Learning Intra-view and Cross-view Geometric Knowledge for Stereo Matching

Rui Gong*[1], Weide Liu[2], Zaiwang Gu[2], Xulei Yang[2], and Jun Cheng†[2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University
[2]Institute for Infocomm Research, A*STAR

Abstract

Geometric knowledge has been shown to be beneficial for the stereo matching task. However, prior attempts to integrate geometric insights into stereo matching algorithms have largely focused on geometric knowledge from single images while crucial cross-view factors such as occlusion and matching uniqueness have been overlooked. To address this gap, we propose a novel Intra-view and Cross-view Geometric knowledge learning Network (ICGNet), specifically crafted to assimilate both intra-view and cross-view geometric knowledge. ICGNet harnesses the power of interest points to serve as a channel for intra-view geometric understanding. Simultaneously, it employs the correspondences among these points to capture cross-view geometric relationships. This dual incorporation empowers the proposed ICGNet to leverage both intra-view and cross-view geometric knowledge in its learning process, substantially improving its ability to estimate disparities. Our extensive experiments demonstrate the superiority of the ICGNet over contemporary leading models. The code will be available at https://github.com/DFSDDDD1199/ICGNet.

1. Introduction

Stereo matching is the task of estimating a disparity map from a pair of rectified images. It stands as one of the cornerstone challenges in a range of computer vision tasks, including augmented reality, autonomous driving, and robotics [18]. With the rapid development of deep learning techniques and network structures in computer vision scenarios, deep stereo matching networks have achieved great success in stereo matching.

Geometric knowledge plays a crucial role in the overall performance of stereo matching networks [21, 45, 55, 69] due to the existence of repeated patterns, textureless areas, and reflective surfaces, which make simple texture-based matching methods ineffective. Consequently, several methods were proposed to introduce geometric knowledge into stereo matching networks through kinds of geometric structures. EdgeStereo [45] proposed a learning network that incorporates edges and edge-disparity relations into the stereo matching network. Normal assisted stereo matching [21, 69] leverages surface normal as geometric constraints to reduce the matching ambiguity for textureless regions. However, the geometric knowledge employed in these studies is mainly intra-view, lacking the cross-view geometric knowledge such as matching uniqueness and occlusion.

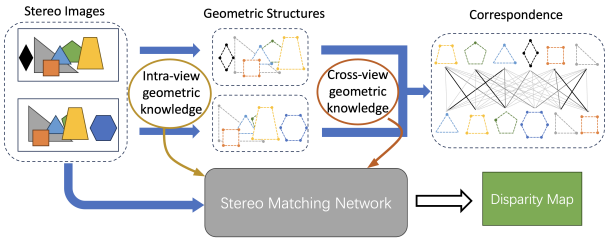

Figure 1. An illustration of our overall framework using synthetic shapes. We leverage intra-view geometric knowledge which is the knowledge of extracting geometric structures, and cross-view geometric knowledge which is the knowledge of the correspondences of these structures, to aid the stereo matching task. Note that the dotted lines are used for illustration and are not used in the method.

To address the absence of cross-view geometric knowledge, we have identified that the local feature matching task [12, 34, 40] can serve as a suitable source of cross-view geometric knowledge for stereo matching. Local feature matching involves the identification of correspondences between two images, typically achieved by initially detecting sparse, geometrically meaningful interest points within each view and then extracting descriptors for these interest points encoded with their visual attributes [12]. Subsequently, cross-view matching is carried out by comparing the interest points between the two images [40]. This matching process solves a partial assignment problem be-

*The work was done during an internship at I²R, A*STAR.
†cheng_jun@i2r.a-star.edu.sg

tween the points which takes matching uniqueness and occlusion into account. Moreover, we also find that the interest points from each image also provide intra-view knowledge of the image. Building upon these insights, we introduce our ICGNet framework to take both the intra-view and cross-view geometric knowledge contained in local feature matching models into the stereo matching network. We leverage a pre-trained local feature matching pipeline, comprising an interest point detector and a point matcher. We propose an intra-view constraint that the features we extract from stereo matching backbone preserve the same interest point information as the pre-trained interest point detector to gain intra-view geometric knowledge. Additionally, we propose a cross-view constraint that these features are compelled to serve as descriptors for the interest points, facilitating the matching process between these points. The matching outcome is then constrained to align with both the ground-truth matching labels provided by ground-truth disparity maps and the matching results obtained from the pre-trained interest point matcher, in order to gain cross-view geometric knowledge. Compared to EdgeStereo [45] and normal assisted stereo matching [21, 69], our method not only additionally introduces cross-view geometric knowledge but also keeps zero overhead when inference.

Our experiments show that our method effectively improves the performance across both the seen and unseen domains. The contributions are summarized as follows:

1. We propose to learn intra-view geometric knowledge for stereo matching networks from an interest point detector by introducing a novel intra-view geometric constraint computed from the interest point detection outcomes.
2. We propose to learn cross-view geometric knowledge for stereo matching networks from an interest point matcher by introducing a novel cross-view geometric constraint. We compute both soft cross-view and hard cross-view geometric losses using pretrained interest point matcher and disparity ground truth.
3. We conduct a series of comprehensive experiments that consistently demonstrate the proposed framework's ability to enhance the overall performance of state-of-the-art stereo matching networks. Our ICGNet achieves state-of-the-art performance on the SceneFlow dataset and ranks 1^{st} on the KITTI 2015 stereo matching benchmark among published peer-reviewed methods at the time of submission, and improves cross-domain generalization of baseline methods on KITTI 2012, KITTI 2015, and Middlebury 2014 datasets.

