S^3Net: Semantic-Aware Self-supervised Depth Estimation with Monocular Videos and Synthetic Data

Supplementary Material

Bin Cheng[1], Inderjot Singh Saggu[2], Raunak Shah[3], Gaurav Bansal[4], and Dinesh Bharadia[5]

[1]Rutgers University, NJ, USA
[2]Plus.ai, Cupertino, USA
[3]Indian Institute of Technology, Kanpur, India
[4]Blue River Technology, Sunnyvale, USA
[5]University of California, San Diego, USA

Abstract. In this supplementary material, we provide additional comparison results between our proposed S^3Net and two representatives of previous works, i.e., GeoNet [1] and T^2Net [2]. We first present additional qualitative comparison of predicted depth maps generated by the three models. It can be noticed that the depth maps generated by our model posses higher visual quality, and suffer less from semantic distortion and object misdetection. We believe this is primarily because our model combines the merits of supervised depth prediction with domain adaption and self-supervised depth prediction. Secondly, we show that due to the semantic-aware design, the synthetic-to-real image translation module in our model can largely preserve the semantic structure during the image translation. Additional, the semantic-aware design also help improve depth predictions of our model for specific semantic categories and enhance the robustness of our model when the brightness of the input images varies. Lastly, we qualitatively compare the performance of the three models using a unseen dataset and demonstrate the better generalization capability of our model.

1 Visualization Results for Depth Estimation

In this section we present additional depth estimation results (see Fig. 1) for the comparison between three models:1) GeoNet [1]; 2) T^2Net [2]; 3) our proposed S^3Net.

[1] GeoNet is a representative of approaches that predict depth using real videos.
[2] T^2Net is a representative of the approaches utilizing synthetic depth ground-truth and image translation networks for depth prediction.

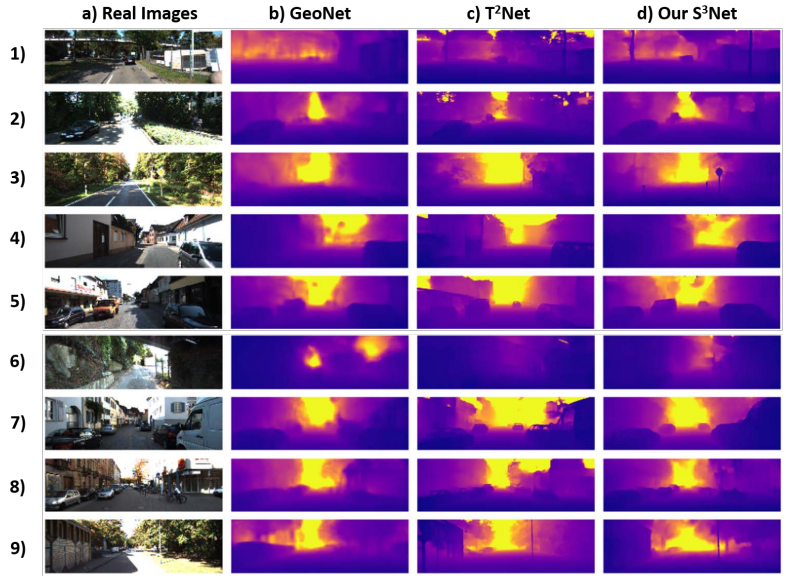

Fig. 1. Additional predicted depth maps: T^2Net produces sharper depth maps than our model, but it fails to predict accurate depth for objects with texture changes. GeoNet doesn't suffer from these drawbacks but produces results which are blurred. Our approach generates sharper depth maps than GeoNet while being more accurate than both GeoNet and T^2Net in its predictions.

