

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

Xiao Fu<sup>1\*</sup>, Shangzhan Zhang<sup>1\*</sup>, Tianrun Chen<sup>1</sup>, Yichong Lu<sup>1</sup>, Lanyun Zhu<sup>2</sup>, Xiaowei Zhou<sup>1</sup>, Andreas Geiger<sup>3</sup>, and Yiyi Liao<sup>1</sup>

<sup>1</sup>Zhejiang University    <sup>2</sup>Singapore University of Technology and Design

<sup>3</sup>University of Tübingen and MPI-IS, Tübingen

<http://fuxiao0719.github.io/projects/panopticnerf/>

**Abstract.** Large-scale training data with high-quality annotations is critical for training semantic and instance segmentation models. Unfortunately, pixel-wise annotation is labor-intensive and costly, raising the demand for more efficient labeling strategies. In this work, we present a novel 3D-to-2D label transfer method, Panoptic NeRF, which aims for obtaining per-pixel 2D semantic and instance labels from easy-to-obtain coarse 3D bounding primitives. Our method utilizes NeRF as a differentiable tool to unify coarse 3D annotations and 2D semantic cues transferred from existing datasets. We demonstrate that this combination allows for improved geometry guided by semantic information, enabling rendering of accurate semantic maps across multiple views. Furthermore, this fusion process resolves label ambiguity of the coarse 3D annotations and filters noise in the 2D predictions. By inferring in 3D space and rendering to 2D labels, our 2D semantic and instance labels are multi-view consistent by design. Experimental results show that Panoptic NeRF outperforms existing semantic and instance label transfer methods in terms of accuracy and multi-view consistency on challenging urban scenes of the KITTI-360 dataset.

**Keywords:** 3D-to-2D label transfer, neural radiance field

## 1 Introduction

Semantic instance segmentation is an important machine perception task for autonomous driving. It is widely acknowledged that large-scale training data with high-quality annotations is critical to propel the performance of segmentation models. However, manual annotation of pixel-accurate segmentation masks is highly expensive and time-consuming. For example, annotating all instances in a single street scene image requires up to 1.5 hours [33]. While leveraging segmentation models pre-trained on existing datasets is promising to speed up the labeling process [65,34,30], human annotators still need to refine each frame individually, and label consistency across frames is hard to achieve.

It is worth mentioning that modern autonomous driving vehicles are equipped with multiple sensors to capture both 2D appearance and 3D geometric information [15,33,7], enabling efficient annotation in 3D space. By annotating the scene

---

\* Equal contribution.

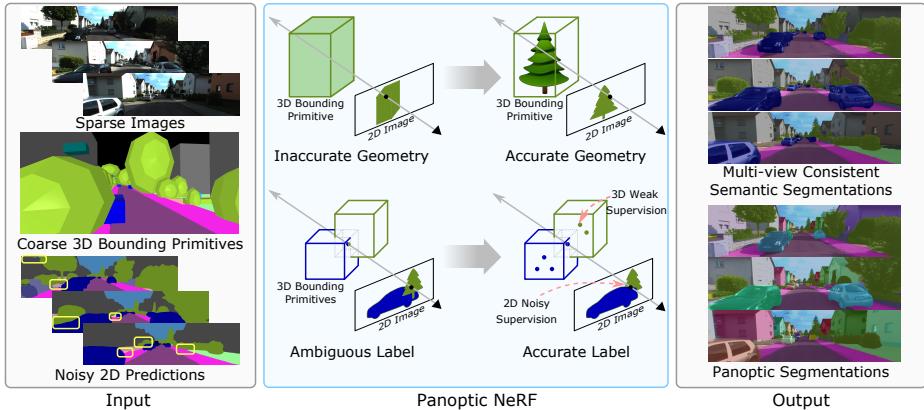


Fig. 1: **Panoptic NeRF** takes as input a set of sparse images, coarse 3D bounding primitives and noisy 2D predictions (yellow boxes highlight inaccurate predictions). By inferring in the 3D space, it generates semantic and instance labels in the 2D image space via volume rendering. This process requires addressing two challenges: 1) The underlying geometry needs to be recovered to render accurate 2D semantic maps over multiple views. 2) The 3D bounding primitives may intersect with each other and the 2D predictions are noisy, resulting in ambiguous labels that need to be resolved.

using coarse 3D bounding primitives (e.g., cuboids and ellipsoids) and projecting them into several 2D frames, annotation time can be significantly reduced to about 0.75 minutes per image as previously shown in [62,33]. Furthermore, it is often easier to separate instances in 3D rather than in the 2D image space (e.g., pedestrian in front of building). However, transferring coarse 3D bounding primitives to per-pixel 2D semantic and instance labels remains challenging. Existing 3D-to-2D label transfer methods either require manual post processing [31] or inference based on coarse 3D reconstructions [62,33], which may lead to inaccurate results.