2. Related Works

2.1. Deep Stereo Matching

Deep-learning-based deep stereo matching techniques have gained widespread adoption and demonstrated promising outcomes [7, 9, 16, 24, 25, 28, 37, 52, 53, 57, 58, 61]. The initial instance of such an approach is DispNetC [35], which presents an end-to-end trainable stereo matching framework. This framework employs the dot product of feature maps from left and right images as correspondence to build a cost volume of three dimensions. However, the drawback of this computational-friendly methodology is its inability to capture sufficient information to achieve satisfactory results. In the pursuit of improved performance, subsequent methods like GC-Net [20] have emerged. Numerous works [3, 65] have followed this trend and incorporated 3D hourglass convolutions to aggregate a 4D cost volume. This volume is created by concatenating features from left and right images. Unfortunately, this approach demands substantial memory usage and computational complexity. A notable enhancement arrives with GwcNet [17], which introduces the concept of group-wise correlation. This innovation enables the construction of a more compact cost volume. Recently, iterative optimization-based methods [23, 27, 63] have achieved splendid performance. Inspired by RAFT [48], RAFT-Stereo [27] iteratively updates the disparity using all-pairs correlation between features extracted. IGEV-Stereo [55] further boosts the performance of stereo matching network by combining cost-volume modules with iterative updating modules. Uncertainty-based methods [8, 51] also gained attention these years. ELFNet [33] leverage uncertainty to fuse multi-scale disparity and stereo network with different structures. SEDNet [6] proposes a soft-histogramming technique to align the distribution of prediction error and prediction uncertainty. Other works focus on improving the domain generalization performance [4, 11, 29, 38, 68] and domain adaptation performance [31, 46, 64, 66] of stereo matching networks.

Recognizing the need for increased efficiency and the integration of richer semantic information, multi-scale cost volume is used in stereo matching networks. HSMNet [59] employs a pyramid of volumes to enable high-resolution stereo matching. AANet [56] takes a lightweight approach, incorporating a feature pyramid and interactions with multi-scale correlation volumes. In a recent development, PCWNet [43] introduces a volume fusion module that directly combines multi-scale 4D volumes. This innovation calculates a multi-level loss, ultimately expediting the model's convergence speed.

2.2. Geometric Structure Guided Stereo Matching

Geometric structures have been used to guide stereo matching networks in previous works. EdgeStereo [45] introduced an edge branch, integrating edge features into the stereo matching branch. Yang et al. [60] also advocated the use of edges for supervising stereo matching through multi-task learning. Additionally, surface normals have been another employed geometric structure to guide stereo match-

ing. Zhang et al. [69] suggested a joint prediction of a surface normal map and a raw disparity map, iteratively refining the disparity map with guidance from the surface normal map. Kusupati et al. [21] incorporated an extra surface normal branch to predict raw surface normals and employed an independently trained consistency module for refinement of both surface normals and raw disparities.

However, these approaches have two limitations. Firstly, they heavily depend on additional branches during inference to improve their performance, which impacts efficiency. In contrast, our proposed framework leverages geometric structures as constraints during training without incurring any overhead in inference, and shows improved performance. Secondly, the previous methods only made use of intra-view geometric structures, whereas our framework takes into account both intra-view and cross-view geometric knowledge through both geometric structure and their correspondences.

2.3. Local Feature Matching

Local feature matching aims at matching images that depict the same scene or object. This is usually done in a two-stage manner: first, detect interest points each associated with a visual descriptor from both images, and then match the detected interest points.

Traditional interest points detection and description algorithms [1, 19, 34, 39] rely on hand-crafted features. Recent research turns to Convolutional Neural Networks (CNN) for both interest point detection and description [22, 30, 32, 49]. SuperPoint [12] is a representative work of CNN-based interest point detector and descriptor, trained in a self-supervised manner. It constructs a synthetic shape dataset by rendering simplified 2D geometries, such as triangles, quadrilaterals, lines, and ellipses. The junctions of lines are labeled as interest points. These labels are then transferred to substantial real-world image datasets using a MagicPoint model and homographic adaptation. The self-supervised training process equips SuperPoint with the ability to be trained on various domains, enabling it to encode geometric information across diverse domains.

Traditional algorithms employ nearest-neighbor classifiers for matching interest points. Lowe's test [34], inlier classifiers [62, 67], or geometric model fitting [2, 13] techniques are subsequently used to filter out unmatched points and incorrect correspondences. However, these methods are highly domain-specific and struggle to handle particularly challenging conditions. In contrast, deep learning-based matchers are trained to simultaneously match interest points and reject outliers. A notable example of this approach is SuperGlue [40], which leverages Transformers [50] to conduct self-attention among interest points within the same view and cross-attention between interest points in different views. The final matching step involves solving an optimal transport problem. This approach allows SuperGlue to acquire robust insights into scene geometry and camera motion, rendering it resilient to extreme changes and effective in generalizing across various data domains. Subsequent works [5, 26, 44] focused on making it more efficient and digging into its design details while keeping its strong matching capability. Consequently, these matchers are great sources of cross-view geometric knowledge. We set up constraints to let the stereo matching network learn geometric knowledge from a pre-trained interest point detector and a pre-trained interest point matcher.

3. Method

We propose ICGNet to introduce geometric knowledge into stereo matching networks through geometric constraints. Intra-view geometric knowledge refers to the understanding that uncovers the geometric structures contained within a single image. Cross-view geometric knowledge refers to the comprehension that enables the alignment of geometric structures from one image to another. Models capable of uncovering and matching geometric structures harness the potential to convey intra-view and cross-view knowledge. Therefore, as depicted in Fig. 2, our framework introduces intra-view and cross-view geometric knowledge into the stereo matching network by aligning our network with intra-view and cross-view geometric knowledge sources. Although there are multiple choices (e.g., edges, points) to extract intra-view knowledge, we choose interest points as it is easy to further compute cross-view knowledge from these points using well-established methods [40] while it is more challenging to match the edges for cross-view knowledge computation. We denote the left-view stereo image as I_l and right-view stereo image as I_r, and the extracted left feature and right feature as F_l and F_r.

