Panoptic NeRF: 3D-to-2D Label Transfer for Panoptic Urban Scene Segmentation

Xiao Fu[1*] Shangzhan Zhang[1*] Tianrun Chen[1] Yichong Lu[1] Lanyun Zhu[2]
Xiaowei Zhou[1] Andreas Geiger[3] Yiyi Liao[1†]
[1]Zhejiang University [2]Singapore University of Technology and Design
[3]University of Tübingen and MPI-IS, Tübingen

Abstract

In this supplementary document, we first give a detailed overview of our network architecture, sampling strategy, evaluation metrics, and training and inference procedure in Section 1. Next, we describe our data preparation process in Section 2, including the stereo depth maps for supervision, and depth maps and semantic labels for evaluation. Finally, we provide additional experiments in Section 3. The supplementary video shows our label transfer results on consecutive video frames as well as novel view label renderings.

1. Implementation Details

1.1. Network Architecture

Fig. 1 shows the trainable part of our Panoptic NeRF model. We adopt the same network architecture in all experiments. The network takes as input the 3D location $\mathbf{x}$ (each element normalized to $[-1, 1]$) and the viewing direction $\mathbf{d}$. Following NeRF [5], both $\mathbf{x}$ and $\mathbf{d}$ are mapped to a higher dimensional space using a positional encoding (PE):

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \cdots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)) \tag{1}$$

To learn high frequency components in unbounded outdoor environments, we set $L = 15$ for $\gamma(\mathbf{x})$ and $L = 4$ for $\gamma(\mathbf{d})$. Our learned semantic field is conditioned only on the 3D location $\mathbf{x}$ rather than the viewing direction $\mathbf{d}$ in order to predict view-independent semantic distributions.

1.2. Sampling Strategy

We sample points within the bounding primitives to skip empty space. As our bounding primitives are convex[1], each ray intersects a bounding primitive exactly twice which determines the sampling interval. For each camera ray, we sort all bounding primitives that the ray hits from near to far and save the intersections offline. To save storage and to speed up training, we keep the first 10 sorted bounding primitives as the rest are highly likely to be occluded. If a camera ray intersects less than 10 bounding primitives, we additionally sample a set of points to model the sky in $[t_{max}, t_{max} + t_{int}]$, where t_{max} denotes the distance from the origin to the furthest bounding primitive in the scene and t_{int} is a constant distance interval.

1.3. Evaluation Metric

We evaluate mIoU and pixel accuracy following standard practice [1, 4]. Here, we provide more details of the multi-view consistency and panoptic quality metrics.

Multi-view Consistency: To evaluate multi-view consistency, we use depth maps obtained from LiDAR points to retrieve matching pixels across two consecutive frames. A similar multi-view consistency metric is considered in [6] where optical flow is used to find corresponding pixel pairs. We instead use LiDAR depth maps as they are more accurate compared to optical flow estimations. The details of generating the LiDAR depth maps will be introduced in Section 2.2. Given LiDAR

[1] The cuboids and ellipsoid are both convex. The extruded 3D plane is convex in a local region.

