Semi-Stereo: A Universal Stereo Matching Framework for Imperfect Data via Semi-supervised Learning

Xin Yue[1,*], Zongqing Lu[1], Xiangru Lin[2*], Wenjia Ren[1*], Zhijing Shao[2†],
Haonan Hu[1], Yu Zhang[2], Qingmin Liao[1]
[1]Tsinghua University [2]Prometheus Vision Technology Co., Ltd.

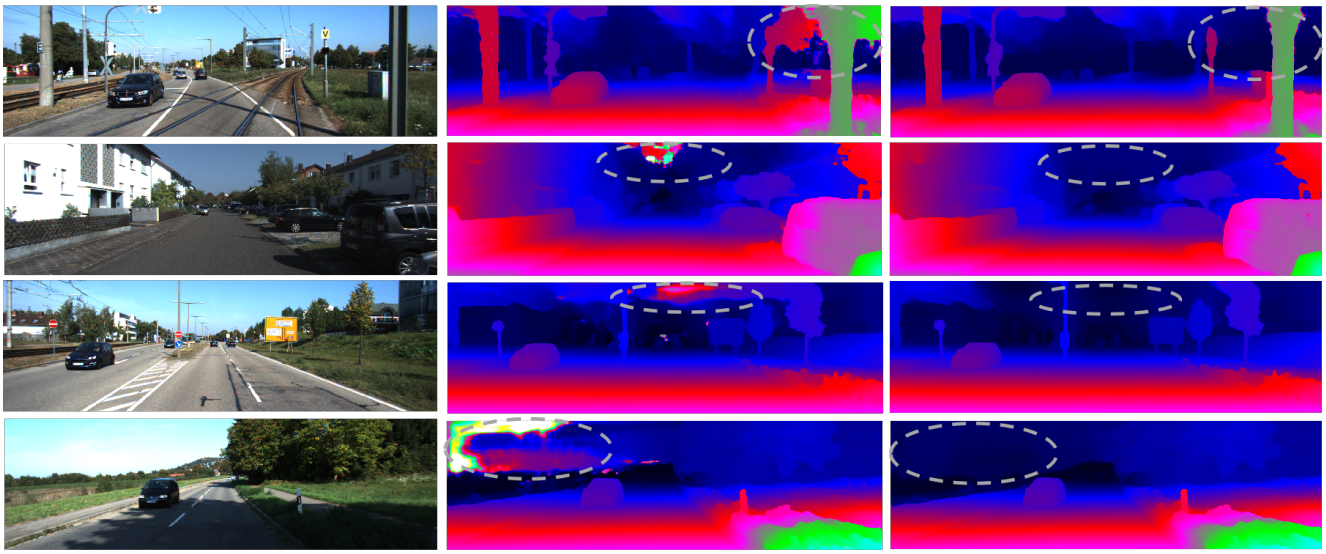


Figure 1: Comparisons of disparity inference results on sparse-annotated data. The inference results of PSMNet (first two rows) and GANet (last two rows) are shown in the middle column. In areas with sparse ground-truth, they experience a substantial prevalence of outliers. Our semi-supervised learning framework could significantly alleviate this phenomenon.

Abstract

Data matters in deep-learning-based binocular stereo matching. Obtaining a perfect dataset for stereo matching is hard and thus imperfect data is common in existing benchmark datasets, such as KITTI, ETH3D and Middlebury. The imperfectness typically has two forms: sparse-labeled data or even unlabeled data. Current stereo matching networks ignore the supervision from these imperfect data itself, even the semi-supervised networks often suffer from confirmation bias in the predictions. Besides, current methods lack a unified solution to utilize the supervision signal from those imperfect data. To mitigate this research gap, we propose Semi-Stereo, the first unified stereo matching framework empowered by the teacher-student paradigm where the teacher and the student networks are trained in a mutual-beneficial manner. To explore the rich knowledge in imperfect data, we propose a consistency regularization module with weak-strong augmentation strategies. Further, in order for the teacher to provide more reliable pseudo labels, we design a confidence module, powered by left-right consistency (LRC) check and disparity distribution entropy (DDE). Extensive experiments demonstrate Semi-Stereo produces accurate and consistent predictions in untrained semantic regions and improves the performance of baseline networks in multiple tasks, including domain adaptation and domain generalization.

1. Introduction

Stereo matching is a fundamental computer vision research topic, which aims to find the horizontal correspondences, *i.e.* disparity, between two rectified images [12, 20]. It has many practical applications such as robotics, UAVs,

*Authors contributed equally.
†Corresponding author.

autonomous driving, and augmented reality. Unlike traditional stereo matching methods, current deep-learning-based stereo matching networks rely heavily on the quality of the training datasets. However, acquiring high-quality ground-truth disparity maps is known to be a costly endeavor, hampered by depth scan devices. Therefore, real-world datasets are usually imperfect. The imperfectness typically has two forms: sparse-labeled data and unlabeled data. Specifically, real-world datasets are generally labeled with incomplete ground-truth. For example, as illustrated in Fig. 3, annotations mostly exist in specific foregrounds such as cars, roads, and trunks, while lost in most backgrounds such as sky, trees, and buildings. The sparse annotation leads to many outliers in untrained semantic regions. It can be observed that the inference has a noticeable error in the sky area. This situation may lead to undesirable consequences, such as recognizing the sky as a wall and then making a wrong decision in autonomous driving. However, the two visually different predictions have similar error maps under the existing benchmark evaluation system.