3.1. Intra-view Geometric Knowledge

Intra-view Geometric Knowledge Decoder is to introduce intra-view geometric knowledge into the stereo matching network. We propose to leverage a pre-trained interest point detector $\mathcal{IP}_{\text{det}}$ [12] in our method. It takes the left-view image I_l and the right-view image I_r as input and generates interest maps $\mathbf{M}_l$ and $\mathbf{M}_r$, where each pixel is either labeled as interest point or non-interest point. These maps convey the intra-view geometric characteristics of the two images. We expect that the backbone features F_l and F_r preserve and unveil the intra-view geometric details present in both the left and right view images, akin to what the pre-trained interest point detector accomplishes. Consequently, we aim to extract two interest point maps, $\mathbf{M}'_l$ and $\mathbf{M}'_r$, from F_l and F_r using a trainable intra-view decoder D_{intra}, aligning with $\mathbf{M}_l$ and $\mathbf{M}_r$ respectively. For the design of D_{intra}, we use a stack of two bottleneck convolutional blocks. Note that decoder D_{intra} is discarded during the test stage.

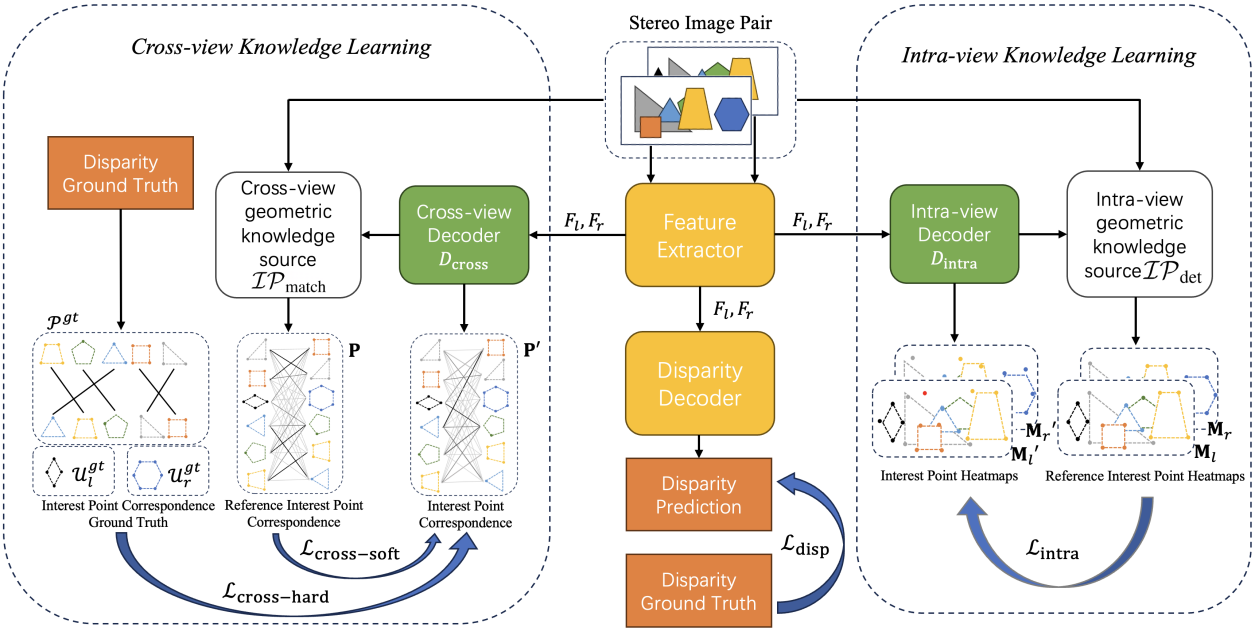

Figure 2. Overall structure of our proposed framework. The model architecture comprises three parts: stereo matching network, cross-view knowledge learning network, and intra-view knowledge learning network. The cross-view knowledge learning network introduces cross-view geometric knowledge by aligning the interest point correspondences $\mathbf{P}'$, $\mathbf{P}$ and $\mathcal{P}^{gt}$ using $\mathcal{L}_{\text{cross-hard}}$ and $\mathcal{L}_{\text{cross-soft}}$. The intra-view knowledge learning network introduces intra-view geometric knowledge through aligning the interest point maps $\mathbf{M}'$ and $\mathbf{M}$ using $\mathcal{L}_{\text{intra}}$. Note that the dotted lines are used just for clear illustration and are not used in our work.

Intra-view geometric loss $\mathcal{L}_{\text{intra}}$ is computed to enforce this alignment:

$$\mathcal{L}_{\text{intra}} = \frac{1}{2}\mathcal{L}(\mathbf{M}_l', \mathbf{M}_l) + \frac{1}{2}\mathcal{L}(\mathbf{M}_r', \mathbf{M}_r), \tag{1}$$

where we compute a focal loss $\mathcal{L}$ due to unbalanced positive and negative samples of interest points.

3.2. Cross-view Geometric Knowledge

Cross-view Geometric Knowledge Decoder enables the stereo matching network to acquire cross-view geometric knowledge by framing it as a learning task. This task is to align the stereo matching network with two complementary knowledge sources: a pre-trained interest point matcher denoted as $\mathcal{IP}_{\text{match}}$ where we use [40] in our method, and the ground truth matching between interest points.

The pre-trained matcher processes two sets of interest points extracted from stereo image pair I_l and I_r, denoted as p_l and p_r, each containing m and n points respectively, along with their associated descriptors. It outputs an assignment matrix $\mathbf{P} \in [0,1]^{(m+1)\times(n+1)}$. This matrix includes an extra row and column to account for the possibility that some points do not match any counterpart in the other view. The assignment matrix $\mathbf{P}$ elucidates the cross-view geometric knowledge via the relations between the matching probabilities of the two sets of points.