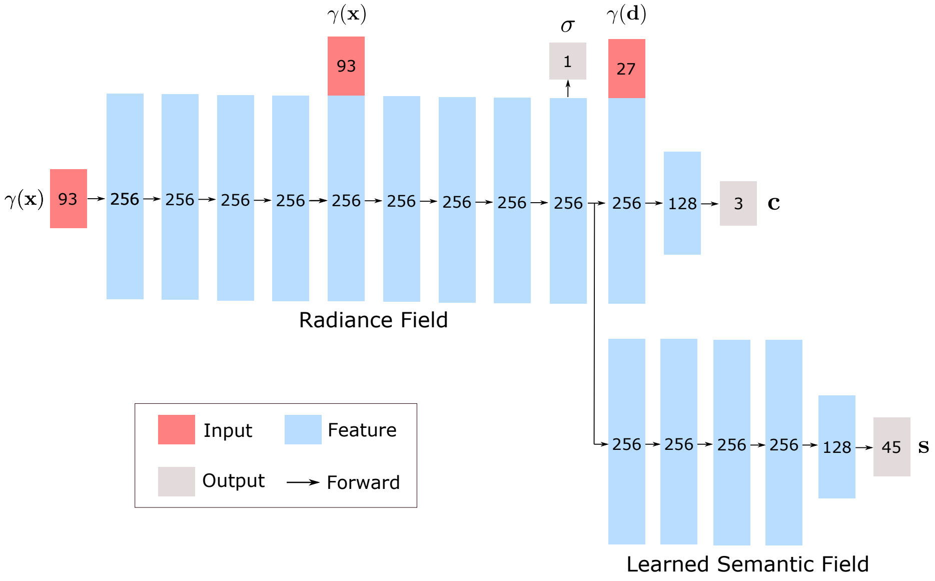

Figure 1: **Trainable Part of Panoptic NeRF.** For the radiance field, we follow the original implementation of NeRF except for setting $L = 15$ for $\gamma(\mathbf{x})$. The learned semantic field predicts semantic logits independent of the viewing direction. The logits are then transformed into categorical distributions through a softmax layer.

depth maps at two consecutive test frames, we first unproject them into 3D space and find matching points. Two LiDAR points are considered matched if their distance in 3D is smaller than 0.1 meters. For each pair of matched points, we retrieve the corresponding 2D semantic labels and evaluate their consistency. The MC metric is evaluated as the number of consistent pairs over all matched pairs. Despite being not 100% accurate as the 3D points may not match exactly in 3D space, we find this metric meaningful in reflecting multi-view consistency.

Panoptic Quality: Following [3], we use the PQ metric to evaluate the performance of panoptic segmentation. As the ground truth panoptic labels are not precise in distant areas and have a lot of small noises of things, we set ground truth labels of areas less than 100 pixels to "void". Correspondingly, segment matching will not be performed in void regions. In addition, Panoptic maps of the 3D-2D CRF and our method are obtained by 3D primitives, thus containing very far objects. In fact, these far objects may only occupy very small areas, usually less than 100 pixels, on 2d images. To avoid being biased by those extremely far objects in the segment matching, we omit them by setting the predicted labels of the areas less than 100 pixels to the "sky" class. To ensure a fair comparison across all methods, we adopt the same evaluation protocol for all baselines and our method.

1.4. Training and Inference

As mentioned in Section 4.2 of the main paper, our total loss function comprises five terms, including three semantic losses $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$, $\mathcal{L}_{\mathbf{S}}^{2D}$, $\mathcal{L}_{\mathbf{s}}^{3D}$, the photometric loss $\mathcal{L}_{\mathbf{c}}$ and the depth loss $\mathcal{L}_{d}$. During per-scene optimization, the photometric loss $\mathcal{L}_{\mathbf{c}}$ is defined on the posed stereo images. The 2D semantic losses $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$, $\mathcal{L}_{\mathbf{S}}^{2D}$ are applied to the left images only. While our method allows for using noisy 2D semantic predictions on the right images, this ensures fair comparison to the 3D-2D CRF which is not capable of using predictions on other viewpoints for inference. We apply the 3D semantic loss $\mathcal{L}_{\mathbf{s}}^{3D}$ directly on 3D points sampled along the camera rays of the left images. The depth loss $\mathcal{L}_{d}$ is also defined on the left images as the information gain is marginal on the right views.

For inference, we compare our method to the baselines on the left views of which the manually labeled 2D Ground Truth is defined. Note that our method is not constrained to the left views during inference. We show label transfer results on the right camera views in Section 3.3 and novel view label synthesis in Section 3.4.

