoder blocks. Here, we verify the impact of removing the fusion blocks, which means that the multi-scale combination volume information is ignored. As shown in the $Multi-scale\ Cost\ Volume\ Fusion$ section of Tab. 3, the D1_all error rate increase from 1.97% (D+E+F(ours)) to 2.09% (D+E) after removing the fusion blocks, which further verifies the necessity of including multi-scale information.

Cost volume construction. The proposed combination volume consists of concatenation volume and group-wise correlation volume. Here, We test the impact of using different cost volumes. As shown in Tab. 3, the proposed combination volume achieves the best result. Moreover, the performance of combination volume without the normalization layer δ is even worse than the usage of single cost volume. Thus, it's essential to add this layer to make the two cost volumes share the same data distribution.

Multi-modal input evaluation. In the disparity refinement module, we employ multi-modal input to help our network learn the residue more purposefully. Here, we test the impact of each input individually. As shown in the *Multi-modal input* section of Tab. 3, each input is indispensable and the 3D warping volume is at the core. Specifically, the improvements of each part are: 0.29% for V_w, 0.08% for $\mathcal{E}_{rec}$, 0.02% for f_i and D_i, respectively. The result verifies all the multi-modal inputs work positively to improve the performance and compared with other inputs, the 3D warping volume V_w achieves the largest gain.

Model Generalization. Moreover, to further verify the generalization of the proposed method, we conduct two more ablation studies. In this setting, all the frameworks are trained on the SceneFlow dataset and evaluated on the SceneFlow testing set and KITTI 2015 training set without finetuning. The comparison results are given in Tab. 4 (a). From the table, we can find that the proposed

Table 3: Ablation Study of the proposed method on the KITTI2015 dataset. V_w, $\mathcal{E}_{rec}$, f_l, D_i denote the 3D warping volume, reconstructed Error, left features, and initial disparity map, respectively. D, E, and F represent decoder blocks, encoder blocks, and fusion blocks, respectively. D1_all is used for evaluation (the lower the better). 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.

Experiment	Method	KITTI D1_all
Multi-scale Cost Volume Fusion	D+E	2.09
	D+E+F (ours)	**1.97**
Cost Volume	concatenation volume	2.04
	group-wise correlation volume	2.13
	combination volume without C_r	2.14
	combination volume (ours)	**1.97**
Multi-modal input	Multi-modal input without V_w	2.26
	Multi-modal input without $\mathcal{E}_{rec}$	2.05
	Multi-modal input without f_l and D_i	1.99
	Multi-modal input (ours)	**1.97**

Table 4: (a) Ablation study of model generalization. (b) Sub-module generalization analysis on three real datasets. All methods are only trained on the synthetic dataset and tested on three real datasets. MSCVF and WVBDF denote the multi-scale cost volume fusion module and warping volume based disparity refinement module, respectively. Raw disparity refers to the disparity estimation result before cost volume fusion.

Method	SceneFlow EPE (px)	KITTI 2015 (w/o finetuning) D1_all (%)
no WVBDF + MSCVF	0.8578	6.18
no WVBDF	0.8387	5.81
PCWNet	**0.7868**	**5.55**

(a)

Different operations	KITTI 2012 D1_all (%)	KITTI 2015 D1_all (%)	ETH3D bad 1.0 (%)
raw disparity	8.60	8.57	16.44
MSCVF	5.62 (**-2.98**)	6.5 (**-2.07**)	9.96 (**-6.48**)
Stacked hourglass	4.49 (-1.13)	5.84 (-0.66)	6.57 (-3.39)
WVBDF	4.23 (-0.26)	5.55 (-0.29)	5.2 (-1.37)

(b)

multi-scale cost volume fusion(MSCVF) and warping volume based disparity refinement (WVBDF) can both promote the generalization ability on KITTI as well as finetuning performance on SceneFlow. The error on the KITTI dataset has been decreased from 6.18% to 5.55%. Moreover, we further analyze the effect of each module on generalization in Tab. 4 (b). As shown, each module works positively for better generalization, and the multi-scale cost volume fusion module (MSCVF) is at the core, which contributes 68.19%, 68.54%, 57.65% error reduction on KITTI2012, KITTI2015, and ETH3D, respectively.

5 Conclusion

In this paper, we have proposed a pyramid combination and warping cost volume based network, *i.e.*, PCW-Net, for accurate and robust stereo matching. Our pyramid cost volume can be divided into two parts. First, we construct combination volumes on the upper levels of the pyramid and develop a cost volume fusion module to integrate them for initial disparity estimation. Second, we construct the warping volume on the last level of the pyramid and employ it to refine the initial disparity. Experimental results show the superiority of PCW-Net across a diverse range of datasets. Specifically, PCW-Net achieves state-of-the-art performance and strong cross-domain generalization at the same time.

Acknowledgements This research was supported by National Key Research and Development Program of China (2018AAA0102803) and NSFC (61871325).