Furthermore, the ground truth disparities can serve as another source of knowledge since it provide accurate matching between these points. Ground truth matching pairs between the points $\mathcal{P}^{gt} = \{(i,j)\} \subset p_l \times p_r$ ($\times$ refers to Cartesian product) and unmatched points in left image $\mathcal{U}_l^{gt} \subseteq p_l$ and right image $\mathcal{U}_r^{gt} \subseteq p_r$ are formed by warping interest points from the left image to the right image using the disparity.

The ground truth is more precise compared to the assignment matrix provided by the interest point matcher. In the ground truth, the relationships between points are binary, classified simply as either a match or a mismatch. On the other hand, the assignment matrix presents matching scores as continuous values ranging from 0 to 1, implicitly indicating the degree of similarity between the points. This makes the two approaches complementary: one provides precise matching categorization, which is 'harder', while the other offers a fine-grained assessment of similarity, which is 'softer'.

To align the stereo matching network with these two sources of cross-view geometric knowledge, a trainable cross-view decoder denoted as D_{cross} is proposed. D_{cross} first use as an Multilayer Perceptron (MLP) encoder [14, 50] followed by four alternative self-attention and cross-attention layers [50]. The self attention layers operate on points within the same image, while the cross-attention

layer operate on points within different images. The decoder is responsible for decoding an assignment matrix $\mathbf{P}'$. This decoded matrix is trained to closely match its corresponding reference assignment matrix $\mathbf{P}$ created by a pre-trained interest point matcher $\mathcal{IP}_{\text{match}}$ and the ground truth matching pairs $\mathcal{P}^{gt}$ and unmatched points $\mathcal{U}_l^{gt}$, $\mathcal{U}_r^{gt}$ from left and right images respectively, serving as the supervision pseudo ground truth during training. The points input p_l and p_r to the decoder D_{cross} are identical to the input to $\mathcal{IP}_{\text{match}}$, but with different descriptors bi-linearly sampled from features F_l and F_r extracted by stereo backbone such that the gradient can flow back to the stereo network. Note that decoder D_{cross} is discarded during the test stage.

Soft cross-view geometric loss $\mathcal{L}_{\text{cross-soft}}$ is proposed to introduce the cross-view geometric knowledge from the reference assignment matrix $\mathbf{P}$ into the predicted assignment matrix $\mathbf{P}'$.

$$\begin{aligned}\mathcal{L}_{\text{cross-soft}} = &\frac{1}{m}\sum_{i=1}^{m}\mathcal{KL}\left(\mathbf{P}_{i,\cdot}/||\mathbf{P}_{i,\cdot}||_1, \mathbf{P}'_{i,\cdot}/||\mathbf{P}'_{i,\cdot}||_1\right)\\ &+\frac{1}{n}\sum_{j=1}^{n}\mathcal{KL}\left(\mathbf{P}_{\cdot,j}/||\mathbf{P}_{\cdot,j}||_1, \mathbf{P}'_{\cdot,j}/||\mathbf{P}'_{\cdot,j}||_1\right),\end{aligned} \tag{2}$$

where $\mathcal{KL}$ stands for KL-divergence, and $\mathbf{P}_{i,\cdot}$ stands for the i^{th} row of $\mathbf{P}$, $\mathbf{P}_{\cdot,j}$ stands for the j^{th} column of $\mathbf{P}$. In this work, we use the interest points filtered from interest point maps $\mathbf{M}_l$, $\mathbf{M}_r$ extracted by pre-trained interest point detector [12] by non maximum suppression as p_l and p_r.

Hard cross-view geometric loss is further computed based on the ground truth matching between interest points $\mathcal{P}^{gt}$, unmatched points in the left image $\mathcal{U}_l^{gt}$, and unmatched points in the right image $\mathcal{U}_r^{gt}$. Hard cross-view geometric loss is utilized as a negative log-likelihood loss balanced between matched points and unmatched points to align $\mathbf{P}'$ and $\mathcal{P}^{gt}, \mathcal{U}_l^{gt}, \mathcal{U}_r^{gt}$:

$$\begin{aligned}\mathcal{L}_{\text{cross-hard}} = &-\frac{1}{|\mathcal{P}^{gt}|}\sum_{(i,j)\in\mathcal{P}^{gt}}\log\mathbf{P}'_{i,j}\\ &-\frac{1}{|\mathcal{U}_l^{gt}|}\sum_{i\in\mathcal{U}_l}\log\mathbf{P}'_{i,n+1}\\ &-\frac{1}{|\mathcal{U}_r^{gt}|}\sum_{j\in\mathcal{U}_r}\log\mathbf{P}'_{m+1,j}.\end{aligned} \tag{3}$$

3.3. Loss

We integrate our proposed $\mathcal{L}_{\text{intra}}$, $\mathcal{L}_{\text{cross-soft}}$, and $\mathcal{L}_{\text{cross-hard}}$ with disparity loss L_{disp}, enabling simultaneously estimating disparity and learning intra-view and cross-view geometric structure knowledge from local feature matching models. The disparity loss, L_{disp}, is determined by the loss function utilized by the base models. The total loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\mathcal{L}_{\text{disp}} + \lambda_{\text{intra}}\cdot\mathcal{L}_{\text{intra}} + \lambda_{\text{cross-soft}}\cdot\mathcal{L}_{\text{cross-soft}}\\ &+\lambda_{\text{cross-hard}}\cdot\mathcal{L}_{\text{cross-hard}},\end{aligned} \tag{4}$$

where $\mathcal{L}_{\text{intra}}$, $\mathcal{L}_{\text{cross-soft}}$, $\mathcal{L}_{\text{cross-hard}}$ are weighting terms. The loss weights are empirically set as $\mathcal{L}_{\text{intra}} = 100$, $\mathcal{L}_{\text{cross-soft}} = 0.5$, $\mathcal{L}_{\text{cross-hard}} = 0.5$ such that they are comparable and none of the items would dominate the results.

