Appendix

1. Hyper-parameter Evaluations

In this section, we will evaluate the sensitivity of hyper-parameter N' in $T_L(\cdot)$, hyper-parameters μ and β in hierarchical visual transformation module, and hyper-parameters λ_1, λ_2 and λ_3 in the final loss.

Sensitivity to N'. In $T_L(\cdot)$, we slice the original image into $N' \times N'$ non-overlapping patches. The number of patches in $T_L(\cdot)$ will increase with larger N'. We use the PSMNet backbone with global and local visual transformation module to analyze the influence of different N'. In Fig. 1, with the increase of N', the performance gradually improves until N' is equal to 5. However, larger N' will take more time during training since we transform the image patches sequentially in $T_L(\cdot)$. Therefore, we choose to set N' to 4 in this work.

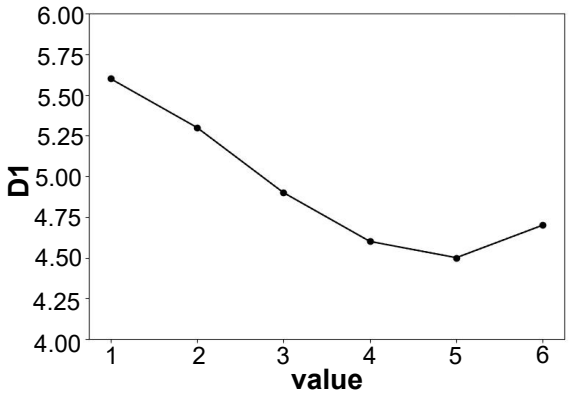

Figure 1. Evaluation on the change of N'. We use the KITTI 2012 dataset as testing dataset with D1 metric.

Sensitivity to μ and β. In hierarchical visual transformation module, we use μ and β to control the degree of global, local and pixel visual transformation. For each level, we use single visual transformation to evaluate the influence of different μ and β based on PSMNet backbone. β equals 0.15 when μ changes and μ equals 0.1 when β changes. As Fig. 2 shown, different μ and β don't lead to drastic performance change since visual transformation is a learnable process and model will adaptively generates the novel images with appropriate difficulty. Therefore, we set μ and β to 0.1 and 0.15 in this work following the best evaluation results in Fig. 2.

Sensitivity to λ_1, λ_2 and λ_3. In the final loss, λ_1 and λ_2 enforce the model to maximize the cross-domain visual discrepancy and the λ_3 enforces the model to minimize the cross-domain feature inconsistency. HVT-PSMNet model with different λ_1, λ_2 and λ_3 is used to evaluate their influence to the final performance. As Fig. 3(a) and Fig. 3(b) shown, the performance gradually improves until λ_1 equals 1 and λ_2 equals 0.5. Model will neglect the learning of domain-invariant features with larger λ_1 and λ_2. As Fig. 3(c) shown, performance nearly remains unchanged until λ_3 is equal to or greater than 0.5. Therefore, we set λ_1, λ_2 and λ_3 to 1, 0.5 and 0.5 in this work.

2. Cross-Domain Feature Consistence

We calculate the feature cosine similarity scores respectively between original feature and original, global transformed, local transformed and pixel transformed features based on the DN-PSMNet model and HVT-PSMNet model, and then plot the histograms of them from Fig. 4 to Fig. 6. The DN-PSMNet model is the baseline PSMNet model with using the domain normalization to replace original normalization in feature extraction module for a fair comparison. If the model learns the robust shortcut-invariant features (such as semantic or structural features), the cosine similarity score will be close to 1 and the frequency of larger similarity scores will be higher. We randomly select images from KITTI 2012 dataset, KITTI 2015 dataset and Middlebury dataset respectively from Fig. 4 to Fig. 6. The ETH3D dataset isn't selected since we can not generate sufficiently different global and local transformed images with only the grayscale images in this dataset.