In this paper, we aim to obtain accurate 2D semantic and instance labels from efficiently-labeled coarse 3D bounding primitives. While bounding primitives only provide weak semantic information in 3D space, state-of-the-art 2D segmentation models trained on existing labeled datasets allow for exploiting 2D RGB information and provide reliable 2D semantic cues on frequently occurring classes. To effectively unify weak 3D semantic information and 2D semantic cues in a single model, our key idea is to utilize Neural Radiance Fields (NeRF) [40] as a differentiable tool to bridge 3D and 2D space. Specifically, we introduce Panoptic NeRF, a novel label transfer method built on NeRF that infers in 3D space to render dense 2D semantic and instance labels, i.e., panoptic segmentation labels [28] (see Fig. 1).

To address this task, two key challenges have to be tackled: (1) *Geometric Reconstruction*: A high-quality geometric understanding of the underlying 3D scene is required in our setting to filter out noisy 2D labels and to render accurate

semantic maps across multiple views. However, obtaining accurate geometry using a vanilla NeRF model is hard, particularly in the driving scenario where input views are sparse; (2) *Semantics Estimation*: The 3D labels provided by the bounding primitives are ambiguous in the overlapping regions while the 2D predictions are noisy.

We tackle both challenges based on fusing 3D and 2D weak semantic information using our *dual semantic fields*. We demonstrate that noisy 2D semantic predictions can be leveraged to improve the underlying geometry when applied to a fixed semantic field determined by the 3D bounding primitives. Based on the improved geometry, the semantic renderings can be further refined by another learned semantic field that fuses information of the 3D bounding primitives and the 2D noisy predictions. As evidenced by our experiments, this fusion procedure is able to resolve the label ambiguity of the 3D bounding primitives and largely eliminate noise in the 2D predictions. Furthermore, Panoptic NeRF enables rendering globally consistent 2D instance maps across multiple frames, where each object has a unique instance index determined by the 3D bounding primitives.

Utilizing 3D bounding primitives of the recently released KITTI-360 dataset, Panoptic NeRF outperforms existing 3D-to-2D and 2D-to-2D label transfer methods. Our method can significantly reduce the labeling effort for human annotators, providing a promising approach to develop large-scale and densely labeled datasets for autonomous driving. We summarize our contributions as follows:

- 1) We formulate the 3D-to-2D label transfer task from the perspective of volume rendering. This allows us to unify easy-to-obtain 3D bounding primitives and noisy 2D semantic predictions in a single model, yielding high-quality panoptic labels.
- 2) By leveraging a novel dual formulation of the semantic fields, Panoptic NeRF effectively improves the geometric reconstruction given sparse views, yielding accurate object boundaries. Moreover, it is able to resolve label ambiguities and eliminates label noise based on the improved geometry.
- 3) Our Panoptic NeRF achieves superior performance compared to existing 3D-to-2D and 2D-to-2D label transfer methods in terms of both semantic and instance predictions. Furthermore, our 2D semantic and instance labels are multi-view and spatio-temporally consistent by design.

## 2 Related Work

**Urban Scene Segmentation:** Semantic instance segmentation is a critical task for autonomous vehicles [47,68]. While learning-based algorithms achieved compelling performance [9,32,66], they rely on large-scale annotated datasets. Unfortunately, annotating images at pixel level is extremely time-consuming and labor intensive, especially for instance-level annotation. While most urban datasets provides labels in 2D image space [5,11,43,26,60], autonomous vehicles are usually equipped with 3D sensors [15,7,17,23,33]. KITTI-360 [33] has demonstrated that annotating the scene using 3D bounding primitives can significantly

reduce the annotation time. However, transferring coarse 3D labels to 2D remains challenging. In this work, we focus on developing a novel 3D-to-2D label transfer method, exploiting recent advances in neural scene representations.

**Label Transfer:** There have been several attempts at improving label efficiency for individual frames [35, 20, 8, 1, 34, 2]. In this paper we focus on efficient labeling of video sequences. Existing works in this area can be divided into two categories: 2D-to-2D and 3D-to-2D. 2D-to-2D label transfer approaches reduce the workload by propagating labels across 2D images [51, 22, 49, 53, 14]. While the aforementioned methods only utilize information in 2D, 3D-to-2D methods exploit additional information in 3D for efficient labeling [23, 37, 42, 61, 6]. To obtain dense labels in 2D image space, some works [33, 62] perform per-frame inference jointly over the 3D point clouds and 2D pixels using a non-local multi-field Conditional Random Field (CRF) model. However, these methods require reconstructing a 3D mesh to project 3D point clouds to 2D. As it is treated as a pre-processing step, the mesh reconstruction is not jointly optimized in the CRF model. Thus, inaccurate reconstruction hinders label transfer performance. In contrast, Panoptic NeRF provides a novel end-to-end approach for 3D-to-2D label propagation where geometry and semantic estimations are jointly optimized.

