Robust Synthetic-to-Real Transfer for Stereo Matching

Jiawei Zhang[1], Jiahe Li[1], Lei Huang[2], Xiaohan Yu[3], Lin Gu[4,5], Jin Zheng[1], Xiao Bai[1*]

[1]School of Computer Science and Engineering, State Key Laboratory of
Complex & Critical Software Environment, Jiangxi Research Institute, Beihang University
[2]SKLCCSE, Institute of Artificial Intelligence, Beihang University
[3]School of Computing, Macquarie University, Australia   [4]RIKEN AIP   [5]The University of Tokyo

Abstract

With advancements in domain generalized stereo matching networks, models pre-trained on synthetic data demonstrate strong robustness to unseen domains. However, few studies have investigated the robustness after fine-tuning them in real-world scenarios, during which the domain generalization ability can be seriously degraded. In this paper, we explore fine-tuning stereo matching networks without compromising their robustness to unseen domains. Our motivation stems from comparing Ground Truth (GT) versus Pseudo Label (PL) for fine-tuning: GT degrades, but PL preserves the domain generalization ability. Empirically, we find the difference between GT and PL implies valuable information that can regularize networks during fine-tuning. We also propose a framework to utilize this difference for fine-tuning, consisting of a frozen Teacher, an exponential moving average (EMA) Teacher, and a Student network. The core idea is to utilize the EMA Teacher to measure what the Student has learned and dynamically improve GT and PL for fine-tuning. We integrate our framework with state-of-the-art networks and evaluate its effectiveness on several real-world datasets. Extensive experiments show that our method effectively preserves the domain generalization ability during fine-tuning. Code is available at: `https://github.com/jiaw-z/DKT-Stereo`.

1. Introduction

Estimating 3D geometry from 2D images is a fundamental problem in computer vision. Stereo matching is a solution that identifies matching correspondences and recovers depth information through triangulation. With the development of deep learning, stereo matching networks have shown impressive performance on various benchmarks.

A major obstacle to their applications is the high cost of collecting real-world annotations. To make up for the lack

---

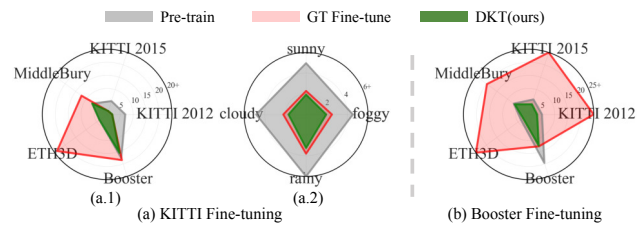*Corresponding author: Xiao Bai (baixiao@buaa.edu.cn).

Figure 1. Target-domain and cross-domain performance of networks pre-trained on synthetic data, fine-tuned with GT, and with our method (best visualized in colors). IGEV-Stereo [44] is used as the baseline model. D1 error (the lower the better) is used for evaluation. (a) and (b) fine-tune networks on the KITTI and Booster datasets, respectively. Our method achieves great performance in target and unseen domains simultaneously. We also evaluate the robustness of KITTI fine-tuned networks on DrivingStereo in (a.2), where our method is more robust to challenging weather.

of real data, existing networks are commonly pre-trained on synthetic data. With recent advancements in building domain generalized networks [7, 17, 21, 55], they have demonstrated strong robustness to unseen domains. However, as shown in Figure 1, fine-tuning pre-trained networks in real-world scenarios degrades the domain generalization ability. We provide visualization results in Figure 2. The degradation of robustness to unseen scenarios can render networks unreliable for real-world applications. To solve it, we explore what degrades the domain generalization ability of stereo networks during fine-tuning and provide a solution by combining Ground Truth (GT) with Pseudo Label (PL) predicted by pre-trained stereo matching networks.

Our exploration stems from the observation that using PL for fine-tuning stereo matching networks can preserve the domain generalization ability (*cf*. Section 3.3.1). Since there is always a gap between predicted PL and GT, we consider whether the difference between them implies certain knowledge. We are also motivated by previous studies [9, 14, 57] of knowledge distillation utilizing wrong response to non-target classes, which they refer to as "dark
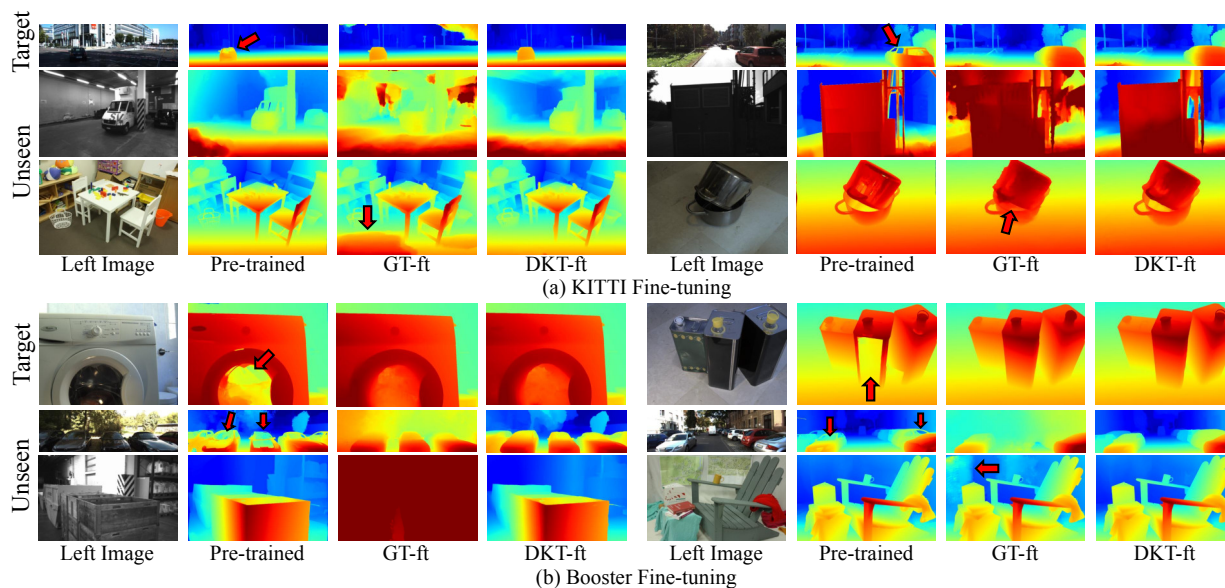
Figure 2. Visualization results on both target and unseen domains. IGEV-Stereo [44] is used as the baseline model. The network pre-trained on synthetic data shows robustness to unseen domains but can still fail to handle unseen challenges such as transparent or mirror (ToM) surfaces. Fine-tuning the network with GT improves the target-domain performance, however, it seriously degrades the domain generalization ability. Our DKT fine-tuning framework performs well on target and unseen domains simultaneously.