Existing convolution-based supervised methods heavily rely on the acquisition of labeled semantic information. Nevertheless, in scenarios where the semantic content of certain regions remains underexplored in sparsely labeled datasets, the network is prone to generating erroneous predictions within these areas [22]. These methods lack tailored designs to solve the problem posed by sparse labeling. For a completely unlabeled dataset, there are two corresponding tasks in the field of stereo matching: domain adaptation and domain generalization. However, most methods are network-specific or task-specific. So far, there is no uniform framework to solve the problems of sparse labeling, domain adaptation task, and domain generalization task together. The semi-supervised network has the potential to unify these issues, while a naive implementation of self-supervision is prone to suffer confirmation bias [27] through iterative finetuning.

To mitigate this research gap, we tackle this problem from a new *teacher-student* perspective. We propose Semi-Stereo, the first unified semi-supervised stereo matching framework empowered by the *teacher-student* paradigm. Specifically, the *teacher* and the *student* models are trained in a mutual-beneficial manner via exponential moving average (EMA), which corrects the confirmation bias in training and better guides the model optimization. Specifically, for the sparse-labeled data, we treat the pixels with ground-truth as labeled samples and those without ground truth as unlabeled samples. Note that we have not introduced additional data, but leveraged the information in unlabeled pixels, which is often ignored in supervised learning. For the unlabeled data, we follow current domain-adapted rules [26] to adapt models from large virtual scenarios [18] to real-world scenarios [5, 19, 23, 24], while treating the virtual data as labeled samples and the real-world data as unlabeled samples. For domain generalization tasks, the virtual data are treated as both labeled and unlabeled samples.

To explore the rich knowledge in imperfect data, we extend the study of consistency regularization from semi-supervised learning to stereo matching. We propose a consistency-based pseudo labeling regularization with weak-strong augmentation strategies. Specifically, 1) inspired by HSMNet [32], we propose a strong augmentation method with multiple thin vertical rectangular blocks in the left or right images to mimic occlusions, which is tailored for stereo matching, 2) to enhance the reliability of pseudo labels from the *teacher* model in Semi-Stereo, we design an effective confidence module powered by the left-right consistency (LRC) check [38] and disparity distribution entropy (DDE). This overall model design forces pixels with similar semantic contents to be consistent in disparity value, which strengthens the supervision signal from imperfect data. This motivation starkly contrasts that of current stereo matching networks. The Semi-Stereo framework(see Fig. 2) is universal and can be equipped with any stereo matching networks without changing their structures. This mechanism allows us to improve a model whether the dataset is dense-labeled or sparse-labeled, even unlabeled. Besides, we propose region-level (Infinity Metric) and image-level (Warp Consistency Metric) metrics to complete the consistency evaluation of stereo matching networks, which has been ignored in previous works. To summarize, our main contributions are:

- We analyze the imperfectness of the real-world datasets in stereo matching and identify the universality of existing benchmarks from the perspective of imperfect data, including disparity estimation, domain adaptation, and domain generalization tasks.
- We propose Semi-Stereo, the first universal stereo matching framework empowered by the *teacher-student* paradigm. It manifests that utilizing the supervision from imperfect data regions coupled with the labeled regions could further improve the performance of existing stereo networks, which has been underexplored in previous works. It is an important step forward to extend the study of semi-supervised learning to stereo matching. We also present two new metrics which can further evaluate the consistency of unlabeled data.
- Extensive experiments under different types of imperfect data with various stereo matching networks have demonstrated the superiority and generality of our simple yet effective Semi-Stereo, including disparity estimation, domain adaptation, and domain generalization tasks.

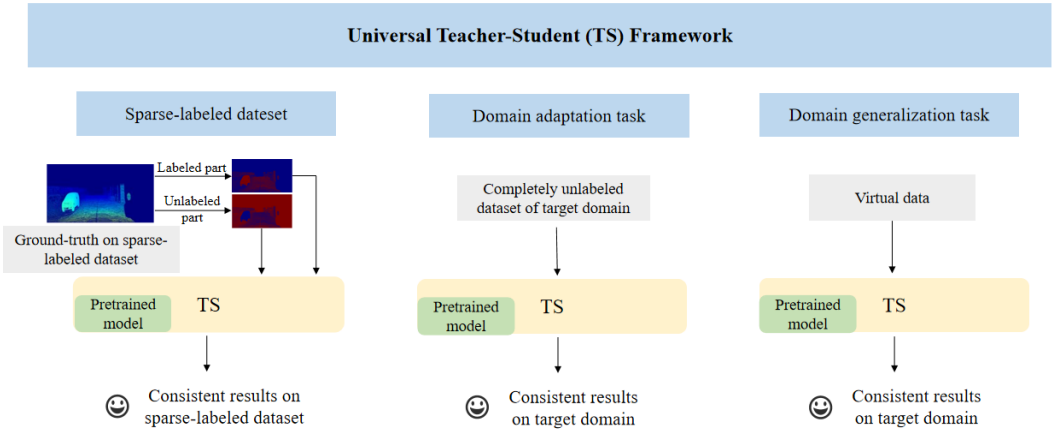
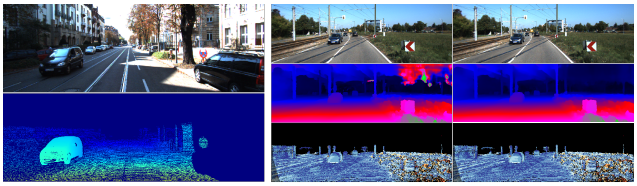


Figure 2: Illustration of our universal framework. Our framework could cope with three tasks: sparse-labeled dataset disparity inference task, domain adaptation task, and domain generalization task. The *teacher-student* (TS) framework is pre-trained on SceneFlow.



