# Semi-Stereo: A Universal Stereo Matching Framework for Imperfect Data via Semi-supervised Learning
## *(Supplementary Document)*

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.

## 1. Implementation Details

**General Settings**. We set $\delta_{lrc} = 1.0$ and $\delta_{dde} = 0.2$ for the confidence module. The *teacher* weights are updated every iteration, and the coefficient of Exponential Moving Average (EMA) for the *teacher* is 0.9996. For the weak augmentation, we set the brightness and contrast to [0.8, 1.2] on both the left and right images. For the strong augmentation, we first randomly set the brightness and contrast to [0.5, 1.5] respectively on each of the image pair. Then we randomly generate 30 thin vertical rectangular blocks on the right image to mimic occlusions, whose $width$ is within [1, 5] and $length$ is within [5, 10].

**Sparse-labeled Data Experiments on KITTI [2, 5]**. In stage 1, we pre-train the model on the large-scale virtual dataset, SceneFlow, following previous works. Then, we conduct the mutual learning of the *teacher* and the *student* models. For instance, the original PSMNet [1] undergoes training for 300 epochs, initially with a learning rate of 0.001 for the first 200 epochs, which is then reduced to 0.0001 for the remaining 100 epochs. Thus, as for our TS-PSMNet, we warm up TS-PSMNet for 30 epochs with the learning rate of 0.001, based on the pre-trained weights on SceneFlow. Then we train our Semi-Stereo model for 300 epochs, with a learning rate of 0.001 for the first 200 epochs and 0.0001 for the last 100 epochs. The submitted results of KITTI 2012 and KITTI 2015 stereo benchmark are qualitatively shown in Fig. 10, Fig. 11 and Fig. 12.

**Domain Adaptation Experiments**. PSMNet is utilized as the baseline network for our study. We follow the same parameter settings described as above. In stage 1, PSMNet is trained on SceneFlow [4] using the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$). The color transfer operation [3] is performed on the images for data pre-processing. Images are randomly cropped to 256×512. The max disparity $D_{max}$ is set to 192. We train our Semi-Stereo with a constant learning rate of 0.001 for 20 epochs. In stage 2, we train our Semi-Stereo on the target domain images without using any ground truth. The target domain images are from the real world. On KITTI 2012 [2], KITTI 2015 [5], Middlebury [6] and ETH3D [7], we train our models for 100, 100, 100, 50 epochs, respectively. For Middlebury, we just use the half-resolution validation set. Errors are the percent of pixels whose end-point errors are greater than the specified threshold. We use the standard evaluation thresholds in our experiments: 3px for KITTI, 2px for Middlebury, and 1px for ETH3D. Fig. 8 shows the qualitative results.

**Domain Generalization Experiments**. We use PSMNet as a baseline network. In stage 1, we follow the same experiment settings as in the domain adaptation experiments but without the color transfer method included. In stage 2, we train our Semi-Stereo on the SceneFlow images without using any ground truth and train our models for 10K iterations. We test our model on KITTI 2012 [2], KITTI 2015 [5], Middlebury [6] and ETH3D [7], with the same metrics as in the domain adaptation experiments. Fig. 9 shows the qualitative results.

## 2. Extended Description of Our Metrics

### 2.1. Infinity Metric

**Pseudo Label Reliability in the Region of the Sky**. The lack of supervision in unlabeled regions such as the sky is due to the measurement limit of the LiDAR. Based on this fact, prior works generally do not perform well in the region of the sky. However, our Semi-Stereo can fit well in these regions under the guidance of the sky pseudo label. Fig. 1 in our main paper shows the improvement in the sky regions. We plot the disparity value of the pseudo label of the sky during training, as shown in Fig. 4. We can see that pseudo labels from TS-PSMNet shows superior performance over PSMNet. Although pseudo labels for those regions are not perfectly accurate, our Semi-Stereo manifests that utilizing the supervision from those regions coupled with the labeled regions could further improve the performance of existing stereo networks. This has been under-explored in previous works.