knowledge" of a transfer set by introducing PL with GT.

Starting from the observation, we aim to identify how GT and the predicted PL behave differently and further answer why fine-tuning with GT degrades the domain generalization ability of stereo matching networks. Specifically, we divide pixels into consistent and inconsistent regions based on the difference between GT and PL (*cf*. Section 3.1). GT and PL of the consistent region are similar, otherwise, it is defined as the inconsistent region. We empirically find two primary causes that degrade the domain generalization ability: (i) In the inconsistent region, networks encounter new knowledge not learned during pre-training on synthetic data. Learning new knowledge without sufficient consistent region for regularization seriously degrades the domain generalization ability. (ii) In the consistent region, stereo matching networks can still overfit the details of GT.

Based on the exploration, we propose the framework utilizing Dark Knowledge to Transfer (DKT) stereo matching networks. DKT consists of a frozen Teacher, an exponential moving average (EMA) Teacher, and a Student network, all initialized with the same pre-trained weights. We use the frozen Teacher to predict PL. The EMA Teacher is updated by the Student's weights, serving as a dynamic measure of consistent and inconsistent regions. With the EMA Teacher's prediction, we propose the Filter and Ensemble (F&E) module to improve disparity maps. For GT, F&E filters out the inconsistent region with a probability, resulting in the remaining GT that has a reduced inconsistent region, avoiding insufficient consistent region for regularization. It also performs an ensemble between GT and the

EMA Teacher's prediction in the consistent region, adding fine-grained permutations to prevent networks from overfitting GT details. For PL, F&E filters out the inconsistent region between PL and the EMA Teacher's prediction and ensembles them in the consistent region, enhancing the accuracy of PL predicted by the frozen Teacher. We train the Student with improved GT and PL jointly. After fine-tuning, we keep the Student for inference.

Our main contributions are as follows:

- To the best of our knowledge, we make the first attempt to address the domain generalization ability degradation for fine-tuning stereo matching networks. We divide pixels into consistent and inconsistent regions based on the difference between ground truth and pseudo label and demonstrate their varied roles during fine-tuning. Our further analysis of their roles identifies two primary causes of the degradation: learning new knowledge without sufficient regularization and overfitting ground truth details.

- We propose the F&E module to address these two causes, filtering out the inconsistent region to avoid insufficient regularization and ensembling disparities in the consistent region to prevent overfitting ground truth details.

- We introduce a dynamic adjustment for different regions by incorporating the exponential moving average Teacher, achieving the balance of preserving domain generalization ability and learning target domain knowledge.

- We develop the DKT fine-tuning framework, which can be easily applied to existing networks, significantly improving their robustness to unseen domains and achieving competitive target-domain performance simultaneously.

We believe this exploration will stimulate further consideration of stereo matching networks' robustness and domain generalization ability when fine-tuning them in real-world scenarios, crucial for their practical applications.

2. Related Work

Stereo Matching Networks. Estimating 3D geometry from 2D images is an important topic in computer vision [37, 39–41, 59], wherein stereo matching computes disparity between images. MC-CNN [51] first introduced a convolutional neural network to compute matching cost and predict disparity maps using SGM [15]. Since then, numerous deep-learning-based methods have been developed for stereo matching [3, 6, 16, 21, 42]. According to their strategy to perform cost construction, they can be divided into two categories. One type builds 3D cost volume and aggregates the cost with 2D convolutions. DispNetC [22] introduces end-to-end regression and builds a correlation-based 3D cost volume. Many works [19, 24, 35, 46, 49] adopt the correlation-based strategy and achieve impressive performance. Another type concatenates features to construct 4D cost volume and perform aggregation with 3D convolutions [3, 5, 16, 43, 53, 56]. Methods in this category can leverage more comprehensive information from the original images. Most recently, stereo matching networks [21, 38, 44, 52, 58] based on iterative optimization [34] are proposed and show state-of-the-art accuracy and strong robustness.

Robust Stereo Matching. Recently, building stereo matching networks that are robust to scenario changes has received increased interest. Existing studies can be categorized into joint generalization and cross-domain generalization types. Joint generalization involves training networks jointly on multiple datasets to perform well with a shared set of parameters [20, 30, 32]. Cross-domain generalization aims to improve generalization performance on unseen scenarios, with a current focus on synthetic-to-real generalization [2, 4, 7, 26, 36, 54, 55]. These methods enhance performance in unseen domains by avoiding networks overfitting synthetic data during pre-training. However, they have not investigated the domain generalization ability after fine-tuning. This paper is unique as it attempts to address the domain generalization degradation during fine-tuning.