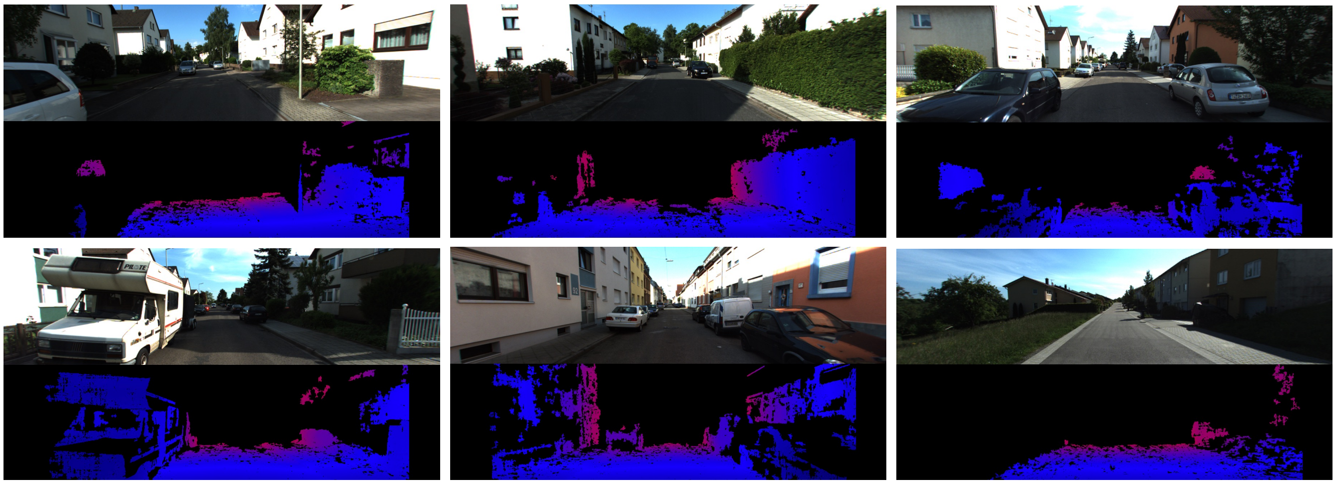

Figure 2: **Depth Maps for Weak Depth Supervision.** Each group shows the RGB image (top) and the corresponding depth maps (bottom) used for supervision.

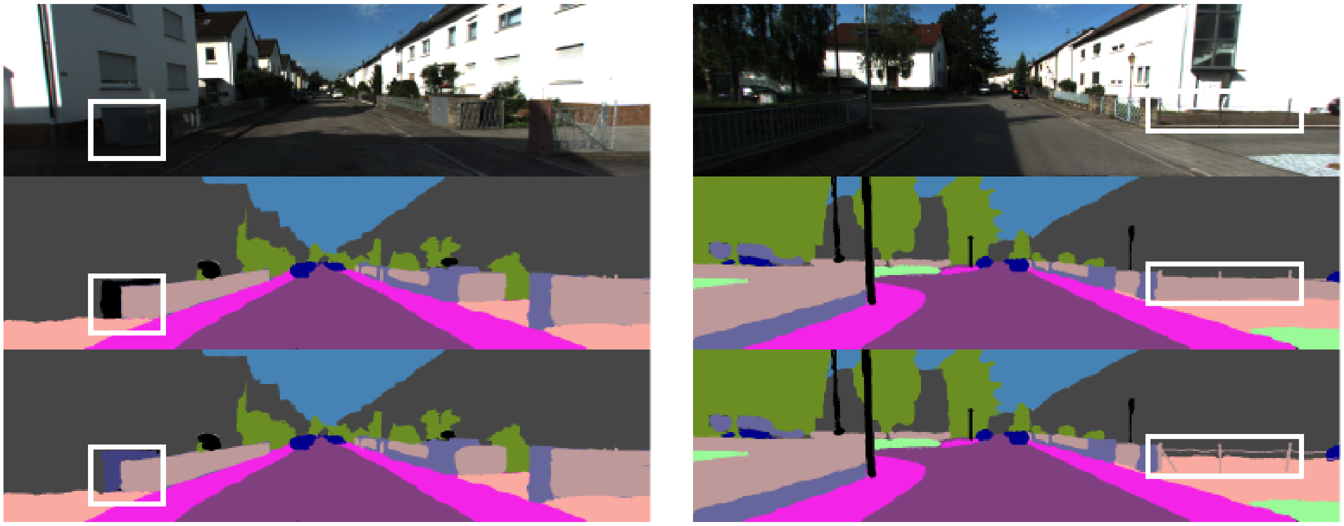

Figure 3: **Examples of Modified Ground Truth.** We correct some GT pixels that were incorrectly labeled in the KITTI-360 dataset. Top: Input RGB images. Middle: Original ground truth. Bottom: Modified ground truth. In the first column, we add the "box" class. In the second column, we correct the "parking" area.

2. Data Preparation

2.1. Stereo Depth for Weak Depth Supervision

To provide weak depth supervision to Panoptic NeRF, we use Semi-Global Matching (SGM) [2] to estimate depth given a stereo image pair. We perform a left-right consistency check and a multi-frame consistency check in a window of 5 consecutive frames to filter inconsistent predictions. We further omit depth predictions further than 15 meters for each frame as disparity is better estimated in nearby regions, see Fig. 2.