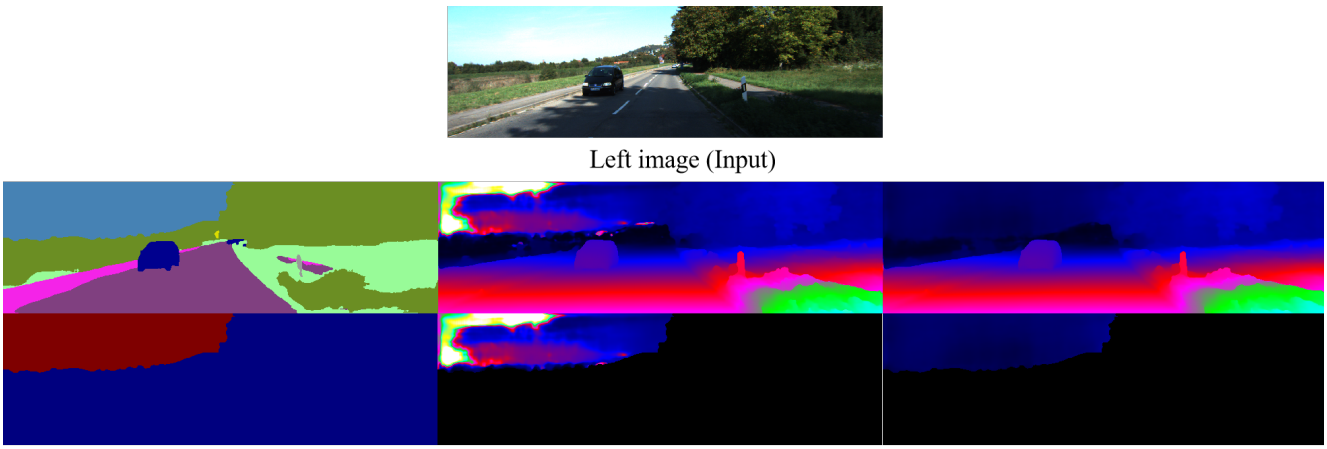



Figure 1: An example of Infinity evaluation on KITTI 2015. (a) represents the result of semantic segmentation, and the sky mask (from top to bottom). (b) denotes the inference result of PSMNet and the disparity of the sky (from top to bottom). (c) represents the inference result of TS-PSMNet and the disparity of the sky (from top to bottom).

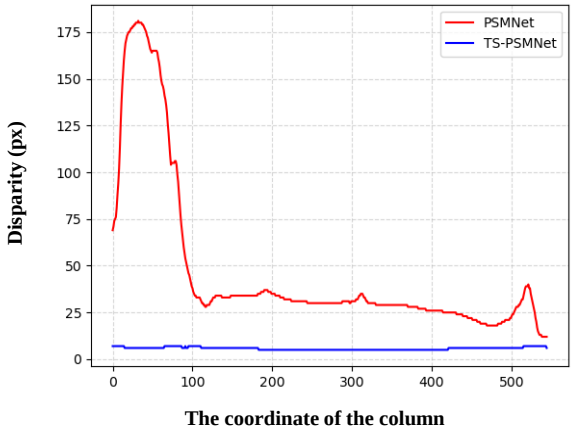



Figure 2: Comparisons of disparity values between PSMNet and TS-PSMNet in the region of the sky. We choose to visualize the result for row 120.

**Visualization**. Our Infinity Metric effectively judges the disparity of pixels with infinite distance, such as the sky. Fig. 1 shows the segmentation of the sky and the inference results from GANet and TS-GANet. Fig. 2 shows the disparity distribution of ours is better and more stable than the baseline.

### 2.2. Warp Consistency Metric

**Comparisons with D1(%)**. In order to evaluate the accuracy of our Warp Consistency Metric, we compare the common D1(%) with our consistency metric. We divide KITTI 2015 into a training set (80%) and a validation set (20%). During the training, we save the model for each epoch and test the D1(%) and consistency metric. From Fig. 5 it can be observed that they share similar trends, which justifies that our metric is representative when no ground-truth labels are available. Therefore, in the absence of ground truth, it is effective to use the warp consistency metric to quantify the analysis of the inference results.

**Visualization**. Fig. 3 is an example of using the double shift with left and right disparity. Zoom in for comparisons. It can be observed that the double-shift result of the baseline has been distorted in detail, while ours maintains strong consistency with the original image.

## 3. Extended Description of Experiments

### 3.1. Ablation of the Confidence Module

We have explored the impact of LRC and DDE in our main paper. Here, we show the experimental results of setting different thresholds for LRC and DDE. We perform our domain-adaptation experiments on KITTI 2012. As Tab. 1 showing, we set the $\delta_{lrc} = 1.0, 2.0, 3.0$ and $\delta_{dde} = 0.1, 0.2, 0.5$, and find our model is insensitive to the threshold. Fig. 6 shows the threshold error rate(%) of the validation set during training.

### 3.2. Ablation of the Imbalance Weak&Strong Augmentation

We verify the effectiveness of our proposed Imbalance Weak&Strong Augmentation with the following experiments: (1) remove the brightness and contrast augmentation (chromatic augmentation); (2) remove the vertical multi-blocks (random occlusion); (3) replace our random occlusions with an asymmetric mask in [8]. We also perform our domain-adaptation experiments on KITTI 2012. As is