4. Experiments

SceneFlow dataset [35] constitutes a synthetic dataset featuring diverse random objects. It contains 35,454 training pairs and 4,370 test pairs with dense disparity labels. Finalpass of the SceneFlow dataset that contains motion blur and defocuses is used.

KITTI 2012 [15] & KITTI 2015 [36] datasets consist of stereo image pairs of real-world driving scenarios. KITTI 2012 contains 194 training and 195 testing image pairs. KITTI 2015 comprises 200 training and 200 testing image pairs. The stereo pairs in both KITTI 2012 and KITTI 2015 possess resolutions of 1226×370 and 1242×375, respectively. Notably, both datasets furnish sparse disparity maps as part of their annotations.

Middlebury 2014 [41] is an stereo dataset consisted of indoor stereo images. It comprises 15 training and 15 testing pairs. All the images are available in three different resolutions. We use the Middlebury 2014 dataset with half resolution to test cross-domain generalization performance.

Evaluation metrics. We use end point error (EPE), the percentage of errors larger than 3px (>3px) on the SceneFlow dataset, and use EPE and D1-3px on KITTI datasets as evaluation metrics for disparity prediction. D1-3px refers to the percentage of disparity estimations with both absolute larger than 3px and relative error greater than 5%. We use EPE and the percentage of errors larger than 2px (>2px) on Middlebury 2014 dataset.

4.1. Implementation Details

Our ICGNet is inherently compatible with most stereo matching models since we do not change the structure of the network during inference. Specifically, our framework's compatibility expands to encompass all stereo matching models incorporating a common feature extractor and a disparity decoder. We experiment with our ICGNet on different baseline models GwcNet [17] and IGEV-Stereo [55]. We train the models end-to-end for all experiments. We pre-train the models on the SceneFlow training set. For the evaluation of cross-domain generalization, we directly test the sceneflow pre-trained models' performance on the KITTI 2012 training set, KITTI 2015 training set, and Middlebury 2014 training set. For evaluation of KITTI 2012

Method	PSMNet [3]	GANet [65]	ACVNet [54]	DLNR [71]	GwcNet [17]	IGEV-Stereo [55]	Ours
EPE(px)↓	1.09	0.84	0.48	0.48	0.765	0.479	**0.447**

Table 1. Comparison with state-of-the-art models on the SceneFlow dataset. Our method outperforms most of the state-of-the-art methods.

Method	KITTI 2012				KITTI 2015 (all pixels)			KITTI 2015 (noc pixels)		
	2-noc↓	2-all↓	3-noc↓	3-all↓	D1-bg↓	D1-fg↓	D1-all↓	D1-bg↓	D1-fg↓	D1-all↓
EdgeStereo [45]	2.32	2.88	1.46	1.83	1.84	3.30	2.08	1.69	2.94	1.89
LEAStereo [10]	1.90	2.39	1.13	1.45	1.40	2.91	1.65	1.29	2.65	1.51
ACVNet [54]	1.83	2.35	1.13	1.47	**1.37**	3.07	1.65	1.26	2.84	1.52
CREStereo [23]	1.72	2.18	1.14	1.46	1.45	2.86	1.69	1.33	2.60	1.54
RAFT-Stereo [27]	1.92	2.42	1.30	1.66	1.58	3.05	1.82	1.45	2.94	1.69
AcfNet [70]	1.83	2.35	1.17	1.54	1.51	3.80	1.89	1.36	3.49	1.72
HITNet [47]	2.00	2.65	1.41	1.89	1.74	3.20	1.98	1.54	2.72	1.74
GANet [65]	1.89	2.50	1.19	1.60	1.48	3.46	1.81	1.34	3.11	1.63
GwcNet [17]	2.16	2.71	1.32	1.70	1.74	3.93	2.11	1.61	3.49	1.92
IGEV-Stereo [55]	1.71	2.17	1.12	1.44	1.38	2.67	1.59	1.27	2.62	1.49
Ours	**1.70**	**2.14**	**1.10**	**1.41**	1.38	**2.55**	**1.57**	**1.26**	**2.56**	**1.47**

Table 2. Finetuning results on KITTI 2012 and KITTI 2015 benchmarks. Our ICGNet achieves state-of-the-art performance on both benchmarks and ranks 1^{st} on the KITTI 2015 stereo matching benchmark at the time of submission.

and KITTI 2015 benchmarks, we finetune SceneFlow pre-trained model on the mix of KITTI 2012 and KITTI 2015 training sets. For GwcNet, we use the GwcNet-g setting for KITTI 2015 pre-training and the GwcNet-gc setting for other experiments, consistent with the original work for a fair comparison. We train the model for 64 epochs on SceneFlow and 1000 epochs on KITTI. The initial learning rate is set to 0.001. The learning rate is decayed by a factor of 2 after epochs 20, 32, 40, 48, and 56 on SceneFlow, and decayed by a factor of 2 after epochs 200, 400, 600, 700, and 800 on KITTI. For IGEV-Stereo, we train the model for 80 epochs on SceneFlow and 1000 epochs on KITTI. The initial learning rate is set as 0.0002 and a one-cycle learning rate schedule is used. We maintain all other settings consistent with the original works for both models. All the codes in this paper would be released in Github.

4.2. Comparison with State-of-the-art

To evaluate the effectiveness of our proposed method, we use the latest IGEV-Stereo [55] as the base model and integrate it with ICGNet to compare with state-of-the-art stereo mathing methods. Tab. 1 presents a comparison of our method with state-of-the-art on SceneFlow. Our proposed method achieved state-of-the-art performance on the SceneFlow dataset. We further finetune the SceneFlow pre-trained model on KITTI12 [15] and KITTI15 [36] datasets, and compare the performance of ICGNet with state-of-the-art methods on KITTI 2012 and KITTI 2015 benchmarks. Results are shown in Tab. 2. Our ICGNet outperforms state-of-the-art methods in most scenarios.