(a) Sparse-annotated data in training sets (b) Inference results from KITTI 2015

Figure 3: (a) It can be observed that large area of regions do not have ground-truth. (b) The top row displays the input. The middle row shows the disparity map, and the bottom row illustrates the evaluated error map. Note that large areas of false inference escapes the evaluation.

2. Related Work

Current Works on Sparse-labeled Data. Traditional stereo matching algorithms like [11, 8] directly produce sparse disparity maps. To fill in the rest missing regions, Ralli et al. [21] diffuse disparity values by using directional masks under the voting scheme while Beucher et al. [4] use hierarchical segmentation to propagate the sparse disparity to a dense and more accurate result. Deep learning based methods produce better predictions through the learning of semantic information. One example is the widely adopted multi-scale pyramid network [2, 31, 6, 25, 33], which can propagate the information from low resolution predictions to obtain high-resolution dense disparity maps. PSMNet [2] generates multi-scale feature maps and then concatenates them to enhance the context information. AANet [31] uses cross-scale cost aggregation to fuse the multi-scale cost volume. Further, the implementations of the cost aggregation help to gather information from adjacent pixels, including the 3D convolution [9, 2, 7, 32], GRU-based recurrent networks [15, 13, 29] and optimal transport [14], where the incomplete labeling issues can be greatly mitigated. However, due to sparse annotations, unlabeled semantic regions do not have ground-truth, and thus the semantic information of this region will not be learned, resulting in outlier during inference.

Domain Adaptation and Domain Generalization. Domain adaptation aims to use the target domain information to achieve better results with the pre-trained model, while domain generalization aims to achieve generalization without using any target domain information. For the domain adaptation task, Tonioni et al. [28] design an unsupervised and continuous online adaptation network with a pyramid strategy. Patrick et al. [10] propose a self-supervised procedure to adapt aerial images without ground truth. StereoGAN [16] designs a joint framework to achieve stereo adaptation with the bidirectional multi-scale feature reprojection and correlation consistency. Recently, Adastereo [26] proposes a novel domain adaptation pipeline to narrow the gaps between the source and target images, with color transfer, cost normalization, and self-supervised reconstruction. For the domain generalization task, Chuah et al. [3] consider that the fundamental problem that prevents domain generalization is shortcut learning. They make features insensitive to the low-level variants of the data through the information bottleneck theory. Zhang et al. [37] argue that maintaining feature consistency between matching pixels is important for promoting generalization ability and thus proposes a contrastive learning strategy across viewpoints. Regardless of domain adaptation or domain generalization tasks, existing network frameworks are network-specific or task-specific, and there is no unified framework to solve this problem.

Semi-supervised Binocular Stereo Matching. Semi-supervised stereo approaches are proposed to handle the problem of sparsely labeled data. Wang et al. [30] propose to utilize a pyramid voting module (PVM) to provide reliable pseudo labels for their OptStereo supervised learning. Patrick et al. [10] address the sparse-labeled problem through an iterative training strategy. In each training, the confident pseudo-labels are selected and sent into the next round of training. Zhou et al. [38] also iteratively update the network parameters under the guidance of left-right consistency check. These three works encounter two major drawbacks: 1) [30] causes the performance ceiling of OptStereo to not exceed PVM and thus the quality of pseudo-labels cannot be improved. In our work, the *teacher* network that gives pseudo-labels is constantly improving, 2) a naive pseudo-labeling [10, 38] causes confirmation bias during training. In our paper, the unbiased *teacher-student* paradigm could avoid this problem.

3. Method

3.1. Problem Definition

We focus on imperfect data in real-world scenarios. Specifically, imperfect data consists of: 1) sparse-annotated data D_s, and 2) completely unlabeled data D_u. Assuming that the mask regions having ground-truth annotations are m_{gt}, we aim to utilize the unlabeled regions in D_s or D_u, *i.e.*, $D_s\left[1-m_{gt}\right]$ or D_u, to ensure the consistency and accuracy of the network inference. Thus, the network is trained in a semi-supervised manner. Domain adaptation and domain generalization are two popular tasks in the field of stereo matching in recent years. They both aim to fix the performance degradation caused by domain gaps and to improve the applicability of stereo networks trained on a large synthetic dataset.