**Coordinate-based Neural Representations:** Recently, coordinate-based neural representations has received wide attention in many areas, including 3D reconstruction [50, 24, 16, 39, 45, 63, 18, 46, 52, 55], novel view synthesis [40, 3, 25, 36], and 3D generative modeling [19, 38, 54]. In this paper, we focus on utilizing coordinate-based representations to estimate the semantics of the scene. A closely related work Semantic NeRF [67] explores NeRF for semantic fusion, label denoising, and label propagation. However, Semantic NeRF takes as input ground truth 2D labels or synthetic noise labels, which fails to produce correct labels when given real-world predictions from pre-trained 2D models. Moreover, Semantic NeRF operates in indoor scenes with dense RGB inputs and struggles in outdoor driving scenarios where input views are sparse, the challenging scenario our method focuses on. Finally, Semantic NeRF is limited to rendering semantic labels, whereas our method can render panoptic labels. A concurrent work NeSF [59] focuses on generalizable semantic field learning from density grids supervised by 2D GT labels, but we concentrate on the 3D-to-2D label transfer task without access to 2D GT.

### 3 Background

#### 3.1 NeRF

NeRF [40] models a 3D scene as a continuous neural radiance field. Specifically, it maps a 3D coordinate  $\mathbf{x}$  and a viewing direction  $\mathbf{d}$  to a volume density  $\sigma$  and an RGB color value  $\mathbf{c}$ :

$$f_\theta : (\mathbf{x} \in \mathbb{R}^3, \mathbf{d} \in \mathbb{S}^2) \mapsto (\sigma \in \mathbb{R}^+, \mathbf{c} \in \mathbb{R}^3) \quad (1)$$

Let  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$  denote a camera ray. The color at the corresponding pixel can be obtained by volume rendering

$$\mathbf{C}(\mathbf{r}) = \sum_{i=1}^N T_i(1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i \quad T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \quad (2)$$

where  $\sigma_i$  and  $\mathbf{c}_i$  denote the density and color value at a point  $i$  sampled along the ray,  $T_i$  denotes the transmittance at the sample point, and  $\delta_k = t_{i+1} - t_i$  is the distance between adjacent samples. Let  $\pi$  denote the volume rendering process of one ray. Enabled by volume rendering, NeRF learns  $f_\theta$  from a set of 2D RGB images with known camera poses.

### 3.2 Problem Formulation

As shown in Fig. 1, Panoptic NeRF aims to transfer coarse 3D bounding primitives to dense 2D semantic and instance labels. In addition to a sparse set of posed RGB images, we assume a set of 3D bounding primitives  $\beta = \{B_k\}_{k=1}^K$  to be available. These 3D bounding primitives cover the full scene in the form of cuboids, ellipsoids and extruded polygons. Each 3D bounding primitive  $B_k$  has a *semantic* label, belonging to either “stuff” or “thing”. For “thing” classes,  $B_k$  is additionally associated with a unique *instance* ID.

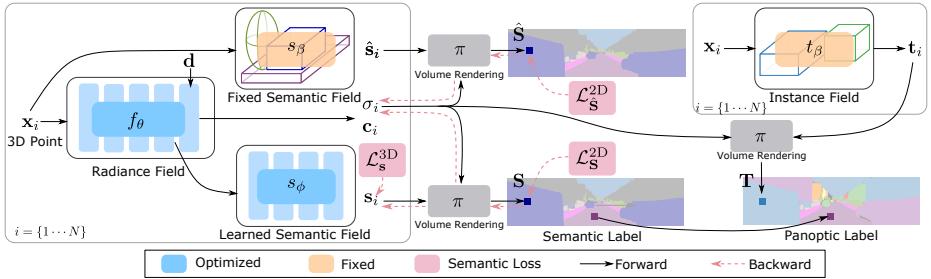
We further apply a semantic segmentation model to the RGB images to obtain a 2D semantic prediction for each image. Despite being noisy, semantic segmentation models usually perform well on frequently occurring classes, e.g., cars and roads. With this input information, our goal is to generate multi-view consistent, semantic and instance labels at the input frames.

## 4 Methodology

Panoptic NeRF provides a novel method for label transfer from 3D to 2D. Fig. 2 gives an overview of our method. We first map a 3D point  $\mathbf{x}_i$  to a density  $\sigma_i$  and a color value  $\mathbf{c}_i$ , as well as two semantic categorical distributions  $\hat{\mathbf{s}}_i$  and  $\mathbf{s}_i$  based on our dual semantic fields (Section 4.1). Correspondingly, for each camera ray, two semantic labels  $\hat{\mathbf{S}}$  and  $\mathbf{S}$  in the 2D image space via volume rendering  $\pi$  is obtained. Based on semantic losses in both 3D and 2D space (Section 4.2), the fixed semantic field  $s_\beta$  serves to improve geometry, while the learned semantic field  $s_\phi$  results in improved semantics. With the 3D bounding primitives, we further define a fixed instance field  $t_\beta$  that allows for rendering panoptic label  $\mathbf{T}$  when combined with the learned semantic field (Section 4.3).