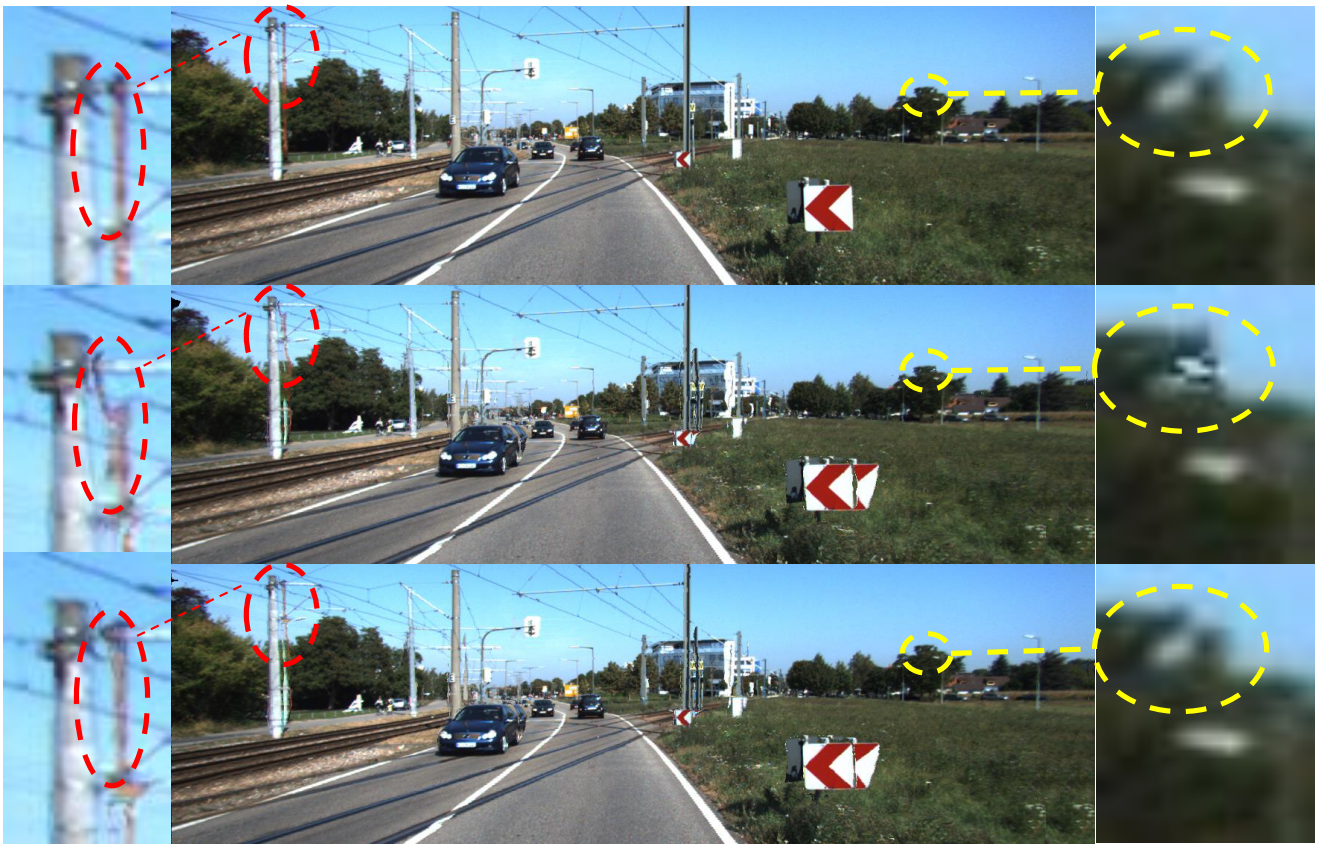


Figure 3: Comparisons of the consistency of the predictions. From top to bottom there are the original left image, the result of PSMNet after the double-shift, and the result of TS-PSMNet after the double-shift.

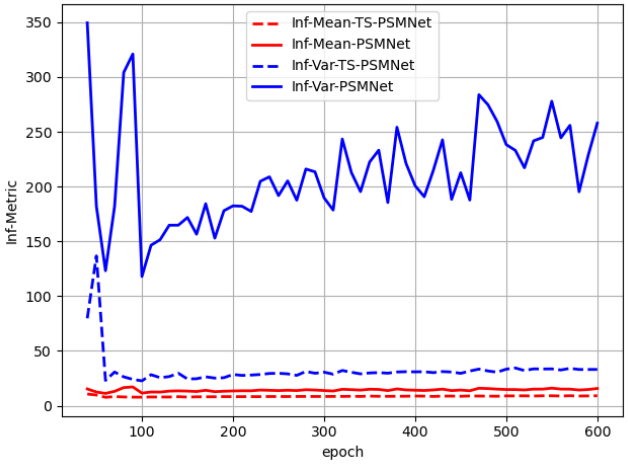



Figure 4: Comparisons of the quality of the pseduo labels from PSMNet and TS-PSMNet. TS-PSMNet shows superior performance over PSMNet across the training.