3.2. Framework Overview

In order to explore the consistency knowledge through the dataset itself, we introduce *teacher-student* paradigm to binocular stereo matching as Fig. 4. The overall process contains two stages. In the first stage, we obtain the model pre-trained from a large scale virtual dataset, *e.g.* SceneFlow. For simplicity, we omit this stage in the figure. In the second stage, we train *teacher* and *student* in the target domain. Based on this framework, we propose an imbalance augmentation and confidence module to let the *teacher-student* paradigm plays a better role in exploring consistency. First, they are both equipped with the pre-trained model and are fed with stereo pairs under varying degrees of augmentation. Second, the confidence module powered by LRC and DDE filters out the unreliable pseudo labels. Loss is conducted between the *teacher* and the *student* inferences and the EMA strategy [27] is adopted to update the weights of the *teacher*. The *teacher* and the *student* step forward mutually thus the quality of pseudo labels is improving continuously.

Motivation for introducing *teacher-student* paradigm comes from the consistency regularization between various inputs. Benefiting from the consistency regularization on the inference, the features on untrained semantic regions maintain stable. In this way, we utilize the unlabeled training data better and thus improve performance.

Thus, our method could be modeled as a data term and a regularization term. The former is from ground truth, while the latter is from *teacher-student* framework.

3.3. Imbalance Weak&Strong Augmentation

Data augmentation is the crucial part to help the *teacher-student* paradigm to explore consistency regularization. When the difference in contrast and chroma of the paired images is large, we require the network to still output consistent results, thus ensuring the realization of consistency regularization.

To enhance the reliability and robustness of the labels generated by the *teacher* network, we employ weak augmentations on the stereo pair inputs. Specifically, we simultaneously apply subtle adjustments ([0.8, 1.2]) to the brightness and contrast of both the left and right images.

In order to make the network have a better consistent regularization effect, we adopt a strong augmentation for the input of the *student* network. Specifically, we 1) randomly adjust the brightness and contrast on the two images to a larger degree ([0.5, 1.5]), noting that the adjusted scale is different for the left view and the right view. 2) randomly generates multiple thin vertical rectangular blocks in the right images to mimic occlusions, as illustrated in Fig. 5.

3.4. Confidence Module

The confidence module is also one of the parts of ensuring consistent inference. It is used to select the inference results of the *teacher* network to provide reliable pseudo-labels for the *student* network. The confidence module consists of the traditional left-right consistency (LRC) check and disparity distribution entropy (DDE) evaluation module designed in this paper.

Left-Right Consistency. We find in practice that the crosscheck of left and right predictions helps to remove most of the unreliable pseudo labels in the occluded areas. For a pixel in the left image with left-to-right disparity d_l, we shift its location horizontally by d_l to get the corresponding value d_{rl}. The consistency module masks out inconsistent pixels if the difference between d_l and d_{rl} is larger than a threshold δ_{lrc} :

$$mask_{lrc} = \left[(d_l - d_{rl}) < \delta_{lrc}\right] \tag{1}$$

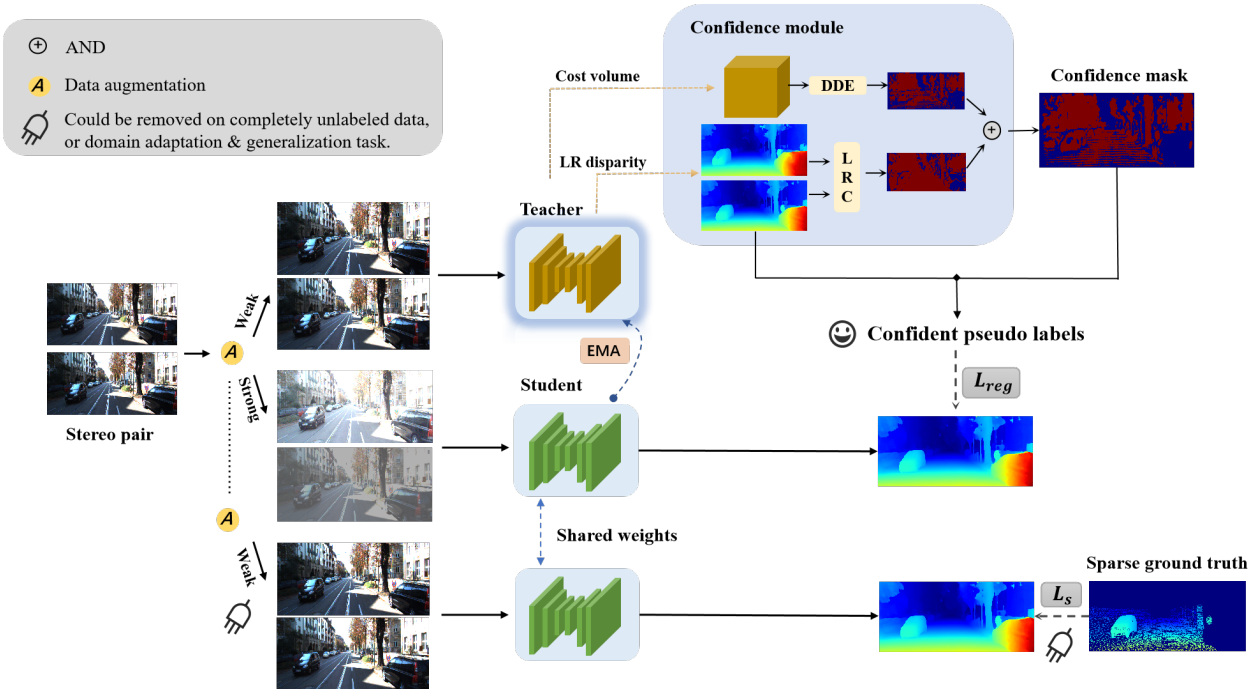
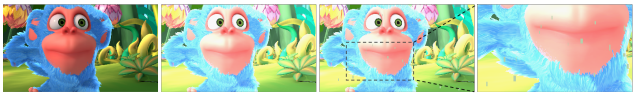