Dark Knowledge. Knowledge Distillation (KD) is explored for transferring knowledge from one teacher model to another student model [11, 14, 48]. In the original form for classification, it constructs a transfer set by using the soft target distribution produced by teachers. Dark knowledge refers to the extra information beyond one-hot GT contained in the modified training set [27], associated with the distribution of responses to non-target classes. Building on this concept, a series of studies [9, 50, 57] explore the effects of incorrect predictions in KD. BAN [9] decomposes the KD into a dark knowledge term and ground truth component
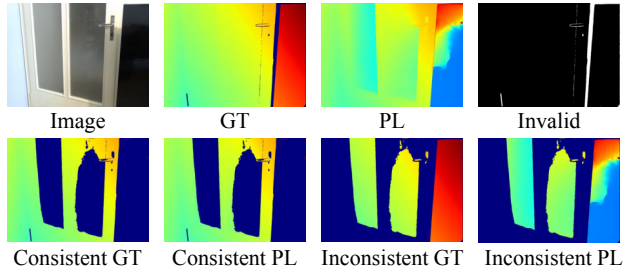

Figure 3. Visualization of each region resulting from our division. We divide GT and PL based on their consistency.

and develops a born-again procedure. DKD [57] reformulates KD into target class and non-target class distillation and decouples the two parts to transfer dark knowledge. We adopt this concept to decouple different regions of disparity maps, analyzing what degrades the domain generalization ability of stereo matching networks during fine-tuning.

3. Fine-tuning Stereo Matching Networks with Ground Truth and Pseudo Label

3.1. Preliminary and Definition

We explore whether stereo matching PL contains valuable knowledge that can preserve the domain generalization ability during fine-tuning and investigate the factors that make GT and PL behave differently. Stereo matching networks based on 2D or 3D cost aggregation are formed to obtain predictions by conducting Soft-Argmin on probability volume [16], while the recently developed iterative optimization methods directly predict continuous values [21, 44, 58] and achieve superior robustness. The difference between GT and PL lies at the level of continuous values, making our study applicable to any form of stereo matching network. Given a rectified stereo pair (I^l, I^r) with GT map D^*, a pre-trained stereo matching network θ predicts a dense disparity map $\hat{D} = \theta(I^l, I^r)$ as PL. We adopt the pixel distance $\hat{D} - D^*$ to measure responses to non-target disparities. To identify what makes GT and PL behave differently, we divide pixels into different regions based on the consistency between PL and GT. The visualization of each region is shown in Figure 3. We define different regions as follows:

Definition 1 *Consistent region $X_c(\tau)$. $X_c(\tau)$ contains pixels where the differences between GT and PL are less than the threshold τ. $X_c(\tau) = \{x||\hat{D}(x_i) - D^*(x_i)| < \tau\}$. This region represents areas where GT and PL align closely.*

Definition 2 *Inconsistent region $X_{inc}(\tau)$. $X_{inc}(\tau)$ contains pixels where the differences between GT and PL are not less than the threshold τ. $X_{inc}(\tau) = \{x||\hat{D}(x_i) - D^*(x_i)| \geq \tau\}$. Stereo matching networks can encounter unseen challenges in $X_{inc}(\tau)$.*

Definition 3 *Invalid region $X_{invalid}$. It contains pixels without available annotations due to the sparsity of GT.*

3.2. Experiment Setup

Dataset. Our experiments are conducted on KITTI [10, 23], DrivingStereo [47], Booster [25], Middlebury [28], and ETH3D [29]. KITTI and DrivingStereo datasets collect outdoor driving scenes, Booster and Middlebury provide indoor scenes, and ETH3D includes indoor and outdoor scenes in grayscale. We use the half resolution of Middlebury and DrivingStereo, and the quarter resolution of Booster. Except for online submissions, we split datasets for local evaluation due to the submission policy. More details are shown in *supplementary materials*. In our tables, we use blue for target domains and red for unseen domains.

Metric. We calculate the percentage of pixels with an absolute error larger than a certain distance for evaluation. 3 pixel is used for KITTI and DrivingStereo datasets. 2 pixel is used for Middlebury and Booster datasets. 1 pixel is used for ETH3D dataset.

Implementation. We adopt IGEV-Stereo [44] as the basic network architecture here, and results of other networks are shown in *supplementary materials*. We initialize networks with SceneFlow [22] pre-trained weights. For KITTI datasets, we fine-tune networks for 50k steps with a batch size of 8. For the Booster dataset, we fine-tune networks for 25k steps with a batch size of 2. We set the learning rate to 2e-4 for KITTI and 1e-5 for Booster. We use the data augmentation strategy in [21] during fine-tuning.

3.3. Ground Truth versus Pseudo Label

3.3.1 PL Preserves Domain Generalization Ability

We compare the fine-tuning of pre-trained networks using GT and PL. Contrast observations lead us to believe that the difference between GT and PL contains valuable information. Analyzing the validation curves presented in Figure 4, we observe a significant degradation of domain generalization ability during GT fine-tuning, occurring at an early stage. In contrast, there is no substantial degradation in domain generalization ability when using PL. This finding demonstrates that PL, as naive knowledge produced by pre-trained stereo matching networks, can preserve the original domain generalization ability.

3.3.2 Investigating the Difference between GT and PL

To understand the role of each region in GT and PL and to delve into the reasons behind the degradation of domain generalization ability, we conduct fine-tuning experiments by isolating different regions. The comparative results across various settings are presented in Table 1.