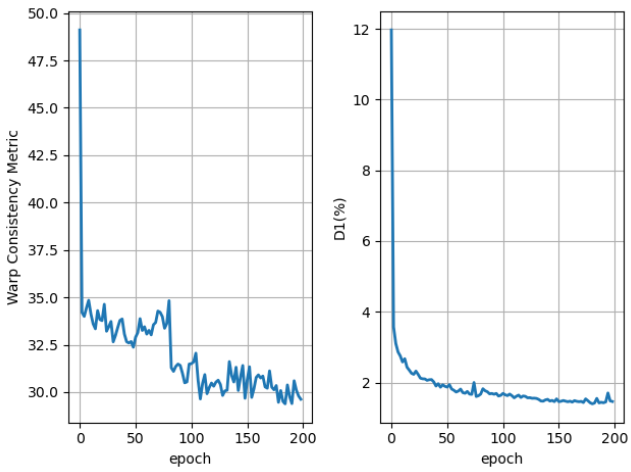



Figure 5: Comparisons between our Warp Consistency Metric and the D1(%) metric.

shown in Tab . 2, the chromatic augmentation is a crucial step, and our vertical multi-block is more effective than the random occlusion method in [8].

Table 1: Ablation study of the confidence module. The model we use is TS-PSMNet. The best performance is obtained by using both confidence modules.

| LRC | DDE | KITTI 2012(3px) |
|---|---|---|
| ✗ | ✗ | 4.5 |
| ✗ | ***0.1*** | 3.88 |
| ✗ | ***0.2*** | 3.78 |
| ✗ | ***0.5*** | 3.98 |
| ***1.0*** | ✗ | 4.05 |
| ***2.0*** | ✗ | 4.10 |
| ***3.0*** | ✗ | 4.16 |
| ***1.0*** | ***0.2*** | **3.53** |

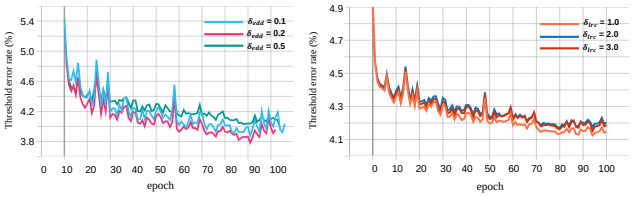



Figure 6: The threshold error rates(%) on validation set.

Table 2: Ablation study of the Imbalance Weak&Strong Augmentation. We use TS-PSMNet in this experiment.

| Chromatic | Random occlusion | | KITTI 2012(3px) |
|---|---|---|---|
| | Asymmetric mask [8] | Vertical multi-blocks | |
| ✗ | ✗ | ✓ | 4.13 |
| ✓ | ✗ | ✗ | 3.71 |
| ✓ | ✓ | ✗ | 3.64 |
| ✓ | ✗ | ✓ | **3.53** |

Table 3: Impact of loss weight on KITTI 2015. EPE is tested on the validation set.

| gt_loss_weight | unsup_loss_weight | EPE |
|---|---|---|
| ***1.0*** | ***1.0*** | **0.76** |
| ***1.0*** | ***0.3*** | **0.76** |
| ***1.0*** | ***3.0*** | 0.87 |

Table 4: Ablation study of the color transformation module.

| Method | Color-Transfor | KITTI 2012(3px) |
|---|---|---|
| PSMNet | ✗ | 15.10 |
| TS-PSMNet | ✗ | 4.30 |
| TS-PSMNet | ✓ | **3.53** |

### 3.3. Ablation of the Weight of $L_{reg}$ and $L_s$

We study the impact of the weight coefficient of $L_{reg}$ and $L_s$ on End-Point-Error (EPE). We divide KITTI 2015 into a training set of 160 pairs and a validation set of 40 pairs. we keep the weight of $L_s$ as 1.0 and vary the weight of $L_{reg}$ as 0.3, 1.0 and 3.0. When the weight of $L_{reg}$ is 0.3 or 1.0, they achieve equal performance quantitatively (see Tab. 3). However, qualitatively, setting the weight of $L_{reg}$ to 0.3 performs inferior to setting the weight of $L_{reg}$ to 1.0 in textureless regions like the sky (see Fig. 7). This manifests that enforcing stronger consistency regularization in our Semi-Stereo could lead to better performance in textureless regions. In this sense, 1.0 is chosen as the final setting.

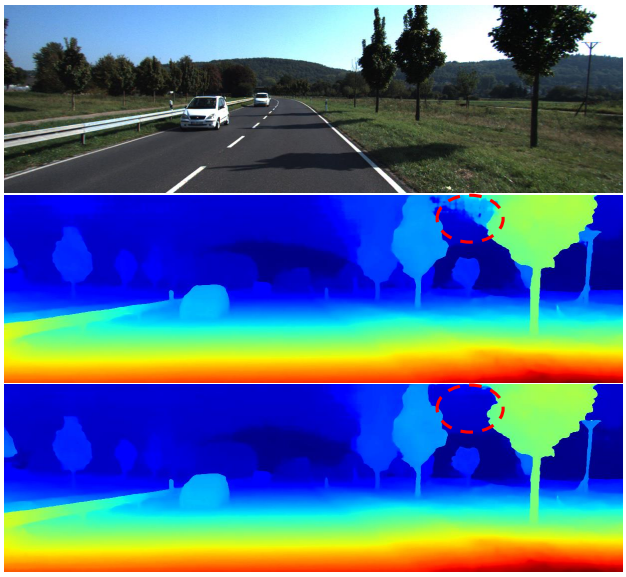


Figure 7: Comparison of using small unsupervised weight and large unsupervised weight. The raw RGB image is presented at the Top. The middle and bottom are the results of loss weights of 0.3 and 1 respectively. This example comes from KITTI 2015 validation set.