Figure 4: Overview of the proposed Semi-Stereo based on the *teacher-student* paradigm. The stereo pair after weak&strong augmentation is fed to the *teacher* and the *student*, respectively. For the sparse annotated data, another weak augmentation is performed and fed into the *student*. The confidence module selects the reliable pseudo labels provided by the *teacher* inference. Loss is conducted between pseudo labels and the inference of the *student*. The weight update of the *teacher* will also take the weight of the *student* into account through the EMA strategy, thus the *teacher* and the *student* are progressing together. Therefore, the quality of pseudo labels improves continuously as the training process.



(a) Original image (b) Brightness & contrast augmentation (c) Random occlusions (d) Enlarged details

Figure 5: Illustration of augmentation.

Disparity Distribution Entropy. Another commonly observed phenomenon in stereo is the multi-modal distribution of the cost volume across the disparity range. A single prominent peak in the cost function suggests that the network is more confident in its prediction, whereas a cost volume exhibiting a multi-peak distribution typically indicates unreliability in the corresponding regions. We propose evaluating the uncertainty of the pseudo-labels by computing the disparity distribution entropy of the cost volume:

$$H(p) = -\sum_{d=0}^{D_{max}} p(d)\log p(d) \tag{2}$$

With the cost volume built on the full disparity range D_{max}, we apply softmax to get $p(d)$ for every disparity candidates. We apply the threshold δ_{dde} to mask out uncertain pixels according to the disparity distribution entropy:

$$mask_{dde} = [H(p) < \delta_{dde}] \tag{3}$$

We set $\delta_{lrc} = 1$ and $\delta_{dde} = 0.2$ in our experiments. The final pseudo-label mask m_t could be obtained by combining results from two confidence modules:

$$m_t = mask_{dde}\ \&\ mask_{lrc} \tag{4}$$

3.5. Loss and Update Strategy

For the inference results of the strong augmentation input of the *student* network, we use pseudo labels for supervision to realize consistent regularization:

$$L_{reg} = \frac{1}{N_u}\sum_{i=0}^{N_u} \|\hat{d}_{stu,i}^{strong} - d_{pse,i}\|_1 \tag{5}$$

where, N_u is the number of unlabeled pixels, $\hat{d}_{stu}^{strong}$ is the inference result of strong augmentation input of the *student*. d_{pse} is the pseudo label generated by the *teacher* network defined as:

$$d_{pse} = \hat{d}_{tch}^{weak} * m_t \tag{6}$$

where, $\hat{d}_{tch}^{weak}$ is the inference result of weak augmentation input of the *teacher* network. We also construct the loss between the output of the weak augmented data through the *student* network and real-world ground truth as follows.

$$L_s = \frac{1}{N_s}\sum_{i=0}^{N_s} \|\hat{d}_{stu,i}^{weak} - d_{gt,i}\|_1 \tag{7}$$

where, N_s is the number of labeled pixels, $\hat{d_{stu}^{weak}}$ is the inference result of weak augmentation input of the *student* network. If the dataset is completely unlabeled, this part could be removed from the framework.

After the gradient backpropagation of the *student* network, we update the weights of the *teacher* to avoid the confirmation bias. The implementation of the EMA strategy is to take the weights of the *student* into account and update the weights of the *teacher* as follows:

$$\theta_t^{'} = \alpha\theta_{t-1}^{'} + (1-\alpha)\theta_t \tag{8}$$

where, θ is the student weights, $\theta^{'}$ is the *teacher* weights, and α controls how much the *student* network update is considered. Through the EMA, we avoid the problem that the weights of the *teacher* network are never updated, which can easily bias the inference results of the network. We set $\alpha = 0.9996$ in our experiments. The quality of pseudo labels could improve continuously as the network training, and the *teacher* and the *student* are progressing mutually (see Sec. 5.4).

4. Consistency Evaluation Metrics

In Fig. 3, we observe that two wildly different predictions have similar error maps. The reason is that the regions of inconsistent inference do not have ground-truth. For a fair comparison of the consistency performance of different networks, two metrics are designed to measure the consistency of the inferences, that is, the infinity metric and the warp consistency metric.

Infinity Metric. The intuition behind the infinity metric is that regions with infinite distance are not rare in real world, especially the sky. The sky part is trained insufficiently due to the lack of semantic information, which causes large outliers. We propose to calculate the mean and variance in this area: the closer the mean and variance are to zero, the better the consistency and accuracy of the network.