Besides quantitative results, qualitative experiments are conducted to show the effectiveness of our proposed method. Fig. 3 visualizes the disparity predictions of our methods. Compared to the state-of-the-art method IGEV-Stereo, our method performs better in reflective regions and unlabelled backgrounds.

4.3. Cross-domain Generalization

We carry out experiments to demonstrate that our approach enhances the domain-generalization capacity of base models, transitioning from the synthetic SceneFlow dataset to real-world datasets. As illustrated in Tab. 3, our proposed ICGNet method boosts the cross-domain generalization performance for IGEV-Stereo which is a strong baseline for cross-domain generalization. Our ICGNet outperforms IGEV-Stereo by 4.8% on EPE in KITTI 2012 and 4.1% in KITTI 2015. Furthermore, from the qualitative results in Fig. 4, we can see that our method outperforms IGEV-Stereo in reflective regions and occluded regions.

4.4. Ablation Studies

We conduct ablation studies to evaluate the effectiveness of our proposed method. To show the generalization of the intra-view and cross-view knowledge learning, two base methods are used: GwcNet [17] and IGEV-Stereo [55].

Effectiveness of Intra-view and Cross-view Knowledge. To verify that our proposed components all contribute to improving the performance, we provide quantitative results on the SceneFlow dataset in Tab. 4. The proposed ICGNet framework introduces an intra-view geometric loss $\mathcal{L}_{\text{intra}}$, a soft cross-view geometric loss $\mathcal{L}_{\text{cross-soft}}$ and a hard cross-view geometric loss $\mathcal{L}_{\text{cross-hard}}$. The intra-view geo-

Method	KITTI 2012		KITTI 2015		Middlebury 2014(H)	
	EPE(px)↓	>3px(%)↓	EPE(px)↓	>3px(%)↓	EPE(px)↓	>2px(%)↓
PSMNet [3]	2.69	15.1	3.17	16.3	7.65	34.2
CFNet [42]	1.04	5.2	1.71	6.0	3.24	15.4
GANet [65]	1.93	10.1	2.31	11.7	5.41	20.3
DSMNet [66]	1.26	6.2	1.46	6.5	2.62	13.8
IGEV-Stereo [55]	1.04	5.18	1.21	6.03	0.91	7.27
Ours	**0.99**	**4.87**	**1.16**	**5.96**	**0.82**	**6.73**

Table 3. Cross-domain generalization from SceneFlow to real-world datasets.

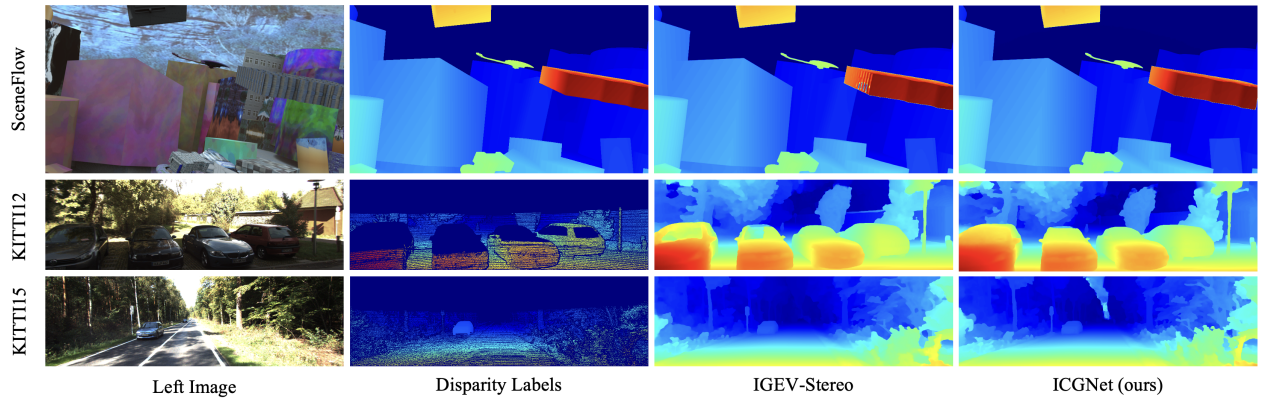

Figure 3. Qualitative results of ICGNet (ours) compared with the state-of-the-art method IGEV-Stereo.

Baseline	Components			SceneFlow	
	$\mathcal{L}_{\text{intra}}$	$\mathcal{L}_{\text{cross-soft}}$	$\mathcal{L}_{\text{cross-hard}}$	EPE(px)↓	>3px(%)↓
GwcNet				0.765	3.30
	✓			0.615	2.61
	✓	✓		0.608	2.56
	✓	✓	✓	**0.597**	**2.55**
IGEV-Stereo				0.479	2.48
	✓			0.452	2.37
	✓	✓		0.449	2.36
	✓	✓	✓	**0.447**	**2.34**

Table 4. Ablation studies of proposed components on SceneFlow dataset on two baseline approaches GwcNet and IGEV-Stereo.

metric loss allows the model to learn intra-image geometric knowledge, and the soft cross-view geometric loss and hard cross-view geometric loss allow the model to learn cross-view geometric knowledge. Notably, our method demonstrates an improvement of over 20% (0.597 vs 0.765) in EPE and over 20% improvement in 3px-error. When applied to the strong base model IGEV-Stereo, our method still shows solid improvements over both EPE and 3px-error. The SceneFlow pre-trained full models of both base models are then finetuned on KITTI 2012 and KITTI 2015 datasets and results are shown in Tab. 5. Our method consistently outperforms the performance of baseline models.