2.2. LiDAR Depth for Evaluation

We evaluate the rendered depth maps against the LiDAR measurements. We refrain from using LiDAR as input as 1) this allows us to evaluate our depth prediction against LiDAR and 2) it makes our method more flexible to work with settings without any LiDAR observations. As LiDAR observations at each frame are sparse, we accumulate multiple frames of

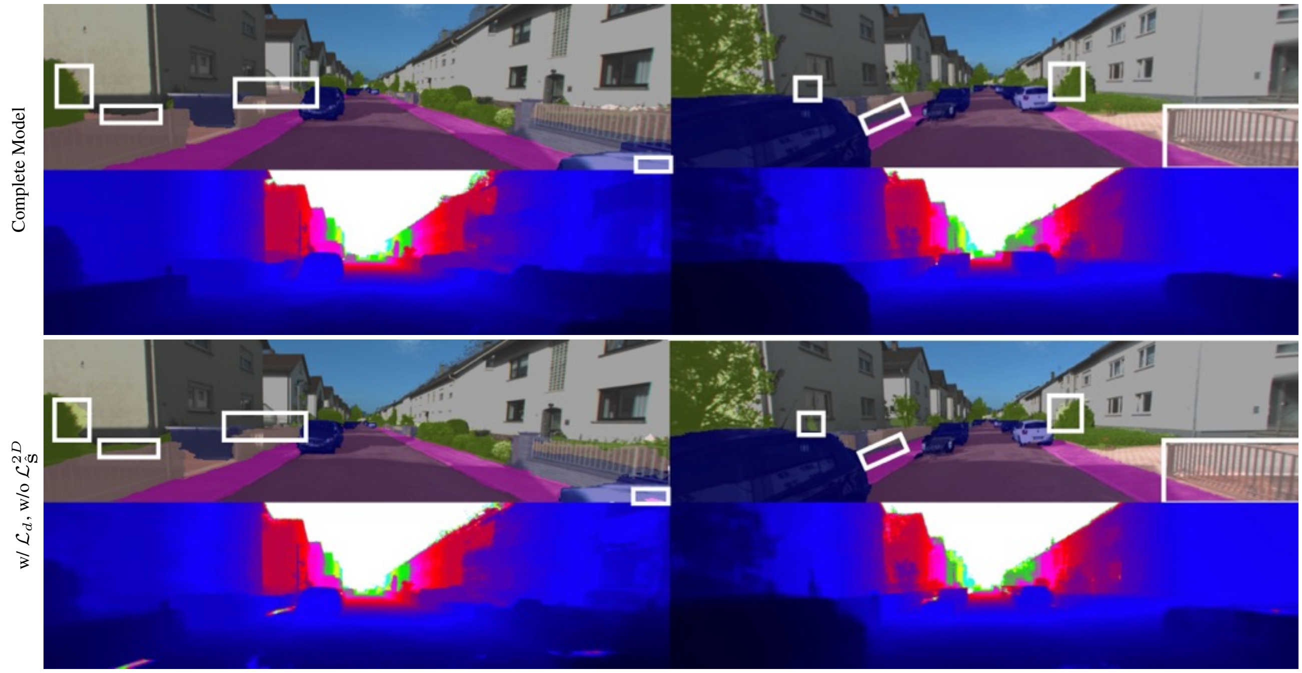
Complete Model
w/ $\mathcal{L}_d$, w/o $\mathcal{L}^{2D}_{\hat{\mathbf{S}}}$

Figure 4: **Qualitative Comparison of Ablation Study.** We visualize the semantic map and depth map of the complete model (top) and the model without fixed semantic field (bottom).

LiDAR observations and project the visible points to each frame similar to [7].

2.3. Manually Annotated 2D GT

The manually annotated 2D ground truth of KITTI-360 [4] is inferior at some regions. For a fair comparison, we improve the label quality by manually relabeling ambiguous classes, see Fig. 3 for illustrations.

3. Additional Experimental Results

3.1. Weak Depth Supervision

We show that using the depth loss $\mathcal{L}_d$ alone is not able to recover accurate object boundaries in Fig. 4. In contrast, adding the semantic loss $\mathcal{L}^{2D}_{\hat{\mathbf{S}}}$ to the fixed semantic field further improves the object boundary. These improvements can be explained as follows: Firstly, the weak stereo depth supervision is not fully accurate, especially at far regions. Furthermore, even with perfect depth supervision, the model receives very small penalty if the predicted depth is close to the GT depth. In contrast, the cross entropy loss $\mathcal{L}^{2D}_{\hat{\mathbf{S}}}$ defined on the fixed semantic field provides a strong penalty as small errors in depth lead to wrong semantics.