Warp Consistency Metric. The warp consistency metric, on the other hand, measures the consistency between the predictions between the left and right images. For a pixel in the left image with RGB color I_l, we 1) shift from its location by the prediction d_l of left-to-right disparity to get the corresponding location on the right image; 2) read the prediction d_r from the right-to-left disparity and shift back; 3) read again from the left image to get RGB value I_{lrl}. The warp consistency metric is measured by the difference between I_l and I_{lrl}:

$$C_{warp} = \begin{cases} 0 & d_l \ll d_{rl} \\ |I_l - I_{lrl}| & Others \end{cases} \tag{9}$$

Note that pixels, where d_l is much less than d_r, are considered possibly occluded regions thus we do not include them in the metric. The warp consistency metric should be considered as a necessary but insufficient condition for reliable disparity predictions.

5. Experiments

5.1. Datasets

We use SceneFlow [18] (Flyingthings3D, Monkaa and Driving), KITTI 2012 [5], KITTI 2015 [19], Middlebury [23], ETH3D [24] and InStereo2K [1] for evaluation. SceneFlow is a large synthetic stereo dataset, containing 35454 training image pairs and 4730 test image pairs. KITTI are collections of real-world driving scenes, containing 194(KITTI 2012)/200(KITTI 2015) training image pairs and 195(KITTI 2012)/200(KITTI 2015) testing image pairs. Middlebury is a high-resolution dataset of indoor scenes, with 23 image pairs for training and/or validation and 15 testing image pairs with full, half, and quarter resolutions. ETH3D contains 27 grayscale image pairs from indoor and outdoor scenes with sparse-labeled ground truth. InStereo2K is a large dataset in indoor scenes, including 2000 pairs for training and 50 pairs for testing.

5.2. Implementation Details

For the sparsely labeled task experiments, we train our semi-stereo network with the backbone pre-trained on SceneFlow. For the cross-domain experiments, we first train the backbone network on SceneFlow, applying the color transformation module [17] on domain adaptation experiments. Then, we train our Semi-Stereo on the target domain images without using any ground truth. The target domain images are from SceneFlow for the domain generalization task or from the real world images for the domain adaptation task. The specific setting of each model is illustrated in the supplementary document.

Table 1: Ablation studies for data augmentation and EMA.

Method	D1_All %
PSMNet	1.81
TS-PSMNet(w/o DataAug)	1.75
TS-PSMNet (w/o EMA)	1.73
TS-PSMNet	**1.71**

5.3. Ablation Study

Ablation of Data Augmentation and EMA. We divide KITTI 2015 into training set (80%) and validation set (20%)

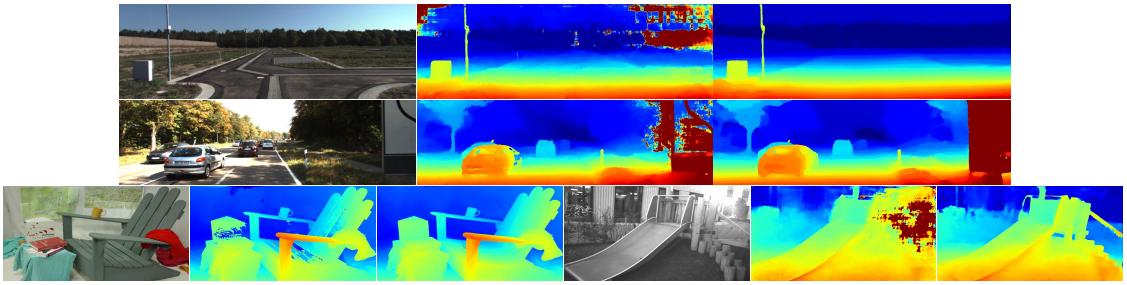


Figure 6: Adaptation examples on four datasets. The top row showcases results on KITTI 2012, followed by the second row presenting results on KITTI 2015. The left side of the third row depicts results on Middlebury, while the right side exhibits those on ETH3D. The middle images of each sub-figure is the predictions from the model just trained on SceneFlow. The right images of each sub-figure is the inference results under our domain adaptation method.

Table 2: Ablation study of the confidence module.

LRC	DDE	Threshold Error Rate(%)	
		KITTI 2012	KITTI 2015
		4.5	6.3
✓		4.1	5.7
	✓	3.8	5.4
✓	✓	**3.5**	**4.1**

Table 3: Semi-supervised learning experiments on InStereo2K.

Method	Training		Testing			
	labeled splits	unlabeled splits	D1	1px	2px	3px
PSMNet	*1*	-	3.23	13.44	6.16	4.19
TS-PSMNet(ours)	*1*	*2-7*	**1.26**	**10.77**	**4.14**	**2.54**
PSMNet	*1-2*	-	2.74	11.87	5.51	3.72
TS-PSMNet(ours)	*1-2*	*3-7*	**1.11**	**10.23**	**3.85**	**2.37**

and compare our framework with or without data augmentation and EMA. Tab. 1 shows the metrics on the validation set. DataAug means Data Augmentation. If we remove both modules, our network will degenerate as the baseline network. The results show that DataAug and EMA are effective respectively. The best performance is achieved by using them together.

Ablation of Confidence Module. We test the effect of the confidence module and explore the impact of LRC and DDE respectively through the domain adaptation testing on KITTI. As presented in Tab. 2, the results show that LRC and DDE are effective respectively. And the best performance is achieved by using LRC and DDE together. Further, we ablate the hyperparameters of LRC and DDE in the supplementary document.