The visualization results show that our HVT-PSMNet model significantly enhances the feature similarity to the same pixels which belong to different domains compared with DN-PSMNet model on three different transformation levels. Besides, we find transformation of pixel level usually causes worse feature consistency compared with two other levels based on HVT-PSMNet model. The reason may be pixel transformed images are more complex and diverse, so as to be more difficult to learn domain-invariant features. However, DN-PSMNet model performs better on pixel level since some pixels keep unchanged due to randomly generated Gaussian distributions in $T_P(\cdot)$. The similarity score of these pixels may be higher to lead this phenomenon.

3. More Qualitative Results

Visualization on outdoor KITTI datasets. In Fig. 7, we show the qualitative results on outdoor KITTI datasets. The SOTA generalized stereo matching method CFNet is selected as baseline backbone. As Fig. 7 shown, our HVT-CFNet model significantly outperforms the baseline CFNet model especially on the **dark textureless areas** which are framed by orange rectangles. The reason is that the shortcut features of the pixels in these areas are extremely indistinguishable and hard to find correct matching pixels. However, our HVT-CFNet model learns the shortcut-invariant features (semantic or structural features) and performs significantly better on these more difficult areas. Besides, both

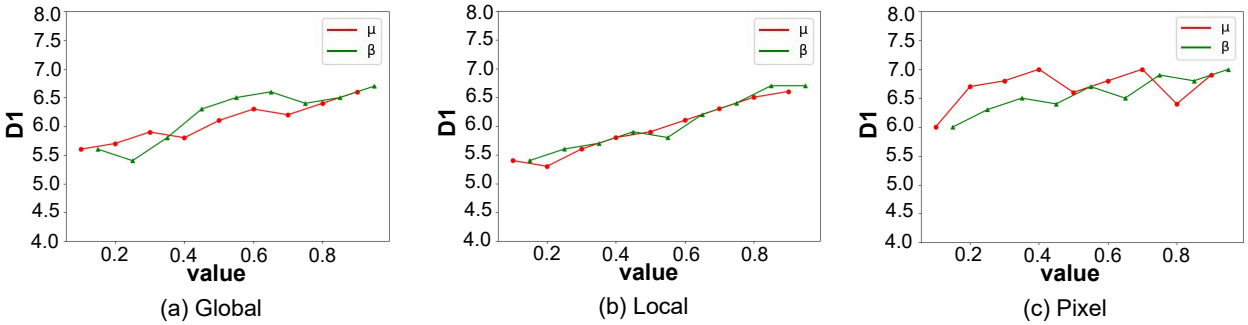

Figure 2. Evaluation on the change of μ and β for global, local and pixel visual transformation respectively. We use the KITTI 2012 dataset as testing dataset with D1 metric.

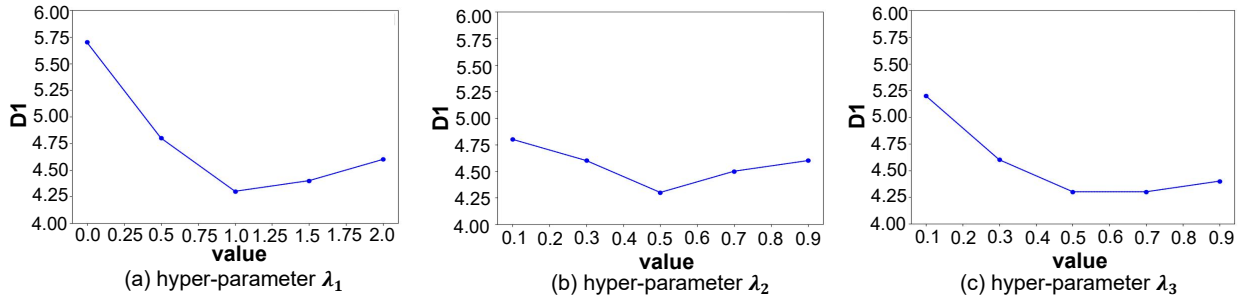

Figure 3. Evaluation on the change of λ_1, λ_2 and λ_3 in the final loss. We use the KITTI 2012 dataset as testing dataset with D1 metric.

robust models can generalize well on other easier areas such as grass with obvious texture of the first image or road surface of two other images, even they also have different data distributions with the SceneFlow training dataset.