### 3.4. Domain Adaptation without Color Transformation

Here we show the effect of removing the color transformation module [3] in Stage 1. We test on KITTI 2012. As is shown in Tab . 4, it is proved that our Semi-Stereo still plays an essential role in the domain adaptation without color transformation.

## References


[1] 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. 1

[2] 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 Con-*

*ference on Computer Vision and Pattern Recognition*, pages 3354 – 3361, 2012. 1

[3] 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. 1, 4

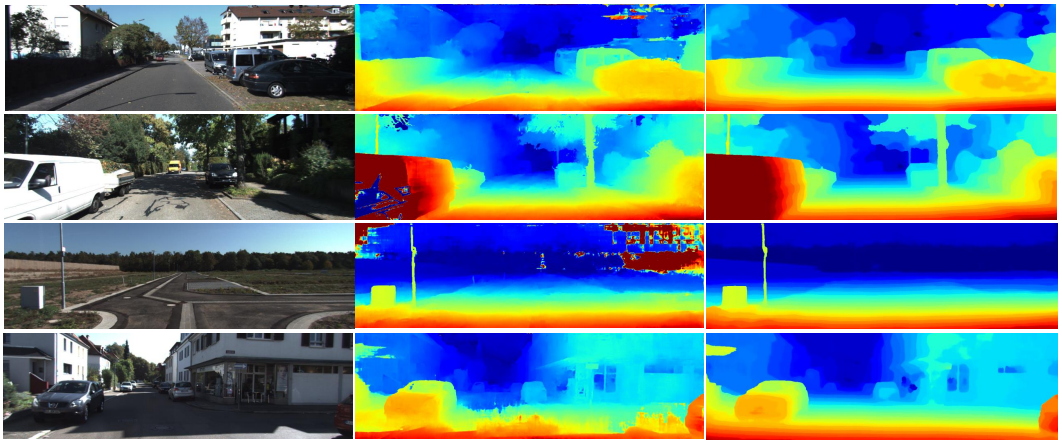
[4] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4040 – 4048, 2016. 1

[5] 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. 1

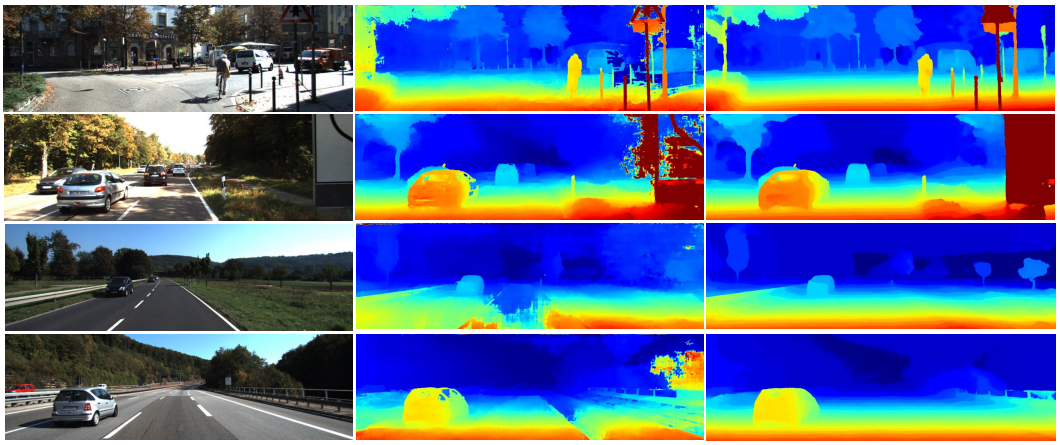
[6] Daniel Scharstein, Heiko Hirschmuller, York Kitajima, Greg Krathwohl, Nera Nei, X. Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8753:31 – 42, 2014. 1

[7] 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. 1

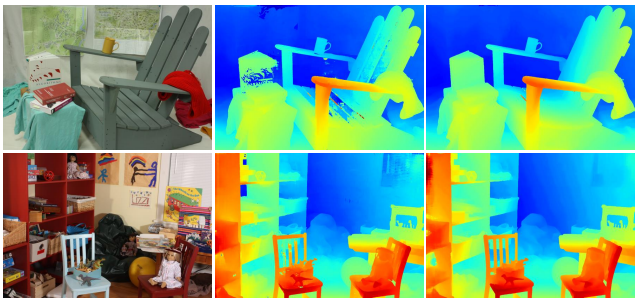
[8] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5510 – 5519, 2019. 2, 3, 4