### 4.1 Dual Semantic Fields

Given only 3D bounding primitives and noisy 2D predictions, both the underlying geometry and the semantic label need to be correctly estimated to render



**Fig. 2: Method Overview.** *Left* (Semantic Segmentation): We leverage dual semantic fields to obtain two semantic labels,  $\hat{\mathbf{s}}_i$  and  $\mathbf{s}_i$ , at each 3D location  $\mathbf{x}_i$ . The 3D semantic labels are accumulated along the ray and projected to the 2D image space via volume rendering, resulting in  $\hat{\mathbf{S}}$  and  $\mathbf{S}$ . The semantic losses applied to both semantic fields improves 1) the volume density  $\sigma_i$  and 2) the 3D semantic predictions  $\mathbf{s}_i$ . *Right* (Panoptic Segmentation): Our method allows for rendering panoptic labels by combining the learned semantic field and a fixed instance field determined by the 3D bounding primitives  $\beta$ .

accurate semantic maps. To jointly improve the geometry and the semantics, we define dual semantic fields, one is determined by the 3D bounding primitives  $\beta$  and the other is learned by a semantic head  $\phi$

$$s_\beta : \mathbf{x} \in \mathbb{R}^3 \mapsto \hat{\mathbf{s}} \in \mathbb{R}^{M_s} \quad s_\phi : \mathbf{x} \in \mathbb{R}^3 \mapsto \mathbf{s} \in \mathbb{R}^{M_s} \quad (3)$$

where  $M_s$  denotes the number of semantic classes. In combination with the volume density defined in Eq. 1, two semantic distributions  $\hat{\mathbf{S}}(\mathbf{r})$  and  $\mathbf{S}(\mathbf{r})$  can be obtained at each camera ray  $\mathbf{r}$  via the volume rendering operation  $\pi$ :

$$\hat{\mathbf{S}}(\mathbf{r}) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) \hat{\mathbf{s}}_i \quad \mathbf{S}(\mathbf{r}) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{s}_i \quad (4)$$

Note that  $\hat{\mathbf{S}}(\mathbf{r})$  and  $\mathbf{S}(\mathbf{r})$  are both normalized distributions when  $\sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) = 1$ . We set the background class to sky if  $\sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) < 1$ . We apply losses to both  $\hat{\mathbf{S}}(\mathbf{r})$  and  $\mathbf{S}(\mathbf{r})$  for training. During inference, the semantic label is determined as the class of the maximum probability in  $\mathbf{S}(\mathbf{r})$ .

**Fixed Semantic Field:** If  $\mathbf{x}$  is uniquely enclosed by a 3D bounding primitive  $B_k$ ,  $\hat{\mathbf{s}}$  is a fixed one-hot categorical distribution of the category of  $B_k$ . For a point  $\mathbf{x}$  enclosed by multiple 3D bounding boxes of different semantic categories, we assign equal probability to these plausible categories and 0 to the others. As explained in Section 4.2, the semantic field  $s_\beta$  is able to improve the geometry but cannot resolve the label ambiguity at the overlapping region.

**Learned Semantic Field:** We add a semantic head parameterized by  $\phi$  to NeRF to learn the semantic distribution  $\mathbf{s}$ . We apply a softmax operation at

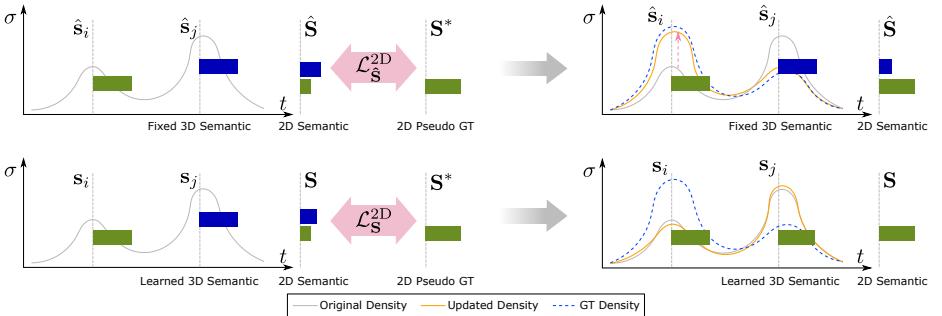


Fig. 3: **Semantically-Guided Geometry Optimization.** The top row illustrates a single ray of the fixed semantic field  $s_\beta$ . It demonstrates that  $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$  can only update the underlying geometry as the semantic distribution  $\hat{\mathbf{S}}$  is fixed. The second row shows a single ray of the learned semantic field  $s_\phi$ . In this case, the network can “cheat” by adjusting the semantic prediction  $\mathbf{s}$  to satisfy  $\mathcal{L}_{\mathbf{S}}^{2D}$  instead of updating the density  $\sigma$ .

each 3D point to ensure that  $\mathbf{s}$  is a categorical distribution. The detailed network structure can be found in the supplementary material.