Visualization on Middlebury and ETH3D datasets. In Fig. 8 and Fig. 9, we show the qualitative results on Middlebury and ETH3D datasets based on CFNet and HVT-CFNet models. In Fig. 8, compared with CFNet model, the HVT-CFNet model generalizes significantly better: 90.4%, 72.0% and 89.9% relative performance improvement on EPE metric to the selected three images from Middlebury dataset. Surprisingly, we also find that most of the performance improvements is in **textureless background areas**, such as black background of the first image, dark brown floor of the second image or white walls of the third image. The pixels of these area are nearly the same especially for their shortcut artifacts. The comparison demonstrates that the features of HVT-CFNet model include more shortcut-invariant components than CFNet model. Besides, in Fig. 9, we find our model also generalizes better in the blurred area of the second image from ETH3D dataset. Our HVT-CFNet model can perform significantly better for the pixels of these hard areas. Though other discriminative areas also have very different data distributions with the SceneFlow training samples, both models can generalize well on these areas.

Visualization on some failure examples. In Fig. 10, we show the qualitative results which our HVT-CFNet model performs worse than CFNet model. The areas which HVT-CFNet model performs worse are framed by the red rectangles. Surprisingly, these areas are similar and located on the windows of the car (with perspective property). The disparity perdiction of these areas with HVT-CFNet model is obviously different with other parts of the car, so as to cause large disparity error. The reason may be HVT-CFNet model has not learned the semantic features of the windows (part of the car). We can study how to learn robust semantic features of the images more explicitly in the future to solve this kind of problem.