3.2. Qualitative Comparison of Label Transfer

Fig. 5 shows additional qualitative comparisons corresponding to the Table 1 of the main paper. Consistent with the quantitative results, our method outperforms all baselines qualitatively. We further show qualitative comparisons to 3D-2D CRF on a set of unlabeled 2D frames, including semantic label transfer in Fig. 6 and panoptic label transfer in Fig. 7.

3.3. Stereo Label Transfer

In Fig. 8, we illustrate our stereo label transfer results. Despite that we only utilize pseudo ground truth on the left views for supervision, our model achieves consistent results on both left and right views.

3.4. Novel View Label Synthesis

Here, we evaluate our performance on novel view label synthesis by applying the photometric loss $\mathcal{L}_{\mathbf{c}}$ to the left images only. This allows us to evaluate novel view appearance and label synthesis on the right view images. As shown in Fig. 9, our method achieves promising results on novel view appearance and label synthesis. More results of appearance and label synthesis on unseen viewpoints can be found in the supplementary video.

3.5. Analysis of 3D-2D CRF

The 3D-2D CRF performs inference based on a multi-field CRF which reasons jointly about the labels of the 3D points and all pixels in the image. To obtain dense 3D points, it accumulates LiDAR observations over multiple frames and project visible 3D points to the image based on a reconstructed mesh. Fig. 10 shows depth maps of the reconstructed mesh corresponding to Fig. 5 of the main paper. As can be seen, the side of the building can hardly be scanned by the LiDAR, leading to incomplete mesh reconstruction. Consequently, 3D-2D CRF lacks 3D information in these regions and needs to distinguish building instances mainly based on 2D image cues. It is not surprising that the 3D-2D CRF fails at overexposed image regions in this case.

3.6. Failure Cases

Our method leverages a deterministic instance field defined by the 3D bounding primitives to render instance labels. Thus, our method struggles to recover accurate instance boundaries where two instance bounding primitives overlap in 3D space. This sometimes occurs on the building class where two buildings are spatially connected to each other as shown in Fig. 11.