Ablation of Hyperparameters. We move the ablations of hyperparameters of the data augmentation, the unsupervised loss weights, and the effect of color transformation module to the supplementary document.

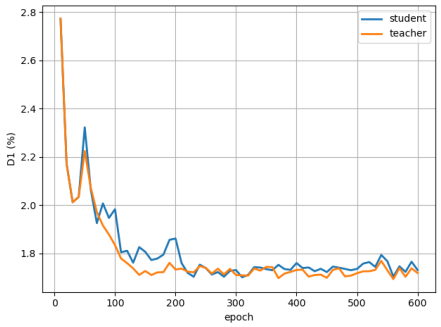


Figure 7: Comparison of the *teacher* and the *student* on KITTI 2015 validation. It shows that the *teacher* and the *student* make progress together.

5.4. Additional Validations

***Teacher-student* Architecture.** We visualize the performance(D1) on KITTI 2015 validation set during each training epoch (Fig. 7). It can be observed that the *teacher* and the *student* progress together in the process of mutual learning, and the *teacher* is good enough to guide the *student*.

Effect on InStereo2K [1]. The training set is divided into 7 parts (split 1 to 7) and we design two settings: (1) Split 1 is chosen as the labeled data and Split 2-7 are selected as the unlabeled data, (2) to further increase the percentage of the labeled data, we choose Split 1 and Split 2 as the labeled data and Split 3-7 as the unlabeled data. As shown in Tab. 3, we compare our Semi-Stereo with the supervised learning baseline (PSMNet) and it can be observed that our Semi-Stereo has the ability to mine useful supervisory signals from the unlabeled data, thus further improving the performance over the supervised baseline.

5.5. Comparisons with Supervised Networks

We integrate our Semi-Stereo approach into PSMNet [2], GANet [35], AANet [31], and subsequently compare its performance with the respective original models. For brevity, we refer to the enhanced models as TS-PSMNet, TS-AANet and TS-GANet.

Table 4: Benchmark results on KITTI test sets.

Method	KITTI 2015		KITTI 2012	
	D1-All	D1-Noc	3px-All	3px-Noc
PSMNet	2.32	2.14	1.49	1.89
TS-PSMNet	**2.06**	**1.86**	**1.30**	**1.71**
GANet	1.81	1.63	1.19	**1.60**
TS-GANet	**1.74**	**1.56**	**1.12**	**1.60**
AANet	2.55	2.32	1.91	2.42
TS-AANet	**2.52**	**2.28**	**1.87**	**2.37**

Table 5: Comparisons of the Infinity-Mean, Infinity-Variance and Warp Consistency on KITTI test sets.

Method	KITTI 2012			KITTI 2015		
	Inf-Mean	Inf-Var	Warp	Inf-Mean	Inf-Var	Warp
PSMNet	15.40	352.46	41.98	10.65	50.44	43.68
TS-PSMNet	**8.26**	**43.46**	**34.77**	**7.32**	**25.53**	**31.50**
GANet	10.15	75.09	40.66	15.40	358.83	37.99
TS-GANet	**9.83**	**43.07**	**36.57**	**8.43**	**40.01**	**36.35**
AANet	10.32	46.23	35.71	11.45	106.43	33.87
TS-AANet	**9.99**	**43.36**	**33.33**	**10.11**	**48.07**	**31.09**

Table 6: Domain generalization (up) and domain adaptation (down) on four validation sets. Threshold error rate (%) is utilized for measurement.

Method	KITTI		Middlebury	ETH3D
	2012	2015		
HD^3[34]	23.6	26.5	37.9	54.2
gwcnet[7]	20.2	22.7	34.2	30.1
GANet[35]	10.1	11.7	20.3	14.1
DSMNet[36]	6.2	6.5	13.8	6.2
PSMNet[2]	15.1	16.3	25.1	23.8
FC-PSMNet[37]	7.0	7.5	18.3	12.8
ITSA-PSMNet[3]	5.2	5.8	12.7	9.8
TS-PSMNet(ours)	**5.0**	**5.4**	**12.0**	**5.6**
StereoGAN[16]	-	12.1	-	-
Ada-ResNetCorr[26]	5.1	5.0	12.7	5.8
Ada-PSMNet[26]	3.6	**3.5**	8.4	4.1
TS-PSMNet(ours)	**3.5**	4.1	**8.1**	**4.0**

Results on KITTI Benchmark. We submit our inference results of our models to KITTI 2012 and KITTI 2015 stereo benchmark for evaluation. As shown in Tab. 4, our methods demonstrate superior performance compared to those trained under the supervised paradigm. Fig. 1 shows the superiority of our method, especially the robustness of the textureless regions. Intuitively, we observe a notable enhancement in most backgrounds, attributable to the consistency achieved by our approach. Since most background pixels are excluded from the evaluation, the enhanced benchmark metric signifies that our proposed framework not only enhances predictions on unlabeled pixels but also boosts performance on labeled ones.

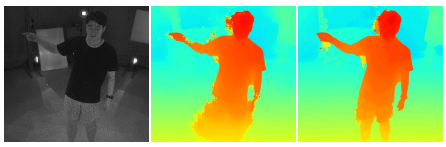


Figure 8: Qualitative comparisons between PSMNet (middle) and TS-PSMNet (right) in the domain adaptation task.