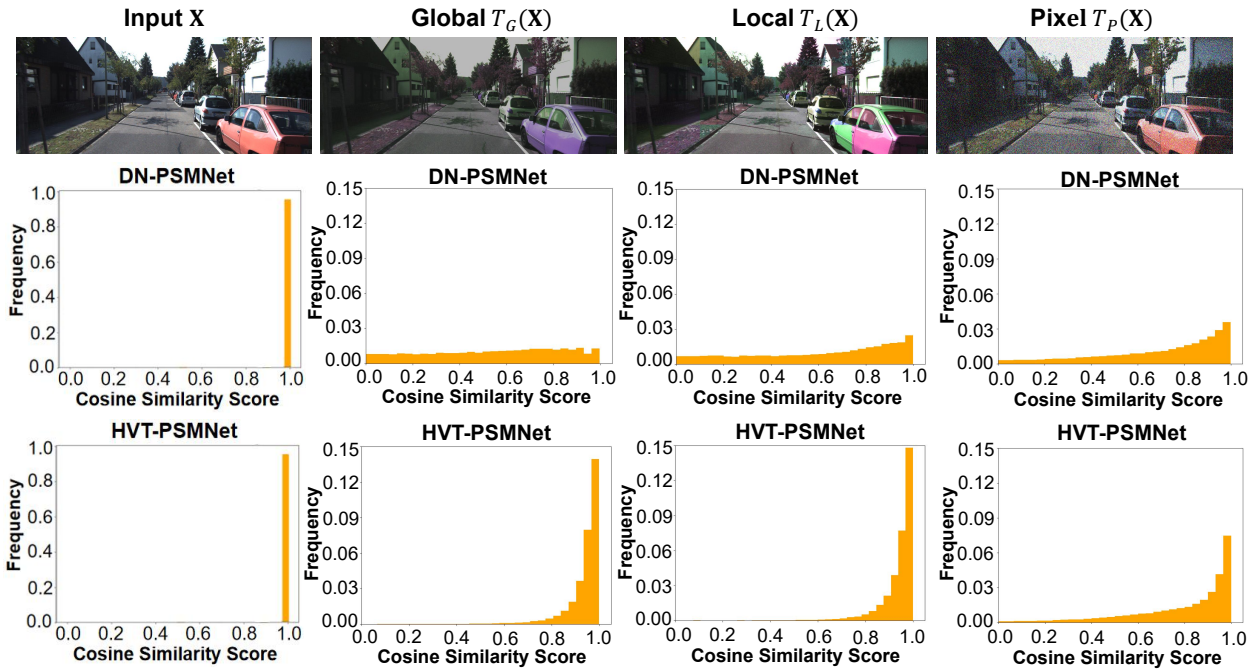

Figure 4. Histograms of feature cosine similarity scores respectively on DN-PSMNet model (see second row) and HVT-PSMNet model (see third row) between original feature and original, global transformed, local transformed and pixel transformed features. The original image is randomly selected from the KITTI 2012 dataset.

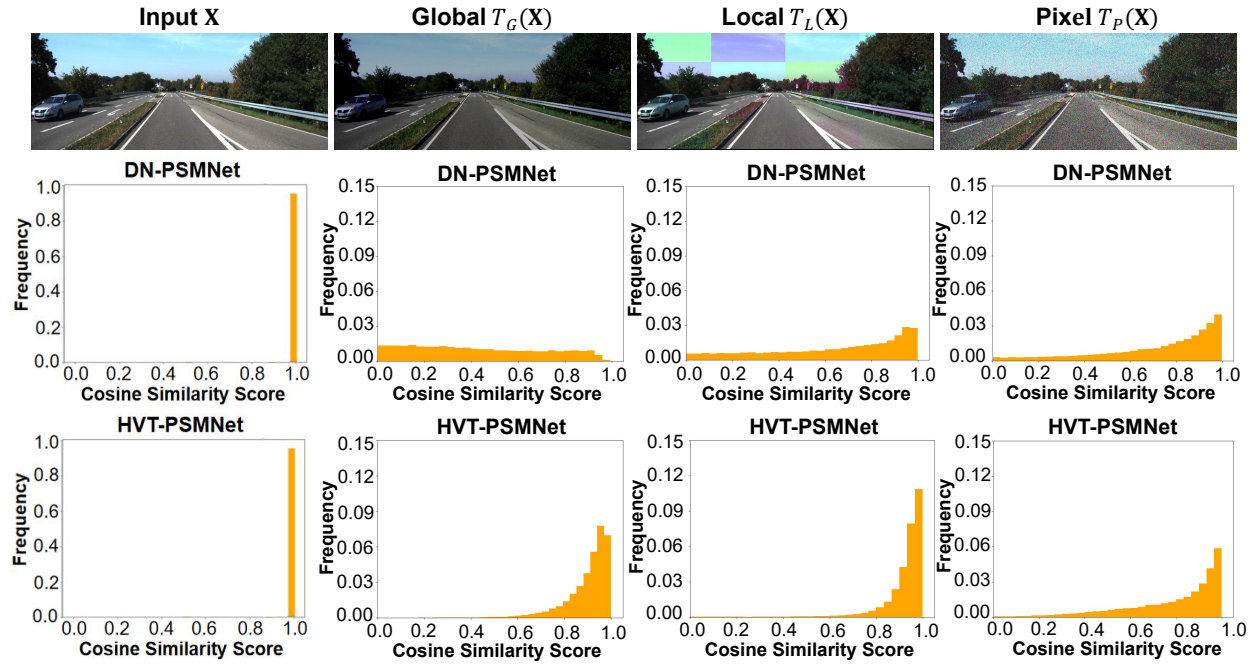

Figure 5. Histograms of feature cosine similarity scores respectively on DN-PSMNet model (see second row) and HVT-PSMNet model (see third row) between original feature and original, global transformed, local transformed and pixel transformed features. The original image is randomly selected from the KITTI 2015 dataset.