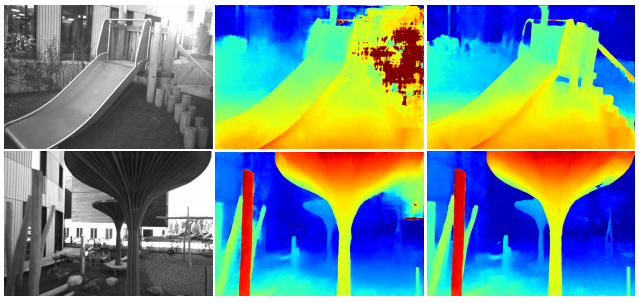
(a)



(b)

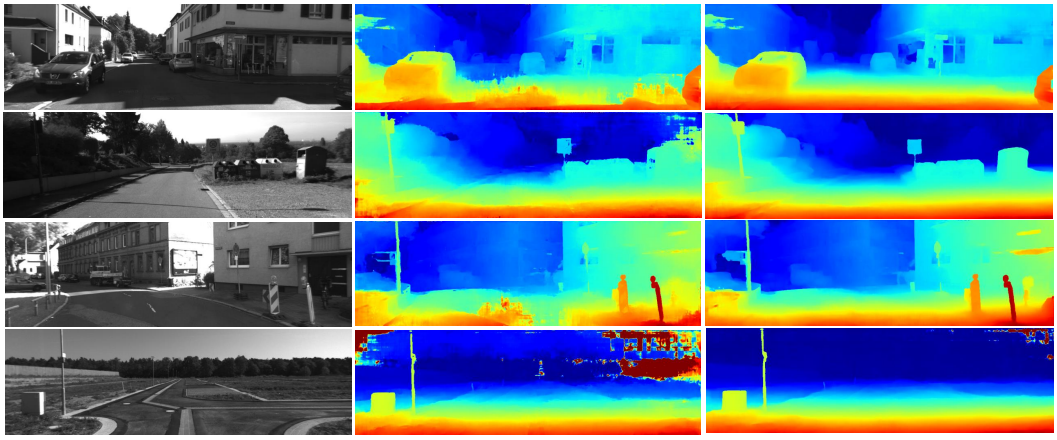


(c)

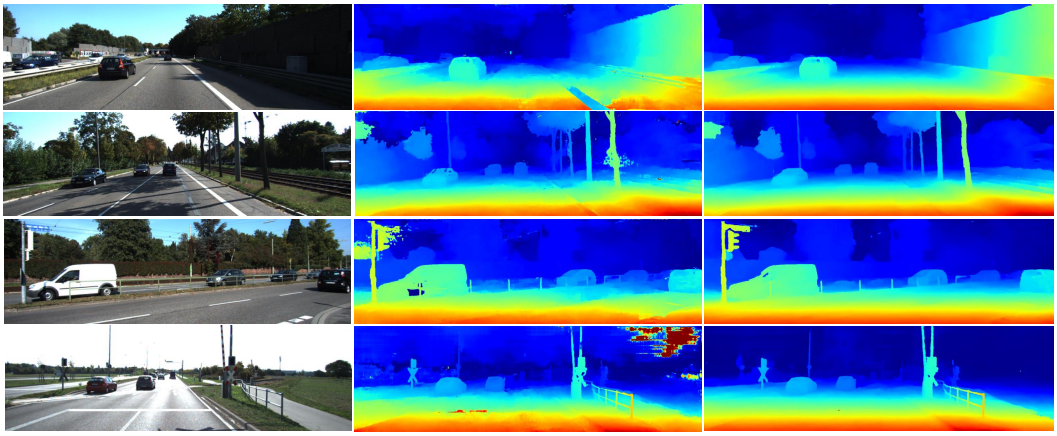


(d)

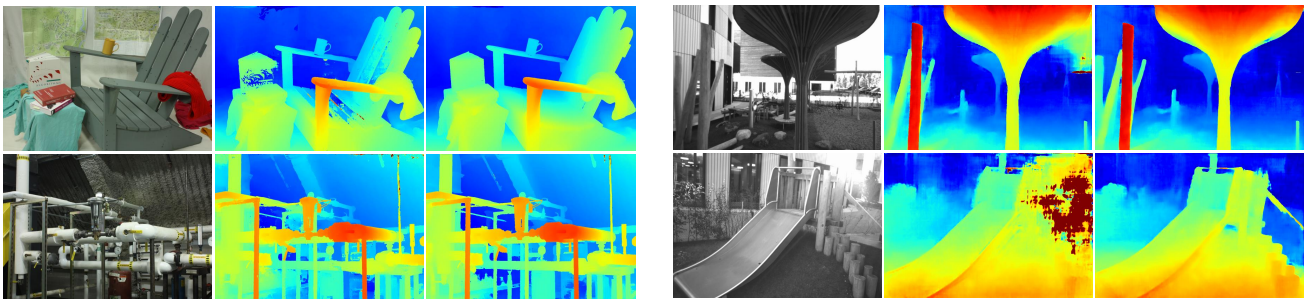
Figure 8: Domain adaptation on four datasets. (a) is on KITTI 2012. (b) is on KITTI 2015. (c) is on Middlebury. (d) is on ETH3D. The middle image of each sub-figure is the prediction from the model just trained on SceneFlow. The right image of each sub-figure is the result of our Semi-Stereo.