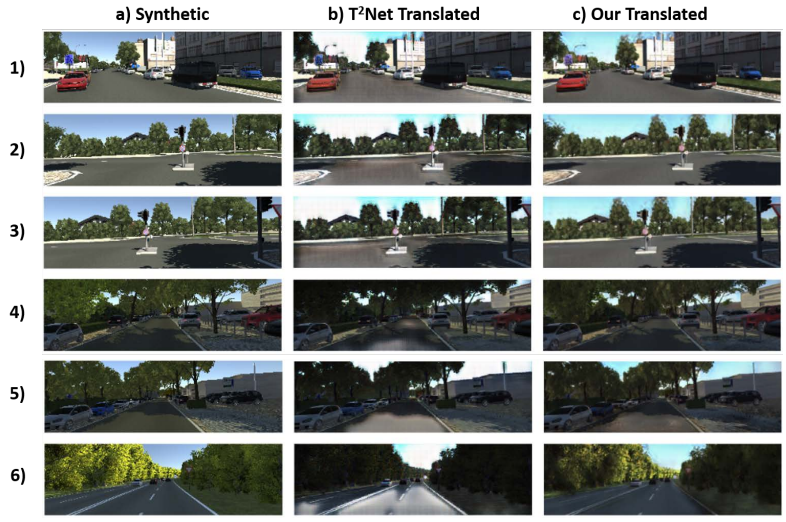

Fig. 2. Normal to bright light conditions : Native T^2Net produces an artifact ridden translated image. Our semantic consistent translator significantly reduces the artifacts and retains the semantic structure of the image.

Column b) in Fig. 1 shows the predicted depth maps for GeoNet which is trained on real-world videos only in self-supervised fashion. GeoNet's results are not very sensitive to edge transitions and thus look blurry overall. T^2Net, on the other hand, can produce visually sharper depth maps but it fails to accurately predict depth for objects with changing texture, e.g., in image 1.c) different depth values are predicted for part of the same object over and under the bridge. This can also be observed in image 7.c) (house with window on the left) and image 9.c) (textured wall on the left) where drastic changes in depth are predicted for the same object due to change in texture. We also observe that T^2Net has troubles in distinguishing pedestrians and road signs from the background (image 2 and 3) whereas our S^3Net generates much better predictions. We also observe that for white shiny surfaces like in image 5 (building on the left) and in image 7 (building on the right) T^2net predicts very high depth value. Looking at the translation results for T^2Net in Fig. 2 we can observe that the translated sky looks very very similar to the white shiny surfaces which can possibly explain the erroneous predictions. Our predictions combine the best of both worlds, they are much sharper than GeoNet and don't have any of the issues discussed for T^2Net. However our model seems to have trouble in predicting depth for sky.

2 Semantic Consistent Synthetic-to-Real Image Translation

We present a qualitative comparison between native T^2Net synthetic-to-real translator network versus our semantic consistent network in Fig. 2 and Fig. 3.

3 Category-Specific Depth Estimation

Fig. 4 presents the depth prediction accuracy in terms of squared relative error for different semantic categories. As shown in the figure, our model outperforms GeoNet and T^2Net in all five selected categories. The improves are 34%, 35%, 37%, 27% and 13% for categories *car*, *road*, *sidewalk*, *person* and *motorcycle*, respectively. It can be noticed that the improvements for *person* and *motorcycle* are comparatively less, since these are overall less frequent categories in both synthetic and real data. We believe these improvements are primarily due to our model is semantic-aware and thus the semantic information is better preserved in depth prediction.

4 Depth Estimation Under Different Brightness Conditions

We also compare the performance of different models for input images under various brightness conditions. Fig 5 shows that our model consistently outperforms GeoNet and T^2Net(the larger the brightness factor is, the brighter the image is). We believe this is also because our model takes semantic information

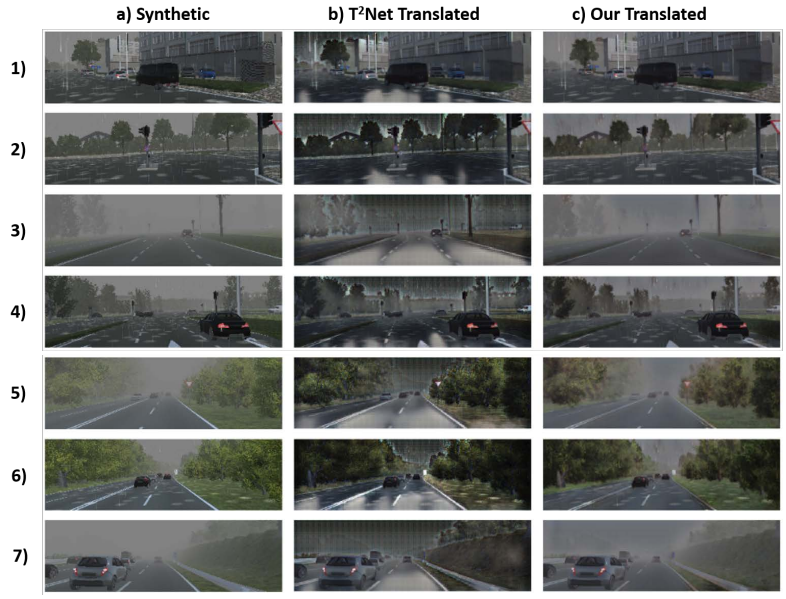

Fig. 3. Low-light, fog and rainy conditions: T^2Net produces unsatisfactory translation results for synthetic images in low-light, foggy and/or rainy conditions. It produces checkered artifacts in the sky and severely distorts the road. Our additional semantic constraint allows translator network to perform well in these varying conditions.