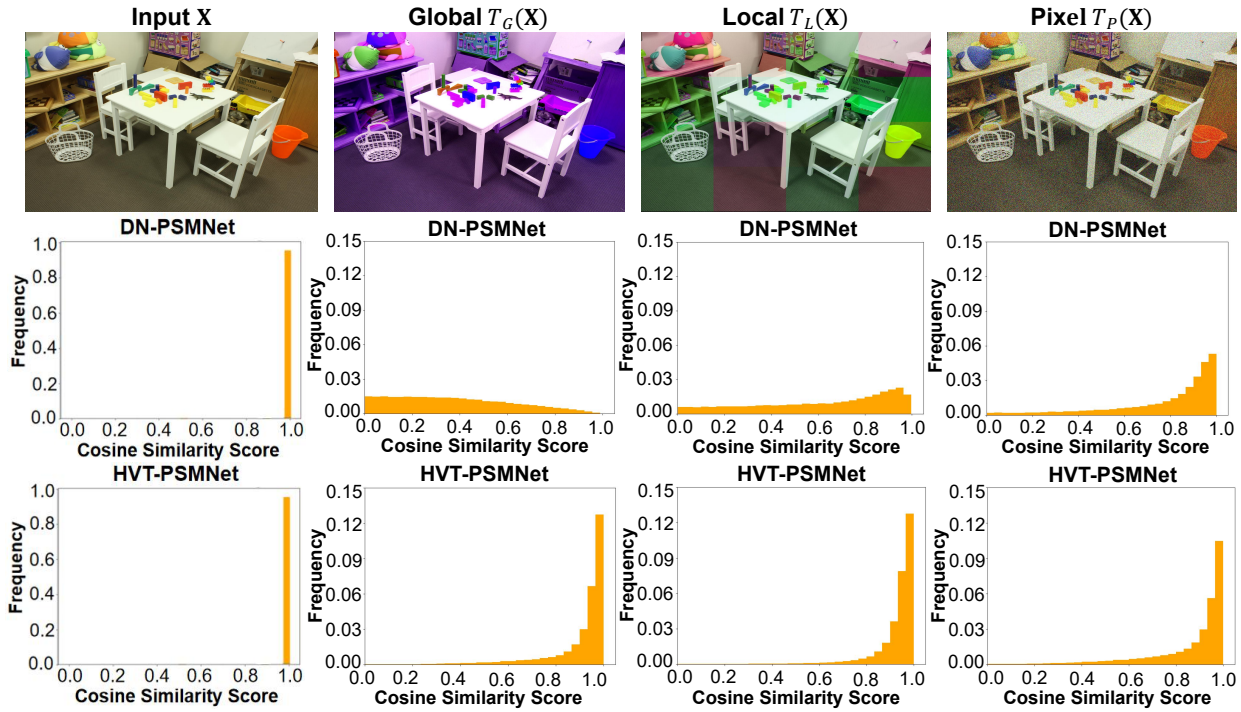

Figure 6. Histograms of feature cosine similarity scores respectively on DN-PSMNet model (see second row) and HVT-PSMNet model (see third row) between original feature and original, global transformed, local transformed and pixel transformed features. The original image is randomly selected from the realistic Middlebury dataset.