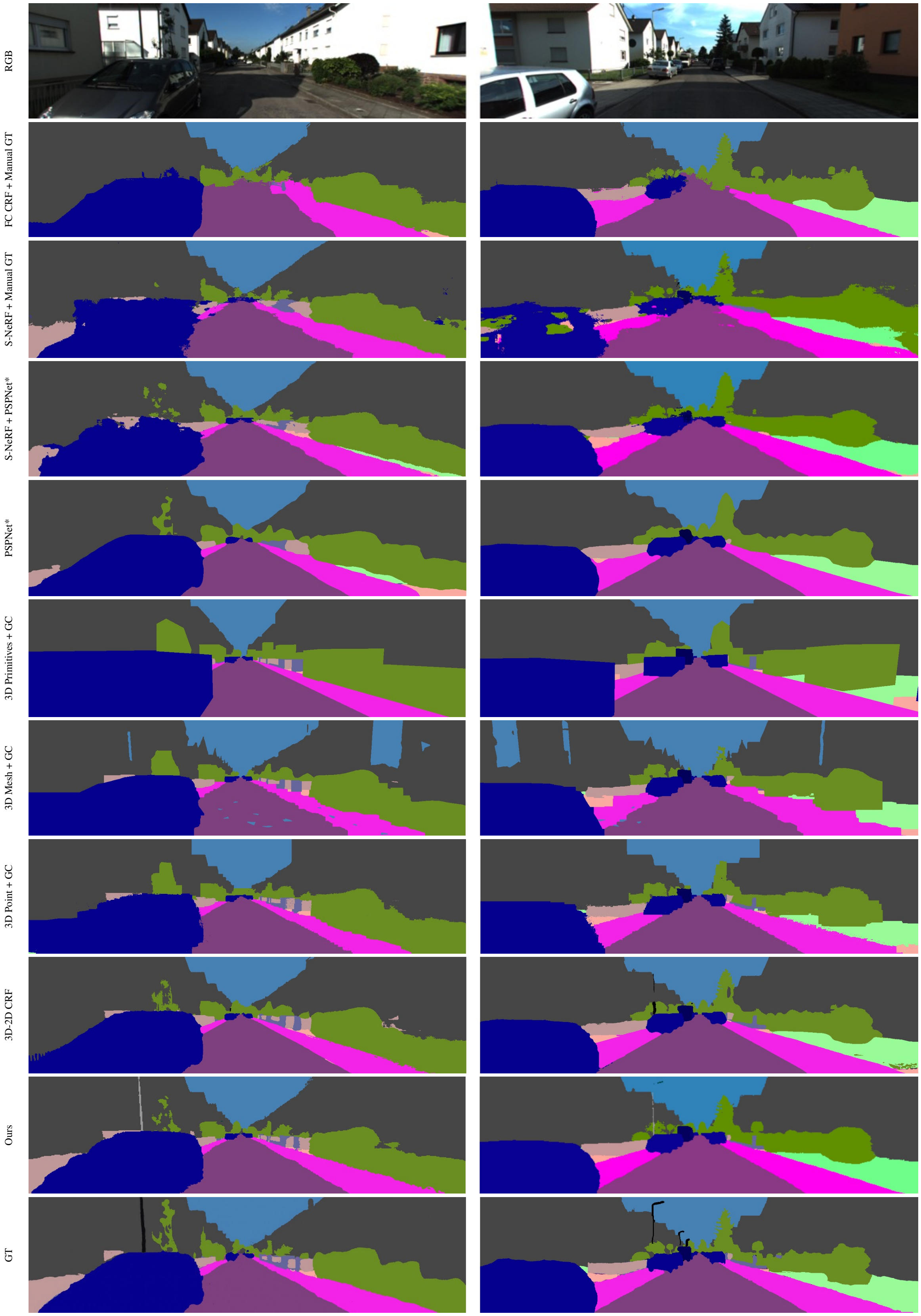

Figure 5: **Qualitative Comparison of Semantic Label Transfer** on frames with manually labeled ground truth.

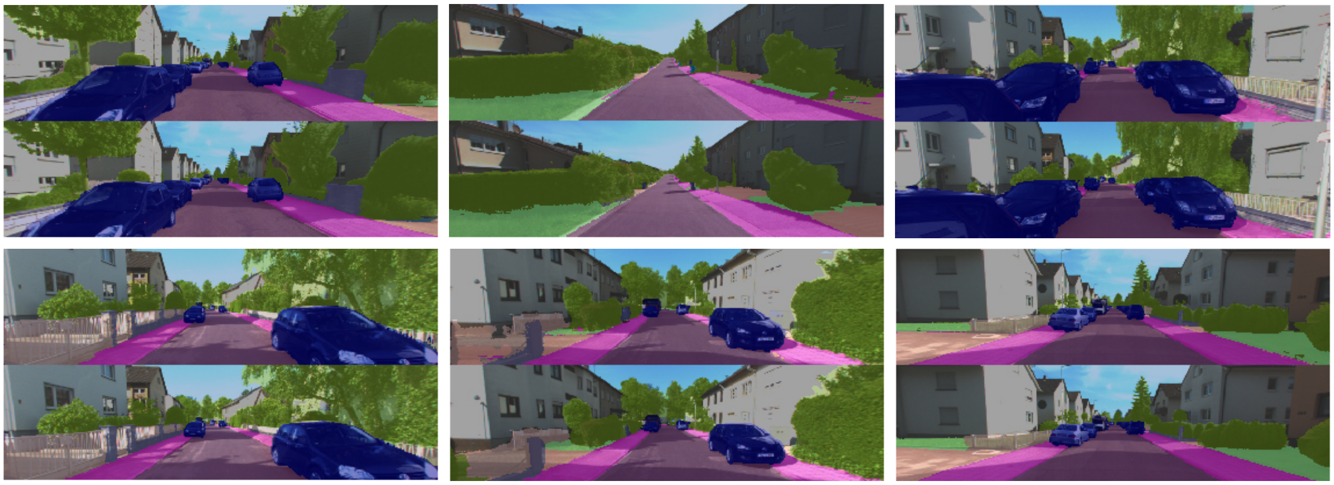

Figure 6: **Qualitative Comparison of Semantic Label Transfer** on frames without manually labeled ground truth. Each group shows the prediction of 3D-2D CRF [4] (top) and ours (bottom).