Results on Consistency Metrics. Tab. 5 shows the consistency metrics on the test sets of KITTI 2012 and KITTI 2015. Our Semi-Stereo surpasses the original ones both in Inf-Mean and Inf-Var. It could be proved that our framework does polish the disparity predictions in the infinite-distance areas. As shown in Tab. 5, the warp consistency mean errors are lower than the original ones.

5.6. Cross-Domain Comparisons

Domain Generalization. For fair comparisons, we choose the same backbone as [26]. Tab. 6 shows the domain generalization results, where we only use SceneFlow for TS framework training, instead of utilizing the target domain. Our TS-framework effectively complements the backbone networks, demonstrating its significant role in enhancing the performance.

Domain Adaptation. Tab. 6 shows the quantitative comparisons while Fig. 6 showcases qualitative examples of domain adaptation. Employing real-world datasets for the target domain, our framework attains optimal performance on KITTI 2012, Middlebury and ETH3D.

Qualitative Evaluation of Domain Adaptation. Apart from validating our approach on public datasets, we also evaluate its domain adaptation capabilities using infrared images, which are commonly encountered in structured light scenarios. We compare the PSMNet which is pre-trained on SceneFlow and the adapted TS-PSMNet on the infrared images without ground truth. After the domain adaptation, we get more accurate predictions with finer details (see Fig. 8).

6. Conclusion

Recognizing the challenge posed by imperfect data in stereo matching, we introduce Semi-Stereo and propose the first universal framework empowered by the *teacher-student* paradigm. This framework could effectively harnesses supervision from both imperfect and labeled data regions, enabling enhanced performance. Besides, we also present Infinity Metric and Warp Consistency Metric to complete the consistency evaluation of stereo matching networks, which is often ignored in previous works. Extensive experiments under different types of imperfect data with various stereo matching networks have demonstrated the superiority and generality of our simple yet effective Semi-Stereo.

References

[1] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63(11), 2020. 6, 7

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5410 – 5418, 2018. 3, 7, 8

[3] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, pages 13012 – 13022, 2022. 3, 8

[4] Sebastien Drouyer, Serge Beucher, Michel Bilodeau, Maxime Moreaud, and Loic Sorbier. Sparse stereo disparity map densification using hierarchical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10225 LNCS:172 – 184, 2017. 3

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3354 – 3361, 2012. 2, 6

[6] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2492 – 2501, 2020. 3

[7] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3268 – 3277, 2019. 3, 8

[8] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328 – 341, 2008. 3

[9] 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

[10] Patrick Knobelreiter, Christoph Vogel, and Thomas Pock. Self-supervised learning for stereo reconstruction on aerial images. In *International Geoscience and Remote Sensing Symposium*, pages 4379 – 4382, 2018. 3, 4

[11] Kurt Konolige. Small vision systems: Hardware and implementation. *Robotics Research*, pages 203 – 212, 1998. 3

[12] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1738 – 1764, 2022. 1

[13] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, pages 16242 – 16251, 2022. 3

[14] 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 International Conference on Computer Vision*, pages 6177 – 6186, 2021. 3

[15] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *Proceedings - 2021 International Conference on 3D Vision*, pages 218 – 227, 2021. 3

[16] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, pages 12754 – 12763, 2020. 3, 8

[17] Haoyu Ma, Xiangru Lin, Zifeng Wu, and Yizhou Yu. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4050 – 4059, 2021. 6

[18] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4040 – 4048, 2016. 2, 6

[19] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3061 – 3070, 2015. 2, 6

[20] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5314 – 5334, 2022. 1

[21] J. Ralli, J. Diaz, and E. Ros. A method for sparse disparity densification using voting mask propagation. *Journal of Visual Communication and Image Representation*, 21(1):67 – 74, 2010. 3

[22] Wenjia Ren, Qingmin Liao, Zhijing Shao, Xiangru Lin, Xin Yue, Yu Zhang, and Zongqing Lu. Patchmatch stereo++: Patchmatch binocular stereo with continuous disparity optimization. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 2315–2325, 2023. 2

[23] Daniel Scharstein, Heiko Hirschmuller, York Kitajima, Greg Krathwohl, Nera Nei, X. Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8753:31 – 42, 2014. 2, 6

[24] Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*, pages 2538 – 2547, 2017. 2, 6

[25] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 13901 – 13910, 2021. 3

[26] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10323 – 10332, 2021. 2, 3, 8

[27] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1196 – 1205, 2017. 2, 4

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

[29] 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

[30] 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. 4

[31] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1956 – 1965, 2020. 3, 7

[32] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5510 – 5519, 2019. 2, 3

[33] Chengtang Yao, Yunde Jia, Huijun Di, Pengxiang Li, and Yuwei Wu. A decomposition model for stereo matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6087 – 6096, 2021. 3

[34] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6037 – 6046, 2019. 8

[35] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H.S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 185 – 194, 2019. 7, 8

[36] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12347 LNCS:420 – 439, 2020. 8

[37] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, pages 12991 – 13001, 2022. 3, 8

[38] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576 – 1584, 2017. 2, 4