References

1. Biswas, J., Veloso, M.: Depth camera based localization and navigation for indoor mobile robots. In: IEEE International Coference on Robtics and Automation (ICRA). pp. 1697–1702 (2011) 2
2. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5410–5418 (2018) 2, 4, 5, 9, 10, 11
3. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2722–2730 (2015) 1
4. Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2524–2534 (2019) 4
5. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **42**(10), 2361–2379 (2019) 10
6. Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Drummond, T., Li, H., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. Advances in Neural Information Processing Systems (NIPS) pp. 22158–22169 (2020) 10, 11
7. Duggal, S., Wang, S., Ma, W.C., Hu, R., Urtasun, R.: Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4384–4393 (2019) 4
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3361. IEEE (2012) 2, 9, 11
9. Girshick, R.: Fast r-cnn. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1440–1448 (2015) 9
10. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2495–2504 (2020) 4
11. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3273–3282 (2019) 2, 4, 5, 6, 9, 10, 11, 13
12. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 2017–2025 (2015) 8
13. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 66–75 (2017) 2, 4
14. Liang, Z., Guo, Y., Feng, Y., Chen, W., Qiao, L., Zhou, L., Zhang, J., Liu, H.: Stereo matching using multi-level cost volume and multi-scale feature constancy. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019) 4, 8
15. Mao, Y., Liu, Z., Li, W., Dai, Y., Wang, Q., Kim, Y.T., Lee, H.S.: Uasnet: Uncertainty adaptive sampling network for deep stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6311–6319 (2021) 4

16. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4040–4048 (2016) 2, 3, 4, 9
17. Menze, M., Heipke, C., Geiger, A.: Object scene flow. ISPRS Journal of Photogrammetry and Remote Sensing **140**, 60–76 (2018) 2, 9, 11
18. Misra, D.: Mish: A self regularized non-monotonic neural activation function. arXiv preprint arXiv:1908.08681 (2019) 10
19. Nie, G.Y., Cheng, M.M., Liu, Y., Liang, Z., Fan, D.P., Liu, Y., Wang, Y.: Multi-level context ultra-aggregation for stereo matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3283–3291 (2019) 2, 4
20. Pang, J., Sun, W., Ren, J.S., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: IEEE/CVF International Conference on Computer Vision workshop (ICCV workshop). pp. 887–895 (2017) 4, 8
21. Rao, Z., Dai, Y., Shen, Z., He, R.: Rethinking training strategy in stereo matching. IEEE Transactions on Neural Networks and Learning Systems (2022) 2
22. Rao, Z., He, M., Dai, Y., Zhu, Z., Li, B., He, R.: Nlca-net: a non-local context attention network for stereo matching. APSIPA Transactions on Signal and Information Processing **9** (2020) 10, 11
23. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision (IJCV) **47**(1), 7–42 (2002) 2
24. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3260–3269 (2017) 2, 9
25. Shen, Z., Dai, Y., Rao, Z.: Cfnet: Cascade and fused cost volume for robust stereo matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13906–13915 (2021) 2, 3, 10, 11, 12, 13
26. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8934–8943 (2018) 4
27. Tonioni, A., Tosi, F., Poggi, M., Mattoccia, S., Stefano, L.D.: Real-time self-adaptive deep stereo. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 195–204 (2019) 4
28. Wang, J., Jampani, V., Sun, D., Loop, C., Birchfield, S., Kautz, J.: Improving deep stereo network generalization with geometric priors. arXiv preprint arXiv:2008.11098 (2020) 2
29. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2.0: Next generation datasets for self-driving perception and forecasting. In: Advances in Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021) 2, 3, 9, 11
30. Wu, Z., Wu, X., Zhang, X., Wang, S., Ju, L.: Semantic stereo matching with pyramid cost volumes. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7484–7493 (2019) 4, 10
31. Yang, G., Manela, J., Happold, M., Ramanan, D.: Hierarchical deep stereo matching on high-resolution images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5515–5524 (2019) 10

32. Yang, J., Mao, W., Alvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4877–4886 (2020) 4
33. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations (ICLR) (2016) 9
34. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 185–194 (2019) 2, 10, 11, 13
35. Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B., Torr, P.: Domain-invariant stereo matching networks. In: the Europe Conference on Computer Vision (ECCV). pp. 420–439 (2020) 2, 5, 12
36. Zhang, S., Wang, Z., Wang, Q., Zhang, J., Wei, G., Chu, X.: Ednet: Efficient disparity estimation with cost volume combination and attention-based spatial residual. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5433–5442 (2021) 4
37. Zhang, Y., Chen, Y., Bai, X., Zhou, J., Yu, K., Li, Z., Yang, K.: Adaptive unimodal cost volume filtering for deep stereo matching. pp. 12926–12934 (2020) 10
38. Zhong, Y., Loop, C., Byeon, W., Birchfield, S., Dai, Y., Zhang, K., Kamenev, A., Breuel, T., Li, H., Kautz, J.: Displacement-invariant cost computation for stereo matching. International Journal of Computer Vision **130**(5), 1196–1209 (2022) 4