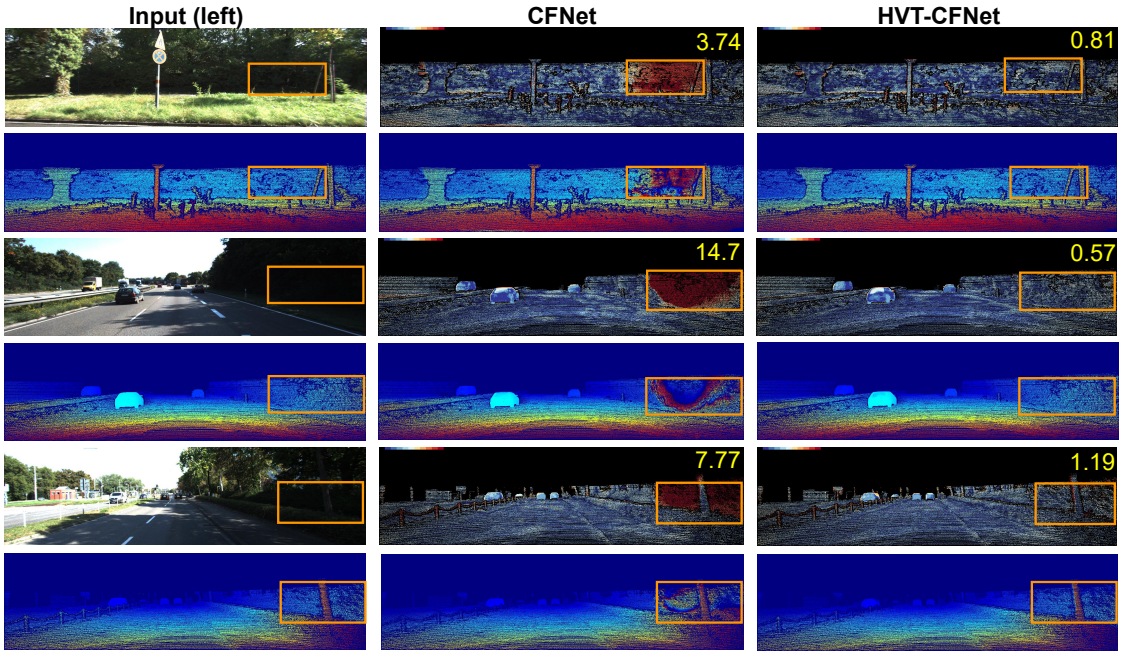

Figure 7. Qualitative results on the KITTI dataset. The first column shows the left image and the corresponding ground truth disparity map. And for each example, the first row shows the EPE error map and the second row shows the predicted colorized disparity map respectively with the CFNet and HVT-CFNet model.