Learning new knowledge without sufficient regularization can degrade domain generalization ability. We start by comparing the baseline, which uses all valid regions of GT, with a setting that uses only the consistent region $X_c(3)$. By removing the inconsistent region $X_{inc}(3)$ of GT,

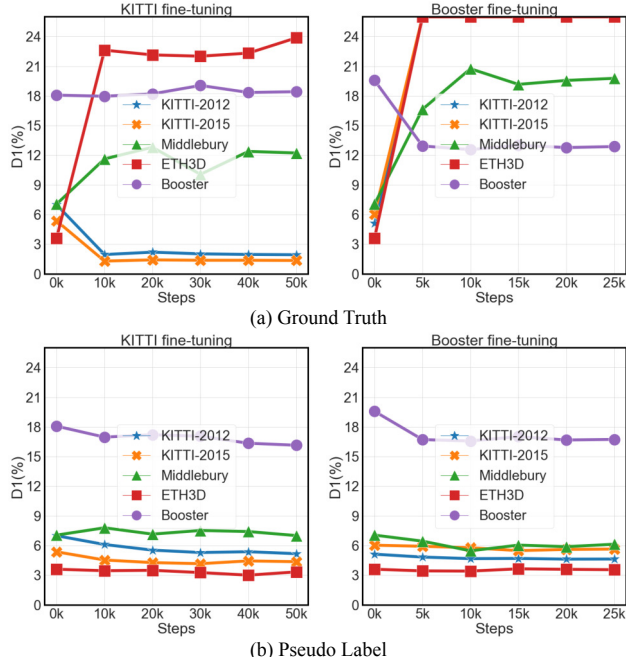(a) Ground Truth

(b) Pseudo Label

Figure 4. Evaluation curves of the target and cross domain performance during fine-tuning with GT or PL.

we observe an effective alleviation of generalization degradation, particularly notable for the Booster dataset. This observation highlights the significant impact of GT in the inconsistent region on degrading generalization ability. Next, we isolate only the inconsistent region $X_{inc}(3)$ for fine-tuning, where we observe a catastrophic degradation in generalization ability. Networks trained with only $GT(X_{inc}(3))$ even suffer difficulties generalizing to the target-domain validation data. This comparison highlights the crucial regularization role played by GT in the consistent region for networks learning to address unseen challenges in the inconsistent region. We conclude that while GT supervision in the inconsistent region is valuable for networks adapting to unseen challenges, learning new knowledge in this region without sufficient regularization can seriously degrade domain generalization ability.

Fine-grained details in GT can cause overfitting, degrading domain generalization ability. Unlike the inconsistent region, where handling is more challenging, networks find handling the consistent region less difficult since most necessary knowledge has been involved in synthetic data. Subsequently, we investigate whether learning in $X_c(\tau)$ can also result in domain generalization degradation. In our observations, a clear degradation in domain generalization ability occurs when using GT compared to PL in the same consistent region $X_c(3)$. Further tightening the threshold for the consistent region to $X_c(1)$ provides some relief, however, we find that even inaccurate components within a one-pixel range can lead to varying domain generalization abilities. We attribute this to stereo matching

Supervision	2012	2015	Midd	ETH3D	Booster
Training set	**KITTI 2012 & 2015**				
zero-shot	7.00	5.37	7.06	3.61	17.62
GT(valid)	**1.94**	**1.36**	12.23	23.88	18.43
GT($X_c(3)$)	2.17	1.69	10.97	19.44	17.39
GT($X_{inc}(3)$)	16.78	22.01	28.06	69.88	36.50
GT($X_c(1)$)	2.58	2.07	9.89	11.76	17.95
PL(all)	5.18	4.37	**7.01**	**3.34**	**16.15**
PL(valid)	5.82	4.90	8.30	3.65	16.46
PL($X_c(3)$)	3.72	3.11	8.46	3.53	16.17
PL($X_{inc}(3)$)	9.27	8.64	11.80	11.67	28.25
PL($X_c(1)$)	3.29	2.92	8.10	3.69	16.26
Training set	**Booster**				
zero-shot	5.13	6.04	7.06	3.61	19.49
GT(valid)	52.30	55.44	19.78	93.31	**12.88**
GT($X_c(3)$)	4.77	7.51	7.26	18.39	14.07
GT($X_{inc}(3)$)	93.35	96.38	50.09	99.24	30.25
GT($X_c(1)$)	4.58	7.66	6.93	17.15	13.78
PL(all)	4.64	5.65	6.14	3.56	16.74
PL(valid)	4.61	5.79	6.61	3.47	16.87
PL($X_c(3)$)	**3.23**	**4.36**	6.15	3.28	14.79
PL($X_{inc}(3)$)	5.70	6.53	7.35	4.55	21.75
PL($X_c(1)$)	3.63	4.89	**6.00**	**3.01**	14.26

Table 1. Results of fine-tuning stereo matching networks using different regions of GT or PL.

Supervision	2012	2015	Midd	ETH3D	Booster
Training set	**KITTI 2012 & 2015**				
GT	**1.94**	**1.36**	12.23	23.88	18.43
GT+PL(all)	3.24	2.75	**7.40**	**3.64**	**16.54**
GT+PL($X_c(3)$)	2.14	1.44	8.82	8.14	17.09
GT+PL($X_c(1)$)	1.97	1.38	8.47	11.29	17.75
Training set	**Booster**				
GT	52.30	55.44	19.78	93.31	12.88
GT+PL(all)	**7.14**	**10.69**	**8.02**	**10.30**	14.29
GT+PL($X_c(3)$)	41.95	45.27	11.74	65.63	**12.18**
GT+PL($X_c(1)$)	45.77	48.16	11.41	72.37	12.21

Table 2. Results of jointly training networks with GT and PL.

networks overfitting fine-grained details in GT, which negatively impacts domain generalization ability.

PL in the consistent region contains valuable knowledge to preserve the domain generalization ability. We compare using all PL with only using the consistent region $X_c(3)$ of PL. Fine-tuning with PL($X_c(3)$) preserves the domain generalization ability. It indicates that details in PL do not contribute to domain generalization degradation, which we consider a positive distinction from the degradation of using GT supervision. Additionally, we observe enhanced performance in the target domain, attributed to the increased accuracy of PL($X_c(3)$) compared to PL(all).