## 4.2 Loss Functions

**Semantically-Guided Geometry Optimization:** In the driving scenario considered in our setting, the RGB images are sparse and the depth range is infinite. We observe that NeRF fails to recover reliable geometry in this setting using only the image reconstruction loss. However, we find that leveraging noisy 2D semantic predictions as pseudo ground truth is able to boost density prediction when applied to the fixed semantic fields  $s_\beta$

$$\mathcal{L}_{\hat{\mathbf{S}}}^{2D}(\theta) = - \sum_{\mathbf{r} \in \mathcal{R}} \sum_{k=1}^{M_s} \mathbf{S}_k^*(\mathbf{r}) \log \hat{\mathbf{S}}_k(\mathbf{r}) \quad (5)$$

where  $\hat{\mathbf{S}}_k(\mathbf{r})$  denotes the probability of the camera ray  $\mathbf{r}$  belonging to the class  $k$ , and  $\mathbf{S}^*$  denotes the pseudo-2D ground truth. As illustrated in Fig. 3, the key to improve density is to directly apply the semantic loss to the fixed semantic field  $s_\beta$ . As  $\hat{\mathbf{S}}$  is jointly determined by the radiance field  $f_\theta$  and the fixed semantic field  $s_\beta$ ,  $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$  can only be minimized by updating  $f_\theta$ . Depending on the correctness of  $\mathbf{S}^*$ , there are several cases: 1) If  $\mathbf{S}^*$  is correct, it increases the volume density of 3D points inside the correct bounding primitive and suppresses the density of others. 2) If  $\mathbf{S}^*$  is wrong and does not match any bounding primitive along the ray, it has no impact on  $f_\theta$ . 3) If  $\mathbf{S}^*$  is wrong but this class exists in one of the bounding primitives along the ray, it means  $\mathbf{S}^*$  corresponds to an occluded bounding primitive. To compensate, we introduce a weak depth supervision  $\mathcal{L}_d$  based on stereo matching to alleviate the misguidance of  $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$ . As shown in our

supplementary, using  $\mathcal{L}_d$  alone provides supervision to improve the overall geometry but fails to produce accurate object boundaries. In contrast, the semantic loss effectively improves the geometry at object boundaries, leading to accurate semantic maps via volume rendering.

**Joint Geometry and Semantic Optimization:** While enabling improved geometry, the 3D label of the overlapping regions remains ambiguous in the fixed semantic field. We leverage  $s_\phi$  to address this problem by jointly learning the semantic and the radiance fields. We apply a cross-entropy loss  $\mathcal{L}_{\mathbf{S}}^{2D}$  to each camera ray based on the filtered 2D pseudo ground truth, where  $\mathbf{u}(\mathbf{r})$  is set to 1 if  $\mathbf{S}^*(\mathbf{r})$  matches the semantic class of any bounding primitive along the ray and otherwise 0. To further suppress noise in the 2D predictions, we add a per-point semantic loss  $\mathcal{L}_{\mathbf{s}}^{3D}$  based on the 3D bounding primitives

$$\mathcal{L}_{\mathbf{S}}^{2D}(\theta, \phi) = - \sum_{\mathbf{r} \in \mathcal{R}} \mathbf{u}(\mathbf{r}) \sum_{k=1}^{M_s} \mathbf{S}_k^*(\mathbf{r}) \log \mathbf{S}_k(\mathbf{r}) \quad \mathcal{L}_{\mathbf{s}}^{3D}(\phi) = - \sum_{\mathbf{r} \in \mathcal{R}} \sum_{i=1}^N \mathbf{u}_i \sum_{k=1}^{M_s} \hat{\mathbf{s}}_i^k \log \mathbf{s}_i^k \quad (6)$$

where  $\mathbf{u}_i$  is a per-point binary mask.  $\mathbf{u}_i$  is set to 1 if (1)  $\mathbf{x}_i$  has a unique semantic label and (2) the density  $\sigma$  is above a threshold  $\sigma_{th}$  to focus on the object surface. As illustrated in Fig. 3,  $\mathcal{L}_{\mathbf{S}}^{2D}(\theta, \phi)$  does not necessarily improve the underlying geometry as the network can simply adjust the semantic head  $s_\phi$  to satisfy the loss. This behavior is also observed in novel view synthesis where NeRF does not necessarily recover good geometry when optimized for image reconstruction alone, specifically given sparse input views [12, 44].