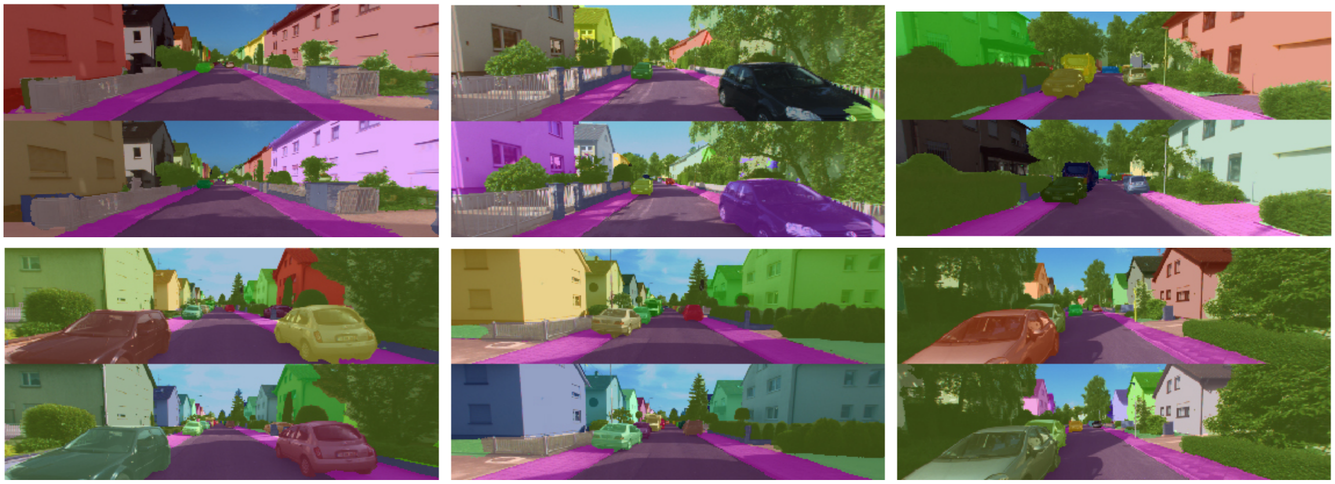

Figure 7: **Qualitative Comparison of Panoptic Label Transfer** on frames without manually labeled ground truth. Each group shows the prediction of 3D-2D CRF [4] (top) and ours (bottom). The colors of the instances do not match due to missing 2D ground truth.

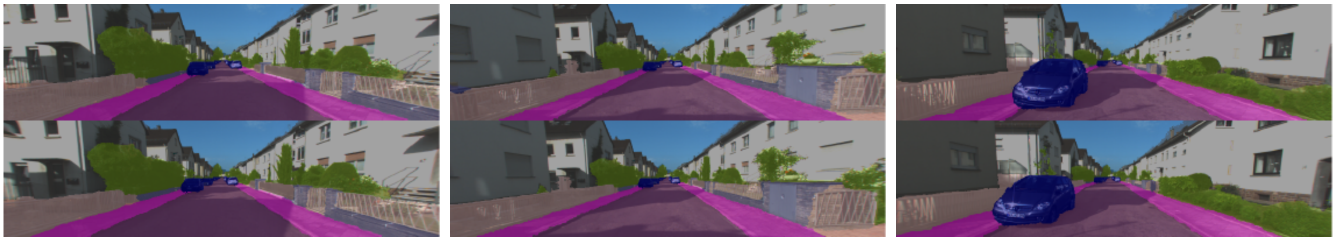

Figure 8: **Qualitative Results for Stereo Label Transfer.** Top: Blended semantic results of left views. Bottom: Blended semantic results of right views.

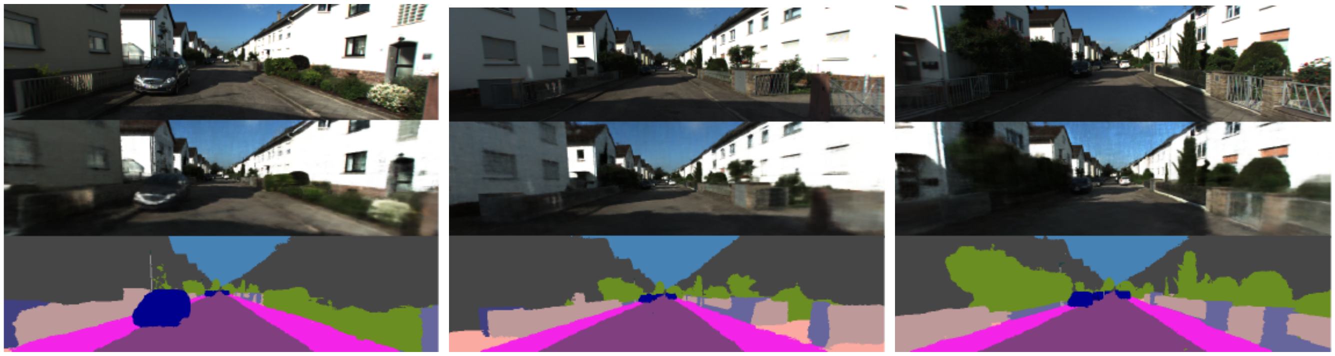

Figure 9: **Qualitative Results for Novel View Label Synthesis.** Top: GT RGB images. Middle: Rendered RGB images. Bottom: Rendered semantic maps.