PL in the inconsistent region negatively impacts target-domain and unseen-domain performances. In the inconsistent region, PL is considered incorrect relative to GT. We explore whether incorrect predictions follow any useful patterns. Compared to using the correct PL($X_c(3)$), training networks with PL($X_{inc}(3)$) degrades both target and unseen domain performances. Consequently, we conclude that PL in the inconsistent region has negative impacts, and keeping the correct PL is still vital.

PL in the invalid region benefits domain generalization ability. The invalid region $X_{invalid}$ contains the potential consistent region and inconsistent region for PL that we cannot directly distinguish due to the unavailable of GT. Based on the conclusion that the inconsistent region contributes negatively, we question whether invalid regions are still necessary and conduct comparisons between using PL(all) and PL(valid). For KITTI datasets, we observe additional benefits when introducing invalid regions, demon-

strating utilizing PL in the valid region can be helpful. For the Booster dataset, we note that GT annotations are very dense and have fewer invalid regions, thus the effect is not that obvious.

3.3.3 PL as Regularization with GT

This section explores the role of PL as a regularization term when combined with GT for fine-tuning. We use PL of different regions with GT and make comparisons between different settings. The results are presented in Table 2.

Using all regions of PL as a naive strategy. We first fine-tune networks jointly with all available regions of GT and PL. While this approach mitigates the degradation in domain generalization, it exhibits inferior performance on target domains compared to using only GT, and its generalization is poorer than using only PL in Table 1. It indicates that this straightforward combination fails to leverage the advantages of both GT and PL.

Utlizing more accurate PL is crucial for satisfactory target-domain performance. The diminishing difference between GT and PL correlates with an enhancement in target-domain performance. In particular, when employing GT in conjunction with PL($X_c(1)$), the network achieves a target-domain performance comparable to one trained solely with GT. It's worth noting that using the consistent region of PL($X_c(\tau)$) for regularization may not be optimal, as it can compromise domain generalization ability compared to using all regions of PL.

More challenging scenarios degrade domain generalization ability more seriously. The evaluation of the Booster dataset, characterized by numerous ToM surfaces, reveals a significant impact on degrading generalization ability. Through a comparison between fine-tuning networks on the Booster and KITTI datasets, we see the degradation on Bosster is more serious. Across various experimental settings, excluding the use of all PL regions, fine-tuning on Booster consistently results in an obvious decline in network robustness, emphasizing the importance of addressing the degradation posed by fine-tuning networks in such challenging scenarios for real-world applications.
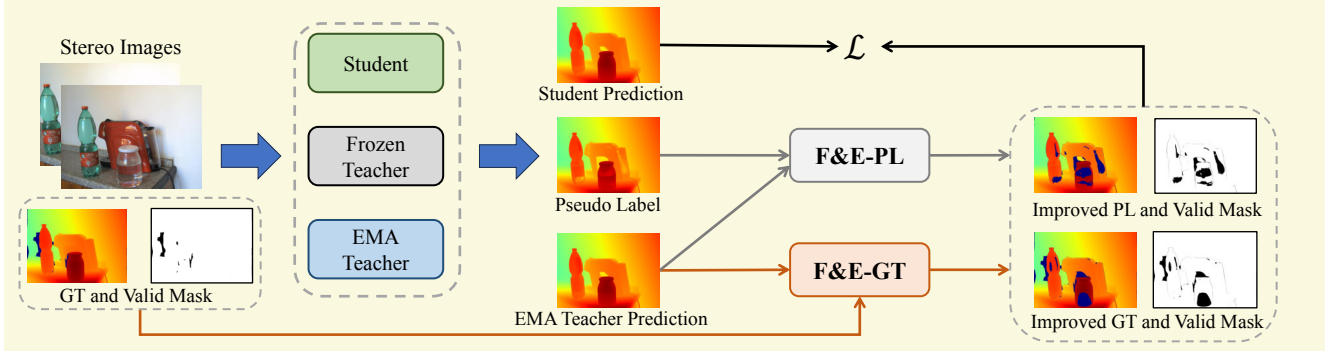
Figure 5. Overview of the DKT framework. It uses the prediction from the EMA Teacher to improve GT and PL during fine-tuning.

3.3.4 Summary

Our investigation reveals that both GT and PL play dual roles during fine-tuning. GT proves crucial for improving target-domain performance, particularly in handling unseen challenges. However, insufficient regularization and over-fitting GT details can degrade domain generalization ability. PL effectively preserves the generalization ability, however, learning incorrect predictions impact negatively. Moreover, a naive combination of GT and PL proves to be a conservative strategy that fails to leverage the full potential of both.

4. DKT Framework

4.1. Architecture

The core idea of DKT is to dynamically adjust the learning of different regions of GT and PL based on what networks have learned during fine-tuning. Figure 5 shows an overview of our framework.

DKT employs three same-initialized networks:

- Student θ_S is trained with GT and PL. After fine-tuning, we keep the Student for inference.
- Teacher θ_T predicts PL and is frozen during fine-tuning. It contains the original knowledge from pre-training.
- EMA Teacher $\theta_{T'}$ is an exponential moving average of Student. Here, it is used to measure what the Student has learned during fine-tuning.

The EMA Teacher is updated as a momentum-based moving average [1, 13, 33] of the Student:

$$\theta_{T'} = m\theta_{T'} + (1-m)\theta_S, \quad (1)$$

where $m \in [0, 1]$ is the momentum decay value for update speed control.

We propose the **Filter and Ensemble (F&E)** module, which utilizes the EMA Teacher to remove harmful regions and keep beneficial regions during fine-tuning. The division of different regions is conditioned on τ, which is set to 3 in our framework. Based on the characteristics of GT and PL, it has two variants termed F&E-GT and F&E-PL. We present the two variants in Figure 6.
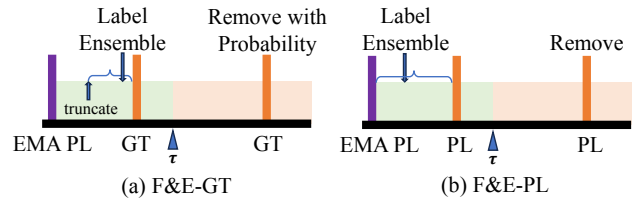

Figure 6. The two variants of Filter and Ensemble (F&E) applied to GT and PL. We use EMA PL to adjust different regions, with a threshold τ to control the operations on GT and PL.