Impact on Occluded and Non-occluded Regions. We further conduct experiments on our method's impact on occluded pixels and non-occluded pixels separately. Tab. 6 shows that our method consistently improves disparity quality in both occluded regions and non-occluded regions. Furthermore, since occluded regions are areas without correspondence in the other figures, intuitively the prediction of disparities on such pixels relies more on intra-view geometric knowledge; on the contrary, predicting disparities at non-occluded pixels relies more on cross-view geometric knowledge [55]. Our experiment results are consistent with this intuition, validating our formulation of the process of extracting interest points as intra-view geometric knowledge and the process of extracting their correspondences as cross-view geometric knowledge.

Complexity of intra-view decoder and cross-view decoder. We conduct experiments to evaluate the impact of the complexity of the intra-view decoder D_{intra} and cross-view decoder D_{cross} on the models' performance. We keep GwcNet as the base model and change the number of layers of the decoders. The number of layers of the intra-view decoder refers to the number of stacked bottleneck convolutional blocks, and the number of layers of the cross-view decoder refers to the number of alternating attention layers. For the cross-view decoder with zero layers, we are directly using cosine similarities between the interest point features plus a softmax layer to form an assignment matrix $\mathbf{P}'$. Results on the SceneFlow dataset with GwcNet [17] as baseline are shown in Tab. 7. Results have shown that an

Baseline	ICGNet	KITTI 2012				KITTI 2015 (all pixels)			KITTI 2015 (noc pixels)		
		2-noc↓	2-all↓	3-noc↓	3-all↓	D1-bg↓	D1-fg↓	D1-all↓	D1-bg↓	D1-fg↓	D1-all↓
GwcNet [17]		2.16	2.71	1.32	1.70	1.74	3.93	2.11	1.61	3.49	1.92
	✓	**1.98**	**2.54**	**1.25**	**1.63**	**1.62**	**3.90**	**2.00**	**1.50**	**3.56**	**1.84**
IGEV-Stereo [55]		1.71	2.17	1.12	1.44	1.38	2.67	1.59	1.27	2.62	1.49
	✓	**1.70**	**2.14**	**1.10**	**1.41**	1.38	**2.55**	**1.57**	**1.26**	**2.56**	**1.47**

Table 5. Ablation studies on KITTI 2012 and KITTI 2015 benchmarks.

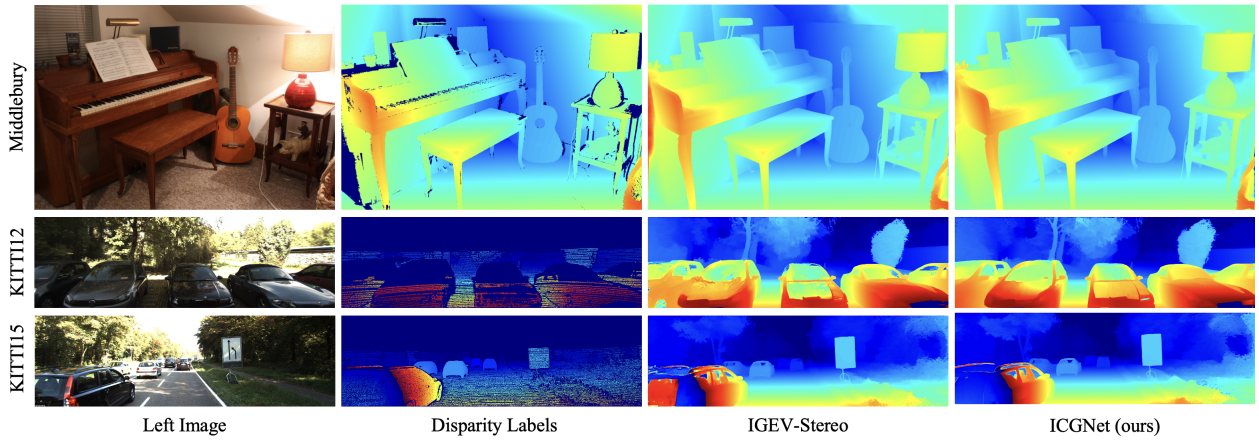

Figure 4. Qualitative results of cross-domain generalization of ICGNet (ours) compared with state-of-the-art baseline IGEV-Stereo [55].

Intra-view	Cross-view	SceneFlow EPE		
$\mathcal{L}_{\text{intra}}$	$\mathcal{L}_{\text{cross-soft}} + \mathcal{L}_{\text{cross-hard}}$	noc(px)↓	occ(px)↓	$\frac{\Delta\text{noc}}{\Delta\text{occ}}$
		0.413	2.58	/
✓		0.334	2.08	6.33
✓	✓	0.318	2.02	3.75

Table 6. Ablation studies on occluded and non-occluded areas on the SceneFlow dataset using GwcNet baseline model. Our method improves disparity prediction accuracy in both occluded areas and non-occluded areas. Compared with cross-view geometric knowledge, intra-view geometric knowledge improves more in occluded regions, consistent with the intuition that occluded areas rely on their surrounding geometric relations to infer their disparities.

Decoder	Layers	SceneFlow	
		EPE(px)↓	>3px↓
intra-view	1	0.631	2.74
	2	**0.615**	**2.61**
	4	0.623	2.68
cross-view	0	0.617	2.61
	4	**0.597**	**2.55**
	6	0.606	2.58

Table 7. Comparison of performance of different decoder complexities.

intermediate number of decoder layers for both intra-view and cross-view decoder yields the best results. This may be due to a too-complex decoder will take up too much of the task of learning the geometric knowledge, making the feature backbone not learn sufficient knowledge, while a too-simple decoder is not sufficient to map the features into the geometric structures, leaving too much for the feature backbone and make it sub-optimal in disparity prediction task.