(a)



(b)



(c) (d)

Figure 9: Domain generalization on four datasets. (a) is on KITTI 2012. (b) is on KITTI 2015. (c) is on Middlebury. (d) is on ETH3D. The middle image of each sub-figure is the prediction from the model just trained on SceneFlow. The right image of each sub-figure is the result of our Semi-Stereo.

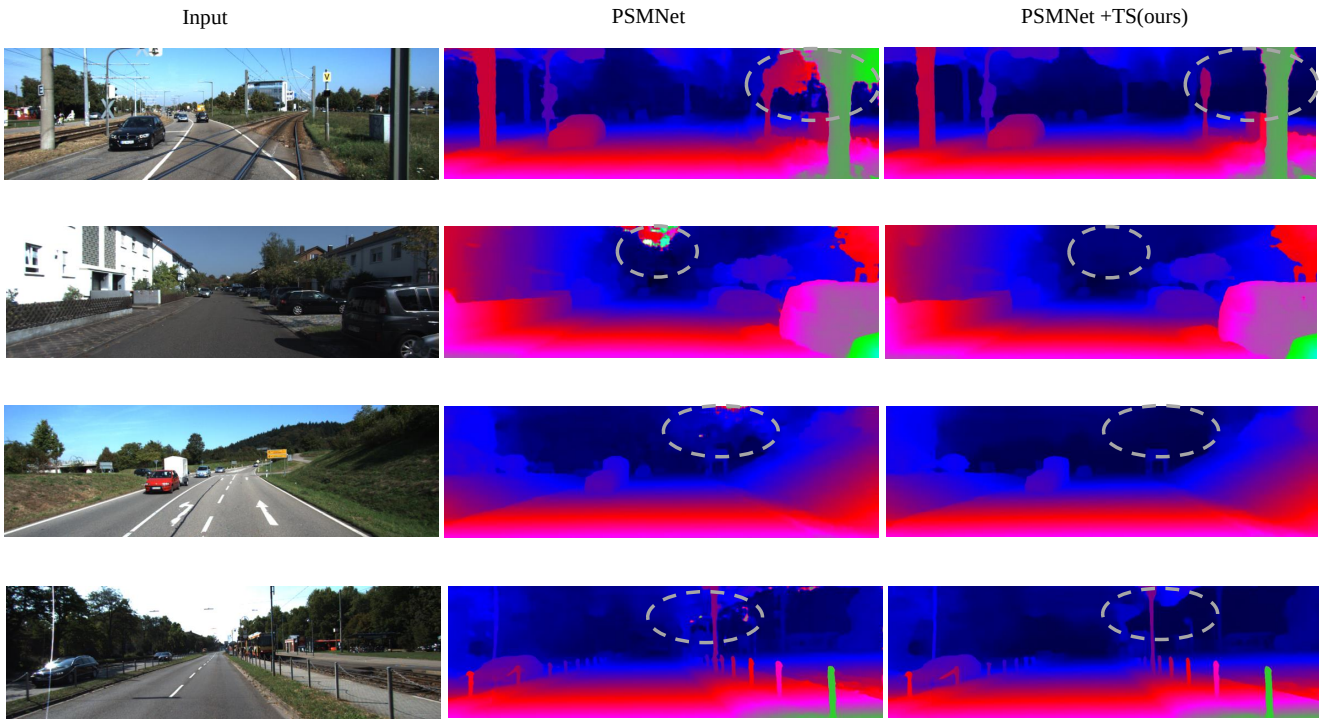



Figure 10: Disparity of sparse-annotated data. The first column is the input and the middle column is the prediction of PSMNet. We can observe large false foregrounds in textureless regions and edge flattening effect at object boundaries. The last column is the result improved by our Semi-Stereo.