F&E-GT. For GT disparity map D^*, we remove the inconsistent region $X_{inc}(\tau)$ to avoid them dominating fine-tuning. However, since we hope networks learn to cope with new challenges, annotations of these challenging regions are especially valuable. Thus F&E-GT removes the challenging inconsistent region with a probability based on the proportion in the whole valid region:

$$\overline{D}_{inc}^* = \begin{cases} D^*, & \text{rand}(0,1) > \frac{|X_{inc}(\tau)|}{|X_{valid}|} \\ \text{invalid}, & \text{otherwise} \end{cases} \quad (2)$$

It uses GT with a smaller probability for the more challenging region, thus avoiding new challenges disproportionately affecting learning. For the consistent region $X_c(\tau)$, we consider it an acceptable deviation and take continuous ensembling between GT D^* and the EMA Teacher's prediction $\hat{D}^{T'}$. We sum them with uniform weights and truncate at 1 pixel apart from GT D^*:

$$\alpha = \text{rand}(0,1) \quad (3)$$

$$\overline{D}_c^* = \alpha * D^* + (1-\alpha) * \hat{D}^{T'}. \quad (4)$$

$$\overline{D}_c^* = \min(\max(\overline{D}_c^*, D^* - 1), D^* + 1). \quad (5)$$

It adds fine-grained permutations to GT which prevents the Student from overfitting GT details. The final improved GT $\overline{D}^*$ is obtained by combining the improved inconsistent and consistent regions:

$$\overline{D}^*(x) = \begin{cases} \overline{D}_{inc}^*(x), & x \in X_{inc}(\tau) \\ \overline{D}_c^*(x), & x \in X_c(\tau) \end{cases} \quad (6)$$
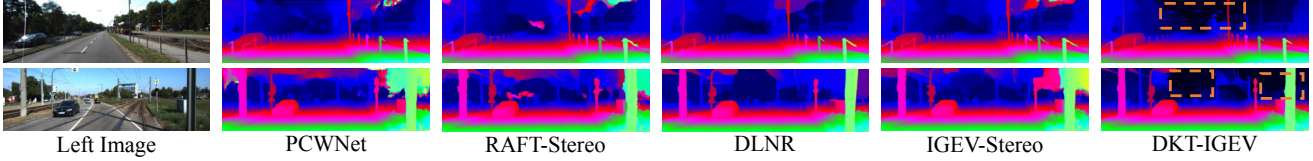
Figure 7. Qualitative results on the KITTI benchmark.

Method	KITTI 2012		KITTI 2015			DrivingStereo					Middlebury	ETH3D	Booster
	noc	all	bg	fg	all	sunny	cloudy	foggy	rainy	avg	>2px(%)	>1px(%)	>2px(%)
PSMNet [3]	1.49	1.89	1.86	4.62	2.32	6.36	4.98	10.63	21.39	10.84	21.85	88.87	32.88
GWCNet [12]	1.32	1.70	1.74	3.93	2.11	3.52	2.84	4.27	9.04	4.92	17.83	31.87	29.08
GANet [53]	1.19	1.60	1.48	3.46	1.81	3.77	3.44	4.26	10.46	5.47	18.79	14.40	30.65
PCWNet [31]	<u>1.04</u>	<u>1.37</u>	<u>1.37</u>	3.16	1.67	3.00	2.41	<u>1.72</u>	<u>3.41</u>	2.64	15.55	18.27	21.34
CGF-ACV [45]	**1.03**	**1.34**	**1.32**	3.08	<u>1.61</u>	3.22	3.19	6.69	17.50	7.65	23.76	36.92	22.42
RAFT-Stereo [21]	1.30	1.66	1.58	3.05	1.82	2.18	1.91	2.74	8.35	3.80	11.78	40.43	23.87
DLNR [58]	-	-	1.60	<u>2.59</u>	1.76	1.94	1.87	2.25	5.31	2.84	8.21	26.18	19.58
IGEV-Stereo [44]	1.12	1.44	1.38	**2.67**	**1.59**	2.28	2.21	2.51	3.80	2.70	11.87	24.28	18.33
DKT-RAFT(ours)	1.43	1.85	1.65	2.98	1.88	**1.85**	**1.46**	**1.32**	5.44	<u>2.52</u>	**7.51**	**2.28**	<u>15.35</u>
DKT-IGEV(ours)	1.22	1.56	1.46	3.05	1.72	<u>1.93</u>	<u>1.71</u>	1.96	**3.26**	**2.22**	<u>7.53</u>	<u>4.23</u>	**15.30**

Table 3. Results on the KITTI benchmark. We evaluate the domain generalization ability with checkpoints provided by authors.

F&E-PL. The frozen Teacher predicts PL $\hat{D}^T$, serving as important regularization during fine-tuning. F&E-PL aims to progressively enhance the accuracy of PL. We use the EMA Teacher's prediction $\hat{D}^{T'}$ to remove the inconsistent region of $\hat{D}^T$:

$$\hat{M} = |\hat{D}^T - \hat{D}^{T'}| < \tau. \tag{7}$$

And we combine $\hat{D}^T$ and $\hat{D}^{T'}$ to get improved PL $\overline{D}^T$.

$$\beta = \text{rand}(0, 1) \tag{8}$$

$$\overline{D}^T = \beta * \hat{D}^T + (1 - \beta) * \hat{D}^{T'}. \tag{9}$$

Training. Our final training loss is a weighted sum of disparity loss with the improved GT $\overline{D}^*$ and PL $\overline{D}^T$ within valid masks M^* for GT and $\hat{M}$ for PL:

$$\mathcal{L} = \mathcal{L}^{disp}(\hat{D}, \overline{D}^*, M^*) + \lambda \mathcal{L}^{disp}(\hat{D}, \overline{D}^T, \hat{M}), \tag{10}$$

where $\hat{D}$ is the Student prediction and λ is the balancing weight. After 5k steps of fine-tuning, we re-initialize the EMA Teacher with the Student weights. The re-initialized EMA Teacher predicts more accurate disparities, resulting in more accurate augmented GT and invalid-region PL.