5. Conclusion

In this work, we introduce ICGNet, a novel approach designed to infuse stereo matching networks with both intra-view and cross-view geometric knowledge. Utilizing interest points and their corresponding matches derived from local feature matching models, we provide rich sources of geometric knowledge to educate the stereo matching network. Our extensive experimental evaluations demonstrate that this methodology enhances the performance of state-of-the-art stereo matching models, boosting both their accuracy and their ability to generalize across diverse domains.

Acknowledgement

This work was supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 3

[2] Luca Cavalli, Viktor Larsson, Martin Ralf Oswald, Torsten Sattler, and Marc Pollefeys. Handcrafted outlier detection revisited. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 770–787. Springer, 2020. 3

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

[4] Tianyu Chang, Xun Yang, Tianzhu Zhang, and Meng Wang. Domain generalized stereo matching via hierarchical visual transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9559–9568, 2023. 2

[5] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6301–6310, 2021. 3

[6] Liyan Chen, Weihan Wang, and Philippos Mordohai. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17235–17244, 2023. 2

[7] Junda Cheng, Gangwei Xu, Peng Guo, and Xin Yang. Coatrsnet: Fully exploiting convolution and attention for stereo matching by region separation. *International Journal of Computer Vision*, pages 1–18, 2023. 2

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 2

[9] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2361–2379, 2019. 2

[10] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33:22158–22169, 2020. 6

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

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 3, 5

[13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[14] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017. 4

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5, 6

[16] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *European Conference on Computer Vision*, pages 263–279. Springer, 2022. 2

[17] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019. 2, 5, 6, 7, 8

[18] Mohd Saad Hamid, NurulFajar Abd Manap, Rostam Affendi Hamzah, and Ahmad Fauzan Kadmin. Stereo matching algorithm based on deep learning: A survey. *Journal of King Saud University-Computer and Information Sciences*, 34(5): 1663–1673, 2022. 1

[19] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, pages 10–5244. Citeseer, 1988. 3

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

[21] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2199, 2020. 1, 2, 3

[22] Jongmin Lee, Byungjin Kim, Seungwook Kim, and Minsu Cho. Learning rotation-equivariant features for visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21887–21897, 2023. 3

[23] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 2, 6

[24] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings*

of the IEEE/CVF international conference on computer vision, pages 6197–6206, 2021. 2

[25] 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. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):300–315, 2019. 2

[26] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. *arXiv preprint arXiv:2306.13643*, 2023. 3

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

[28] Biyang Liu, Huimin Yu, and Yangqi Long. Local similarity pattern and cost self-reassembling for deep stereo matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1647–1655, 2022. 2

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

[30] Jiazhen Liu, Xirong Li, Qijie Wei, Jie Xu, and Dayong Ding. Semi-supervised keypoint detector and descriptor for retinal image matching. In *European Conference on Computer Vision*, pages 593–609. Springer, 2022. 3

[31] Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, and Hongsheng Li. Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12757–12766, 2020. 2

[32] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Gift: Learning transformation-invariant dense visual descriptors via group cnns. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[33] Jieming Lou, Weide Liu, Zhuo Chen, Fayao Liu, and Jun Cheng. Elfnet: Evidential local-global fusion for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17784–17793, 2023. 2

[34] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 1, 3

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

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

[37] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. Multi-level context ultra-aggregation for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3283–3291, 2019. 2

[38] Zhibo Rao, Bangshu Xiong, Mingyi He, Yuchao Dai, Renjie He, Zhelun Shen, and Xing Li. Masked representation learning for domain generalized stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5435–5444, 2023. 2

[39] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 430–443. Springer, 2006. 3

[40] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 3, 4

[41] 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 *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 5

[42] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 7

[43] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *European Conference on Computer Vision*, pages 280–297. Springer, 2022. 2

[44] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12517–12526, 2022. 3

[45] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, 128:910–930, 2020. 1, 2, 6

[46] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. Adastereo: A simple and efficient approach for adaptive stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10328–10337, 2021. 2

[47] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021. 6

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[49] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 3

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[51] Chen Wang, Xiang Wang, Jiawei Zhang, Liang Zhang, Xiao Bai, Xin Ning, Jun Zhou, and Edwin Hancock. Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition*, 124:108498, 2022. 2

[52] Hengli Wang, Rui Fan, Peide Cai, and Ming Liu. Pvstereo: Pyramid voting module for end-to-end self-supervised stereo matching. *IEEE Robotics and Automation Letters*, 6(3): 4353–4360, 2021. 2

[53] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Semantic stereo matching with pyramid cost volumes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7484–7493, 2019. 2

[54] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022. 6

[55] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 1, 2, 5, 6, 7, 8

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

[57] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

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

[59] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 2

[60] Xiaowei Yang, Zhiguo Feng, Yong Zhao, Guiying Zhang, and Lin He. Edge supervision and multi-scale cost volume for stereo matching. *Image and Vision Computing*, 117: 104336, 2022. 2

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

[62] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2666–2674, 2018. 3

[63] Jiaxi Zeng, Chengtang Yao, Lidong Yu, Yuwei Wu, and Yunde Jia. Parameterized cost volume for stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18347–18357, 2023. 2

[64] Chenghao Zhang, Kun Tian, Bin Fan, Gaofeng Meng, Zhaoxiang Zhang, and Chunhong Pan. Continual stereo matching of continuous driving scenes with growing architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18901–18910, 2022. 2

[65] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 2, 6, 7

[66] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 420–439. Springer, 2020. 2, 7

[67] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5845–5854, 2019. 3

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

[69] Shuangli Zhang, Weijian Xie, Guofeng Zhang, Hujun Bao, and Michael Kaess. Robust stereo matching with surface normal prediction. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2540–2547. IEEE, 2017. 1, 2, 3

[70] 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*, pages 12926–12934, 2020. 6

[71] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1327–1336, 2023. 6