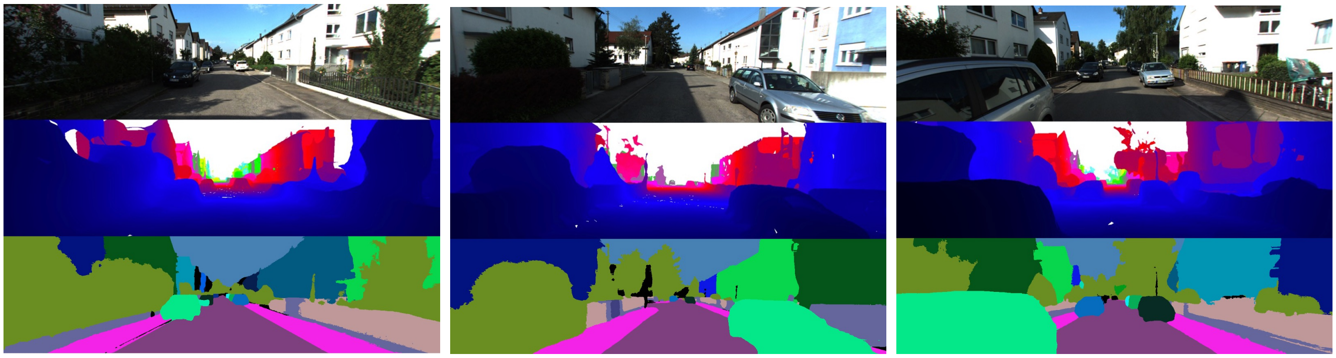

Figure 10: **Qualitative Results of 3D-2D CRF.** Top: Input RGB images. Middle: 3D-2D CRF mesh depth. Bottom: Panoptic label transfer results of the 3D-2D CRF method.

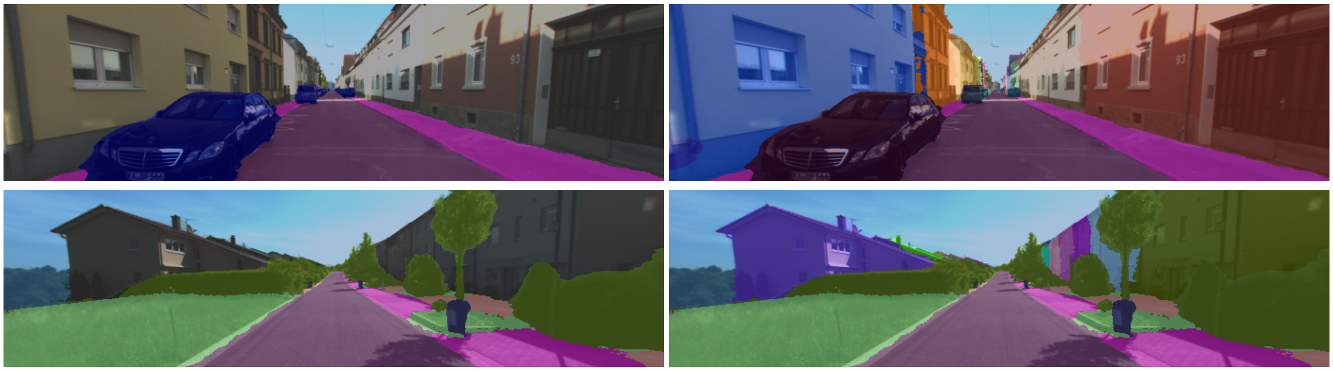

Figure 11: **Failure Cases**. Although our semantic map (left) is correct, the boundary of two adjacent buildings is not well segmented in the panoptic map (right).

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 1

[2] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2007. 3

[3] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 2

[4] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv.org*, 2109.13410, 2021. 1, 4, 7

[5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, pages 405–421. Springer, 2020. 1

[6] Xin Tong, Xianghua Ying, Yongjie Shi, He Zhao, and Ruibin Wang. Towards cross-view consistency in semantic segmentation while varying view direction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021. 1

[7] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 11–20, 2017. 4