4.2. Overall Performance

4.2.1 KITTI Benchmark

We compare the performance of DKT and published SOTA methods [31, 44, 45, 58] on the KITTI benchmark by considering target-domain and cross-domain performance. We utilize checkpoints provided by the authors to assess the performance of domain generalization. For methods to submit separate checkpoints on 2012 and 2015 benchmarks, we calculate the average of the two checkpoints. Table 3 shows the results. For target KITTI scenarios, we find our methods can achieve good but slightly worse performance than
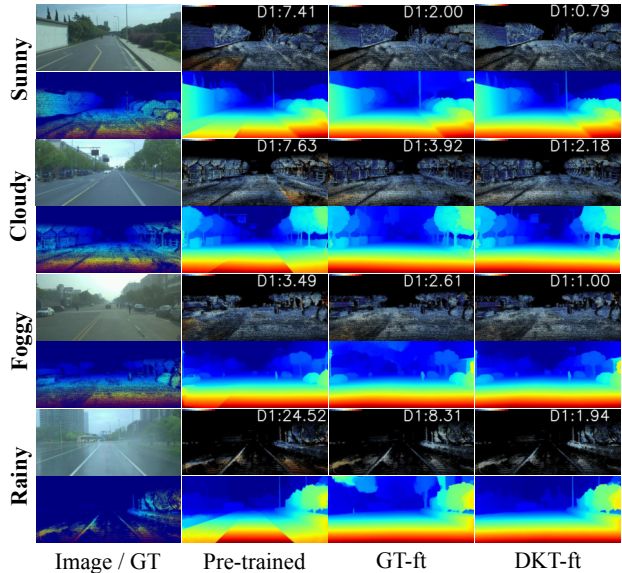

Figure 8. Qualitative results on the DrivingStereo dataset. We fine-tune the networks with KITTI datasets.

the base networks we use. We show the qualitative results in Figure 7, compared to other methods, DKT demonstrates reasonable predictions for weak texture regions. We evaluate how these models generalize to DrivingStereo, which contains unseen driving scenarios with challenging weathers. Among these published models, DKT achieves the strongest robustness by considering four challenging weathers. We visualize results in Figure 8. We evaluate if these models can generalize well to Middlebury, ETH3D, and Booster datasets. We see that DKT is the only method that can generalize well to these scenarios after fine-tuning.

$\mathcal{L}_{GT}$	$\mathcal{L}_{PL}$	F&E-GT	F&E-PL	2012	2015	Midd	ETH3D	Booster
Training set						KITTI 2012 & 2015		
✓				1.94	1.36	12.23	23.88	18.43
✓		✓		1.93	1.38	9.62	12.31	17.46
✓	✓			3.24	2.75	7.40	3.64	16.54
✓	✓	✓		3.26	2.79	7.03	3.24	15.19
✓	✓		✓	1.97	1.43	8.08	4.23	15.11
✓	✓	✓	✓	1.98	1.39	7.11	3.64	15.51
Training set						Booster		
✓				52.30	55.44	19.78	93.31	12.88
✓		✓		9.42	11.04	7.56	6.11	12.76
✓	✓			7.14	10.69	8.02	10.30	14.29
✓	✓	✓		3.69	4.88	7.01	4.32	14.21
✓	✓		✓	23.17	24.58	10.56	47.91	12.55
✓	✓	✓	✓	3.49	4.71	6.61	2.60	12.63

Table 4. Ablation study of each component of DKT framework on the KITTI 2012 and 2015, Middlebury, ETH3D, and Booster training sets.
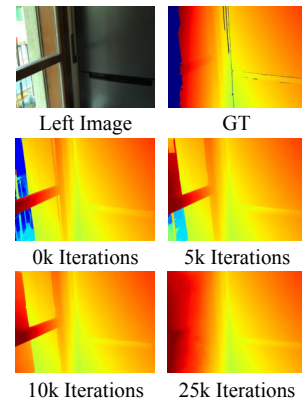

Figure 9. Predictions of the EMA Teacher during fine-tuning.

Method	Booster	2012	2015	Midd	ETH3D
CFNet [30] *	38.32	4.97	6.31	15.77	5.47
RAFT-Stereo [21] *	17.44	4.34	5.68	8.41	**2.29**
CFNet(ft) [30] †	29.65	51.52	68.24	18.42	79.98
RAFT-Stereo(ft) [21] †	10.73	17.02	23.33	15.98	89.05
PCVNet(ft) [52] ‡	**9.03**	35.54	38.60	22.88	75.21
DKT-RAFT(ours)	10.32	**3.69**	4.95	**5.94**	2.57
DKT-IGEV(ours)	14.11	3.76	**4.78**	6.94	3.13

Table 5. Results on Booster benchmark. For target-domain performance, we report their online results. For generalization evaluation: * uses authors' weights, †reproduces models in the same setting as online submissions, ‡the original implementation for submission uses CREStereo dataset [18] to augment Booster and we retrain it with only Booster training set to avoid unfair comparison.