**Total Loss:** Together, the total loss takes the form as

$$\mathcal{L} = \lambda_{\hat{\mathbf{S}}} \mathcal{L}_{\hat{\mathbf{S}}}^{2D} + \lambda_{\mathbf{S}} \mathcal{L}_{\mathbf{S}}^{2D} + \lambda_{\mathbf{s}} \mathcal{L}_{\mathbf{s}}^{3D} + \lambda_{\mathbf{C}} \underbrace{\sum_{\mathbf{r} \in \mathcal{R}} \|\mathbf{C}^*(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2}_{\mathcal{L}_p} + \lambda_d \underbrace{\sum_{\mathbf{r} \in \mathcal{R}} \|\mathbf{D}^*(\mathbf{r}) - \mathbf{D}(\mathbf{r})\|_2^2}_{\mathcal{L}_d} \quad (7)$$

where  $\mathcal{L}_p$  and  $\mathcal{L}_d$  denote the photometric loss and the depth loss, respectively.  $\lambda_{\hat{\mathbf{S}}}$ ,  $\lambda_{\mathbf{S}}$ ,  $\lambda_{\mathbf{s}}$ ,  $\lambda_{\mathbf{C}}$ , and  $\lambda_d$  are constant weighting parameters.  $\mathbf{C}^*(\mathbf{r})$  and  $\mathbf{C}(\mathbf{r})$  are the ground truth and rendered RGB colors for ray  $\mathbf{r}$ .  $\mathbf{D}^*(\mathbf{r})$  and  $\mathbf{D}(\mathbf{r})$  are pseudo ground truth depth generated by stereo matching and rendered depth, respectively. Please refer to the supplementary for more details of  $\mathbf{D}^*$ .

### 4.3 Rendering of Panoptic Labels

Based on our learned semantic field  $s_\phi$  and the 3D bounding primitives  $\beta$  with instance IDs, we can easily render a panoptic segmentation map.

Specifically, for a camera ray  $\mathbf{r}$ , the panoptic label directly takes the class with maximum probability in  $\mathbf{S}(\mathbf{r})$  if it is a “stuff” class. For “thing” classes, we render an instance distribution  $\mathbf{T}(\mathbf{r})$  based on the bounding primitives  $\beta$  to replace  $\mathbf{S}$  with  $\mathbf{T}$ . Our instance field is defined as follow

$$t_\beta : \mathbf{x} \in \mathbb{R}^3 \mapsto \mathbf{t} \in \mathbb{R}^{M_t} \quad (8)$$

where  $C_t$  is the number of the things in the scene and  $\mathbf{t}$  denotes a categorical distribution indicating which thing it belongs to. Note that  $\mathbf{t}$  is determined by the bounding primitives and is a one-hot vector if  $\mathbf{x}$  is uniquely enclosed by a bounding primitive of a thing. As overlap often occurs at the intersection of stuff and thing region, the bounding primitives of things rarely overlap with each other. Thus, this deterministic instance field leads to reliable performance in practice. The instance distribution at a camera ray  $\mathbf{T}$  can be obtained via volume rendering. To ensure that the instance label of this ray is consistent with the semantic class defined by  $\mathbf{S}$ , we mask out instances belonging to other semantic classes by setting their probabilities to 0 in  $\mathbf{T}$ . Finally, the instance label can be determined as the class with maximum probability within  $\mathbf{T}$ .

#### 4.4 Implementation Details

**Sampling Strategy and Sky Modeling:** With the 3D bounding primitives covering the full scene, we sample points inside the bounding primitives to skip empty space. Specifically, for each camera ray, we sort all bounding primitives that the ray hits from near to far and save the intersections offline. As our bounding primitives are convex, each ray intersects a bounding primitive exactly twice which determines the sampling interval. 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 and  $t_{int}$  is a constant distance interval. Our sampling strategy allows the network to focus on the non-empty region. As evidenced by our experiments, this is particularly beneficial in unbounded outdoor environments.

**Training:** We optimize one Panoptic NeRF model per scene, using a single NVIDIA 3090. For each scene, we set the origin to the center of the scene. We use Adam [27] with a learning rate of 5e-4 to train our models. We set loss weights to  $\lambda_{\hat{\mathbf{s}}} = 1, \lambda_{\mathbf{S}} = 1, \lambda_{\mathbf{s}} = 1, \lambda_{\mathbf{C}} = 1, \lambda_d = 0.1$ , and the density threshold to  $\sigma_{th} = 0.1$ . We optimize the total loss  $\mathcal{L}$  for 80,000 iterations.