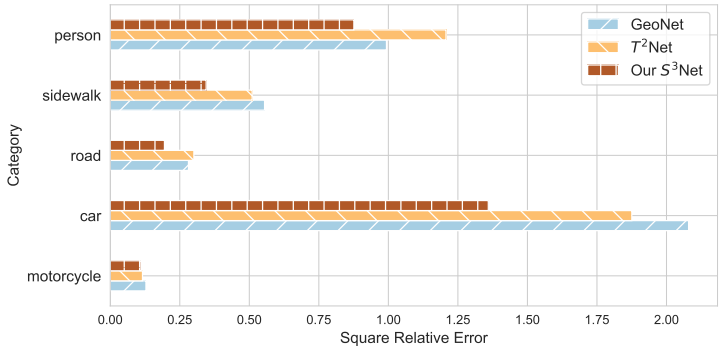

Fig. 4. Squared relative error of predicted depths on KITTI dataset for various semantic categories

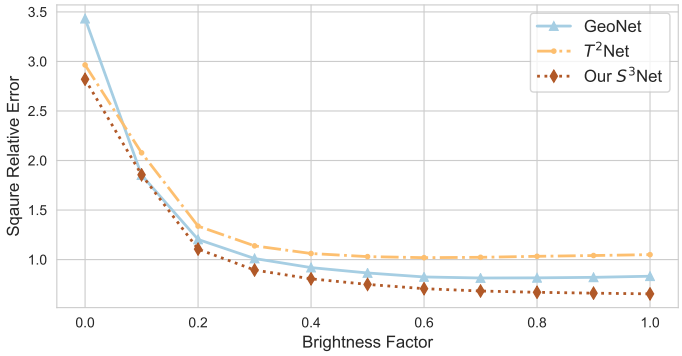

Fig. 5. Squared relative error of predicted depths on KITTI dataset under various brightness conditions

into consideration. The semantic information is robust under different brightness conditions, which can further enhance the robustness of our model.

5 Qualitative Generalization on Make3D

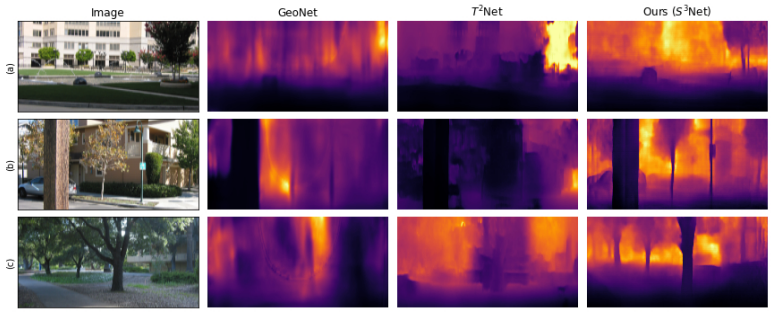

Fig. 6. Examples of our S^3Net's generalization when as compared to a standard unsupervised model (GeoNet) and a standard syn to real model (T^2Net). Images taken from the Make3D test dataset.

We present an example of our model's generalization capability through a qualitative comparison on the Make3D dataset, since this dataset is unseen and the images in this dataset are significantly different from the usual KITTI/vKITTI images. As shown in Figure 6, our model can combine the merits of both supervised depth estimation with domain adaptation and self-supervised depth estimation. For example for input image (b), GeoNet partially succeeds in predicting the depths of the house and poles, whereas T2Net manages to understand variation of depths in intricate locations. Our model outperforms these two models in terms of both predicting visually more accurate depths of objects and clearly reflecting the depth changes in intricate locations.

References

1. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1983–1992 (2018) 1
2. Zheng, C., Cham, T.J., Cai, J.: T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 767–783 (2018) 1