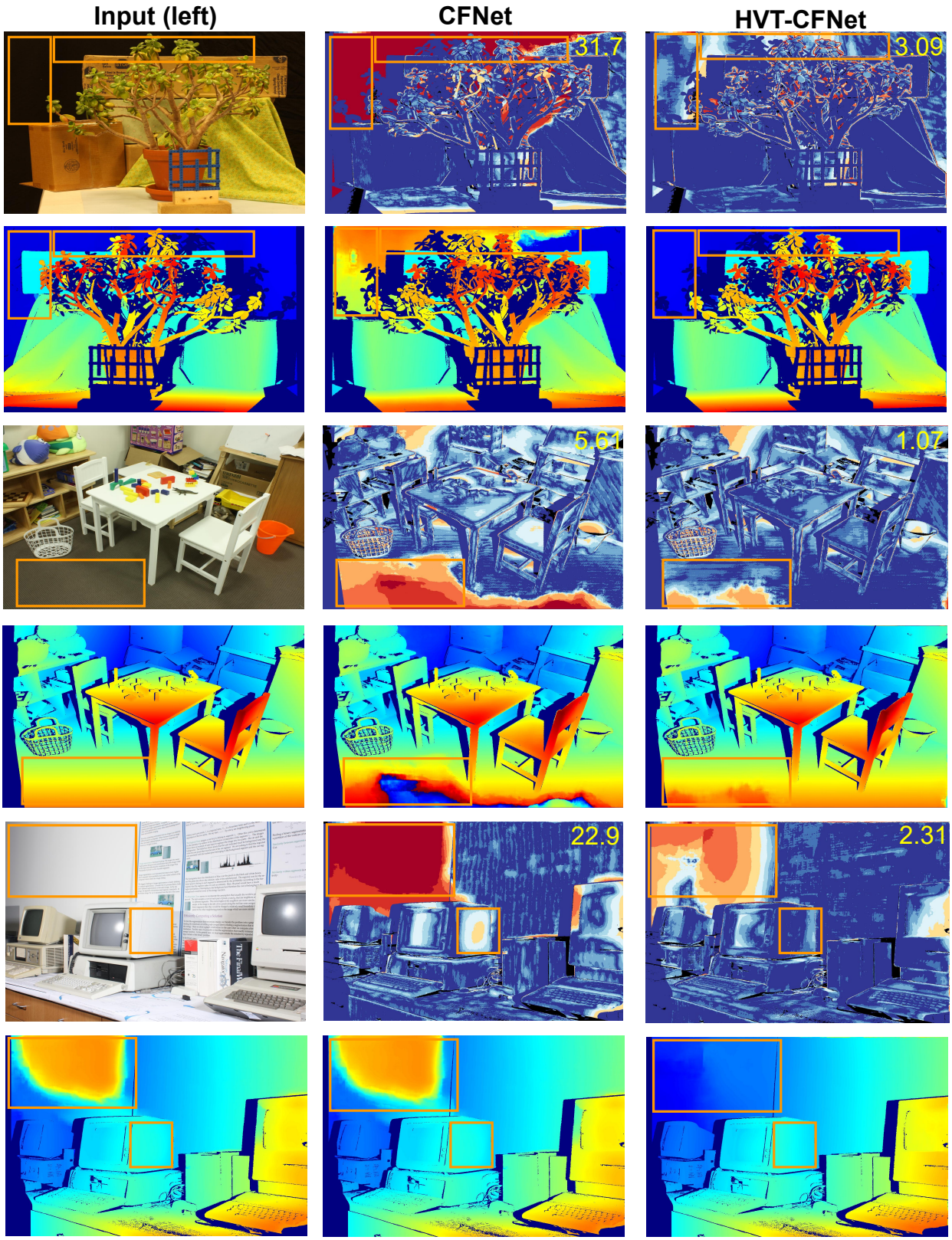

Figure 8. Qualitative results on the Middlebury training dataset. The first column shows the left image and the corresponding ground truth disparity map. And for each example, the first row shows the EPE error map and the second row shows the predicted colorized disparity map respectively with the CFNet and HVT-CFNet model.