## 5 Experiments

**Dataset:** We conduct experiments on the recently released KITTI-360 [33] dataset. KITTI-360 is collected in suburban areas and provides 3D bounding primitives covering the full scene. Following [33], we evaluate Panoptic NeRF on manually annotated frames from 5 static suburbs. We split these 5 suburbs into 10 scenes, comprising 128 consecutive frames 128 consecutive frames each with

---

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

an average travel distance of 0.8m between frames. We leverage all 128 pairs of posed stereo images for training. KITTI-360 provides a set of manually labeled frames sampled in equidistant steps of 5 frames. We use half of the manually labeled frames for evaluation and provide the other half as input to 2D-to-2D label transfer baselines. We improve the quality of the manually labeled ground truth which is inaccurate in ambiguous at ambiguous regions. A qualitative comparison of original and modified labels can be found in the supplementary.

**Baselines:** We compare against top-performing baselines in two categories: (1) *2D-to-2D label transfer baselines*, including Fully Connected CRF (FC CRF) [29] and Semantic NeRF [67]. For both baselines, we provide manually annotated 2D frames as input, sampled at equidistant steps of 10 frames. Note that labeling these 2D frames takes similar or longer compared to annotating 3D bounding primitives [62]. As these 2D annotations are extremely sparse, we further provide the same pseudo-2D labels used in our method to Semantic NeRF (denoted as Semantic NeRF\*). (2) *3D-to-2D label transfer baselines*, including PSPNet\*, 3D Primitives/Meshes/Points+GC [33], and 3D-2D CRF [33]. All these baselines leverage the same 3D bounding primitives to transfer labels to 2D. Here, PSP-Net\* is considered 3D-to-2D as it is pre-trained on Cityscapes and fine-tuned on KITTI-360 based on the 3D sparse label projections. The second set of baselines first project 3D primitives/meshes/points to 2D and then apply Graph Cut to densify the label. The 3D-2D CRF densely connects 2D image pixels and 3D LiDAR points, performing inference jointly on these two fields with a set of consistency constraints.

**Pseudo 2D GT:** We use PSPNet\* to provide pseudo ground truth in our main experiment to supervise our dual semantic fields. This ensures fair comparison to the 3D-2D CRF, which takes the predictions of PSPNet\* as unary terms. While PSPNet\* reduces the domain gap by fine-tuning on KITTI-360, the 3D sparse labels are obtained from LiDAR points and rely on a 3D mesh [33] to determine non-occluded points. To further simplify the entire process, in the ablation study we investigate the performance of our method using pre-trained models on Cityscape without any fine-tuning, including PSPNet [66], Deeplab [9] and Tao *et al.* [56].

**Metrics:** We evaluate semantic labels by the mean Intersection over Union (mIoU) and the average pixel accuracy (Acc) metrics. To quantitatively evaluate multi-view consistency (MC), we utilize LiDAR points to retrieve corresponding pixel pairs between two consecutive evaluation frames. The MC metric is then calculated as the ratio of pixels pairs with consistent semantic labels over all pairs. For evaluating panoptic segmentation, we report Panoptic Quality (PQ) [28], which can be decomposed into Segmentation Quality (SQ) and Recognition Quality (RQ). To verify that Panoptic NeRF is able to improve the underlying geometry, we further evaluate the rendered depth compared to sparse depth maps obtained from LiDAR. Following common practice [4,13], we evaluate depth predictions using Root Mean Squared Error (RMSE) and the ratio of

Table 1: **Quantitative Comparison of Semantic Label Transfer.** Metrics are averaged over the 10 scenes from the KITTI-360 dataset.

Method	Road	Park	Sdwlk	Terr	Bldg	Vegt	Car	Trler	Crvn	Gate	Wall	Fence	Box	Sky	mIoU	Acc	MC
FC CRF [29]	90.3	49.9	67.7	62.5	88.3	79.2	85.6	48.9	78.1	23.4	35.3	46.5	42.0	92.7	63.6	89.1	85.41
S-NeRF [67]	87.0	35.8	64.7	58.2	83.4	76.3	70.3	93.5	76.5	41.4	44.0	52.6	29.0	92.0	64.6	86.8	88.98
S-NeRF* [67]	94.6	52.9	77.7	65.0	88.0	80.8	87.9	58.3	86.0	36.0	44.1	56.8	42.2	90.9	68.6	90.5	93.42
PSPNet* [66]	95.5	49.7	77.5	66.7	88.9	82.4	91.6	46.5	83.1	24.2	43.3	51.3	51.1	89.3	67.2	90.7	91.79
3D Prim.+GC	81.7	31.0	45.6	22.5	59.6	56.7	63.0	61.7	37.3	61.6	28.8	50.6	39.5	50.3	49.3	73.4	86.56
3D Mesh+GC	91.7	53.1	67.2	31.4	81.3	72.1	85.2	93.5	86.0	65.2	40.7	59.7	54.4	65.6	67.7	86.0	94.99
3D Point+GC	93.5	59.0	76.1	37.2	82.0	74.1	87.5	94.7	85.7	66.7	59.4	65.9	58.6	68.0	72.0	87.9	<b>96.51</b>
3D-2D CRF [33]	95.2	64.2	83.8	67.9	90.3	84.2	92.2	93.4	90.8	68.2	<b>64.5</b>	70.0	55.8	92.8	79.5	92.8	94.98
Ours	<b>95.6</b>	<b>68.4</b>	<b>84.1</b>	<b>69.5</b>	<b>91.0</b>	<b>84.4</b>	<b>93.0</b>	<b>94.8</b>	<b>93.7</b>	<b>71.6</b>	63.1	<b>74.4</b>	<b>59.7</b>	91.7	<b>81.1</b>	<b>93.2</b>	95.02