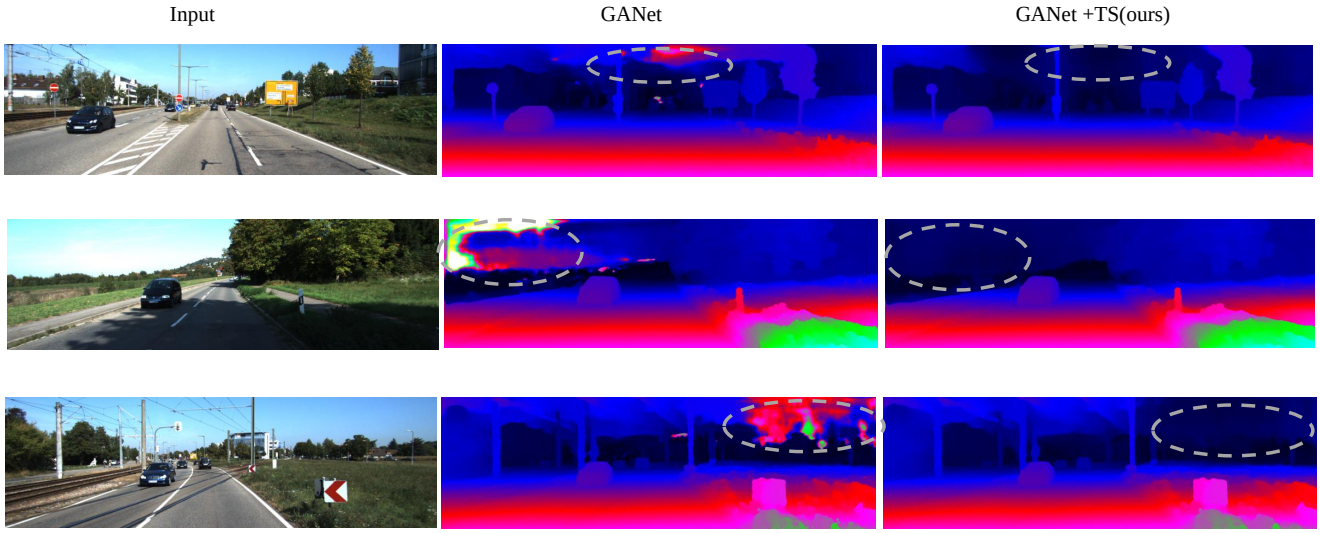



Figure 11: Disparity of sparse-annotated data. The first column is the input and the middle column is the prediction of GANet. We can observe large false foregrounds in textureless regions and edge flattening effect at object boundaries. The last column is the result improved by our Semi-Stereo.

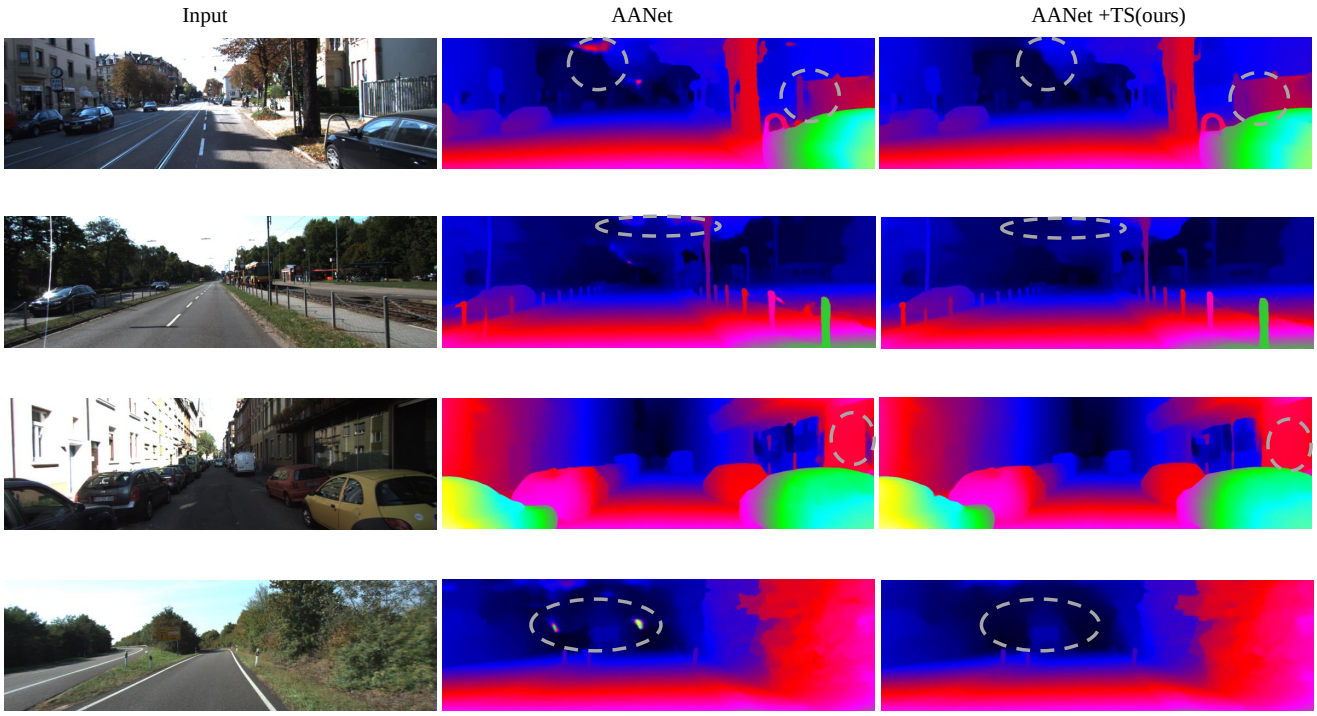



Figure 12: Disparity of sparse-annotated data. The first column is the input and the middle column is the prediction of AANet. We can observe large false foregrounds in textureless regions and edge flattening effect at object boundaries. The last column is the result improved by our Semi-Stereo.