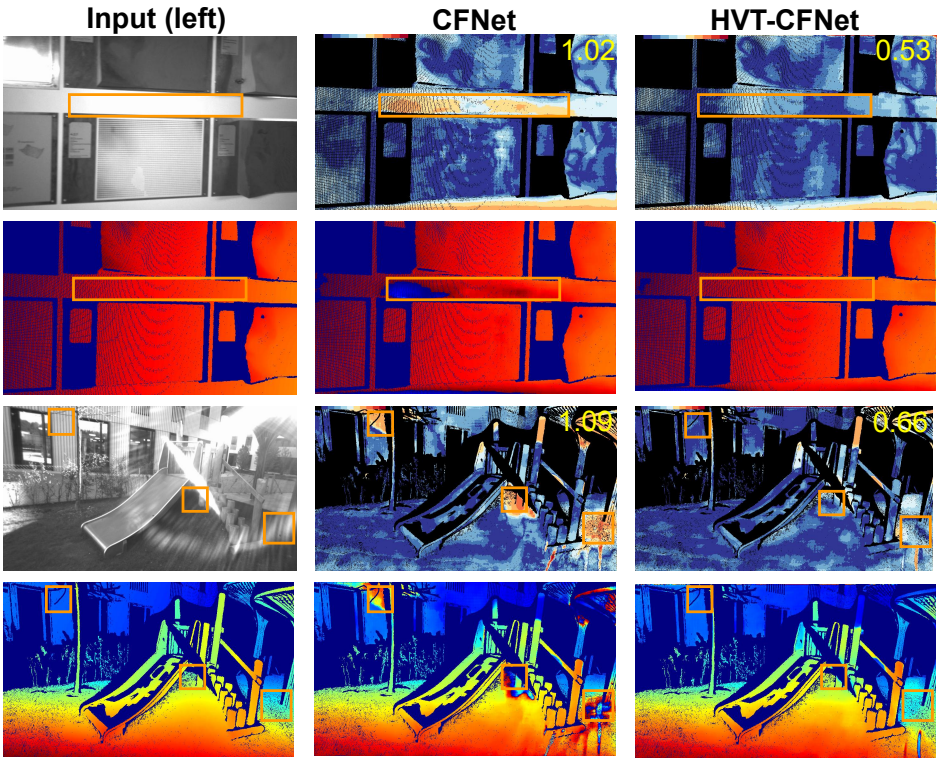

Figure 9. Qualitative results on the ETH3D training dataset. The first column shows the left image and the corresponding ground truth disparity map. And for each example, the first row shows the EPE error map and the second row shows the predicted colorized disparity map respectively with the CFNet and HVT-CFNet model.

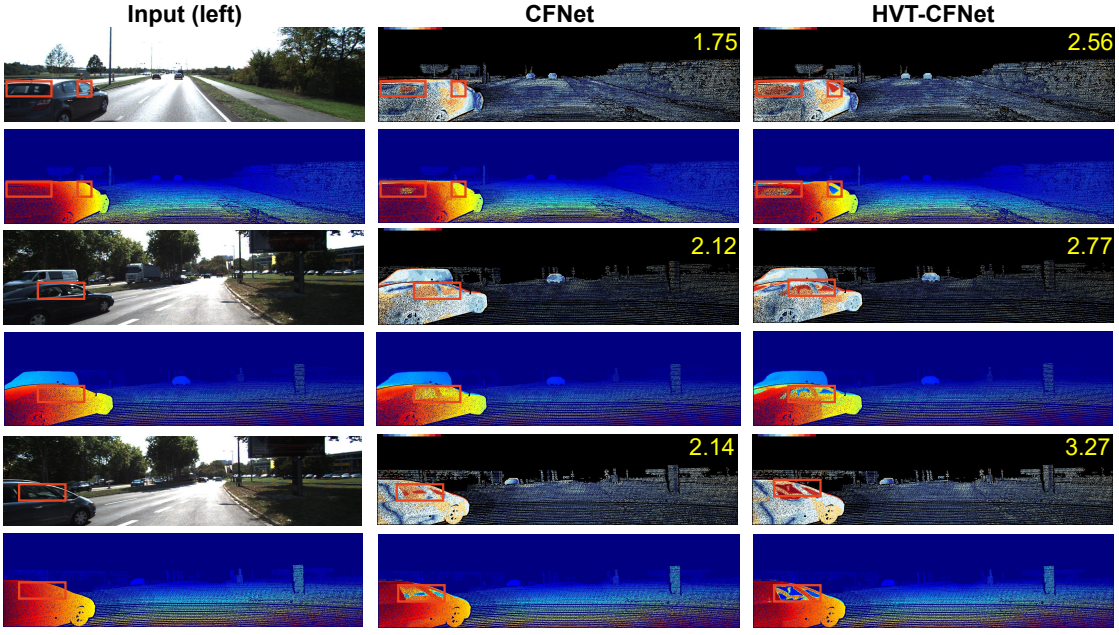

Figure 10. Qualitative results on the KITTI dataset which our model performs worse than baseline model. The first column shows the left image and the corresponding ground truth disparity map. And for each example, the first row shows the EPE error map and the second row shows the predicted colorized disparity map respectively with the CFNet and HVT-CFNet model.