4.2.2 Booster Benchmark

We follow [8, 25, 52] to fine-tune networks on the Booster dataset, which contains very challenging real-world scenarios. The results are presented in Table 5. We see fine-tuning networks with GT improves the predictions in the target domain compared to the synthetic data pre-trained networks. However, it also introduces a notable degradation in generalization ability when fine-tuning networks in such challenging scenarios. Using our DKT framework to fine-tune networks achieves competitive target-domain performances compared to the baseline fine-tuning. Moreover, DKT can help networks generalize knowledge learned from the challenging Booster data to unseen real-world scenes, improving the performance on KITTI and Middlebury than the synthetic data pre-trained models.

4.3. Ablation Study

We evaluate the effectiveness of DKT in utilizing the EMA Teacher for improving PL and GT in Table 4. Additionally, we provide visual insights into the EMA Teacher progressively predicts more accurate disparities in Figure 9.

F&E-GT. It improves GT by combining GT and the prediction of the EMA Teacher. It removes the inconsistent region of GT with certain probabilities and adds fine-grained permutations to GT. It also works when we only use GT supervision $\mathcal{L}_{GT}$ and do not introduce PL supervision $\mathcal{L}_{PL}$: only using F&E-GT to improve GT alleviates the domain generalization ability degradation and little impacts the target-domain performances. For using both PL supervision $\mathcal{L}_{PL}$ and F&E-GT, we show this combination can effectively preserve the domain generalization ability. However, using $\mathcal{L}_{PL}$ negatively impacts the target-domain performances because $\mathcal{L}_{PL}$ can introduce incorrect supervision, which we solve with F&E-PL.

F&E-PL. It plays an important role in improving target-domain performances by removing the inconsistent region of PL. It progressively identifies incorrect disparities of PL according to EMA Teacher's prediction, instead of directly comparing PL with GT which can degrade the domain generalization ability in Section 3.3.3. We show that only F&E-PL can alleviate the domain generalization degradation in KITTI, however, it cannot work well alone in the challenging Booster scenarios where using F&E-GT is necessary.

5. Conclusion

We aim to fine-tune stereo networks without compromising robustness to unseen domains. We identify that learning new knowledge without sufficient regularization and overfitting GT details can degrade the robustness. We propose the DKT framework, which improves fine-tuning by dynamically measuring what has been learned. Experiments show that DKT can preserve robustness during fine-tuning, improve robustness to challenging weather, and generalize knowledge learned from target domains to unseen domains.

References

[1] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. *arXiv preprint arXiv:1806.05594*, 2018. 6

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

[5] 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. 3

[6] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33:22158–22169, 2020. 3

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

[8] Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9244–9255, 2023. 8

[9] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 1, 3

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 4

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. 3

[12] 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. 7

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 6

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 3

[15] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 807–814. IEEE, 2005. 3

[16] 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. 3

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

[18] 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. 8

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

[20] 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. 3

[21] 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. 1, 3, 4, 7, 8

[22] 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. 3, 4

[23] 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. 4

[24] Matteo Poggi, Alessio Tonioni, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Continual adaptation for deep stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[25] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21168–21178, 2022. 4, 8

[26] 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. 3

[27] Peter Sadowski, Julian Collado, Daniel Whiteson, and Pierre Baldi. Deep learning, dark knowledge, and dark matter. In *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, pages 81–87. PMLR, 2015. 3

[28] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014. 4

[29] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017. 4

[30] 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. 3, 8

[31] 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. 7

[32] Zhelun Shen, Xibin Song, Yuchao Dai, Dingfu Zhou, Zhibo Rao, and Liangjun Zhang. Digging into uncertainty-based pseudo-label for robust stereo matching. *arXiv preprint arXiv:2307.16509*, 2023. 3

[33] Antti Tarvainen and Harri Valpola. Weight-averaged consistency targets improve semi-supervised deep learning results. corr. abs/1703.01780 (2017). *arXiv preprint arXiv:1703.01780*, 2017. 6

[34] 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. 3

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

[36] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 855–866, 2023. 3

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

[38] 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. 3

[39] Xiang Wang, Chen Wang, Bing Liu, Xiaoqing Zhou, Liang Zhang, Jin Zheng, and Xiao Bai. Multi-view stereo in the deep learning era: A comprehensive review. *Displays*, 70: 102102, 2021. 3

[40] Xiang Wang, Haonan Luo, Zihang Wang, Jin Zheng, and Xiao Bai. Robust training for multi-view stereo networks with noisy labels. *Displays*, 81:102604, 2024.

[41] Zihang Wang, Haonan Luo, Xiang Wang, Jin Zheng, Xin Ning, and Xiao Bai. A contrastive learning based unsupervised multi-view stereo with multi-stage self-training strategy. *Displays*, page 102672, 2024. 3

[42] Philippe Weinzaepfel, Vaibhav Arora, Yohann Cabon, Thomas Lucas, Romain Brégier, Vincent Leroy, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Improved cross-view completion pre-training for stereo matching. *arXiv preprint arXiv:2211.10408*, 2022. 3

[43] 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. 3

[44] 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, 3, 4, 7

[45] Gangwei Xu, Huan Zhou, and Xin Yang. Cgi-stereo: Accurate and real-time stereo matching via context and geometry interaction. *arXiv preprint arXiv:2301.02789*, 2023. 7

[46] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 3

[47] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 4

[48] Penghui Yang, Ming-Kun Xie, Chen-Chen Zong, Lei Feng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Multi-label knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17271–17280, 2023. 3

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

[50] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020. 3

[51] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, pages 1592–1599, 2015. 3

[52] 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. 3, 8

[53] 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. 3, 7

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

[55] 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. 1, 3

[56] 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. 3

[57] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 1, 3

[58] 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. 3, 7

[59] Xiaoqing Zhou, Xiang Wang, Jin Zheng, and Xiao Bai. Adaptive spatial sparsification for efficient multi-view stereo matching. *Acta Electronica Sinica*, 51(11):3079–3091, 2023. 3