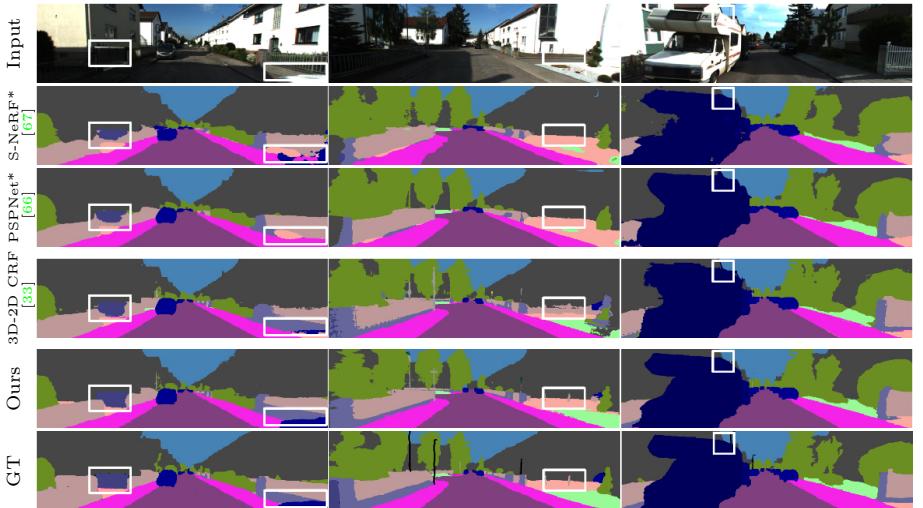


Fig. 4: **Qualitative Comparison of Semantic Label Transfer.** Our method achieves superior performance in challenging regions compared to the baselines, e.g. in under- or over-exposed regions, by recovering the underlying 3D geometry, see box next to the building (left) and building above the caravan (right).

accurate predictions ( $\delta_{1.25}$ ), where a pixel is considered accurate if the relative error is below 1.25.

## 5.1 Label Transfer

We evaluate our model on KITTI-360 and compare it to our baselines in terms of both semantic and instance label transfer. As most baselines are not designed for panoptic label transfer, we first compare the semantic predictions of all methods and then compare our panoptic predictions to the 3D-2D CRF. We provide additional results on both left and right stereo views in the supplementary.

**Semantic Label Transfer:** As evidenced by Table 1, our method achieves the highest mIoU and Acc over a line of baselines. Specifically, compared to 3D-2D CRF, we obtain an absolute improvement of 1.6% (79.5%  $\rightarrow$  81.1%) on mIoU.

Among 14 classes, we outperform the baseline on 12 classes, especially on “Park” (4.2%), “Terrain” (4.2%), “Caravan” (2.9%), “Gate” (3.4%) and “Fence” (4.4%). Despite PSPNet\* being finetuned on KITTI-360 which reduces the performance gap, our method outperforms PSPNet\* by a significant margin. Supervised by the extremely sparse manually annotated GT, Semantic NeRF struggles to produce reliable performance. Using pseudo labels of PSPNet\*, Semantic NeRF\* is capable of denoising and thus improving performance ( $67.2\% \rightarrow 68.6\%$ ). However, both Semantic NeRF and Semantic NeRF\* are inferior in the urban scenario when the input views are sparse. Moreover, in terms of MC, our method slightly outperforms the 3D-2D CRF and significantly surpasses 2D-to-2D label transfer methods. While our method slightly lags behind 3D Point+GC in terms of MC, it is reasonable as the label consistency is evaluated on the sparsely projected 3D points which GC takes as input to generate a dense label map.

Table 2: **Quantitative Comparison of Panoptic Label Transfer.** Metrics are averaged over all 10 test scenes from the KITTI-360 dataset.

Method		PQ	SQ	RQ	Method		PQ	SQ	RQ
3D-2D CRF	All	62.2	79.1	76.9	Ours	All	<b>64.4</b>	<b>79.3</b>	<b>79.6</b>
	Things	60.7	79.5	<b>75.2</b>		Things	<b>61.9</b>	<b>80.7</b>	<b>75.2</b>
	Stuff	63.0	<b>78.9</b>	77.9		Stuff	<b>65.8</b>	78.6	<b>82.0</b>

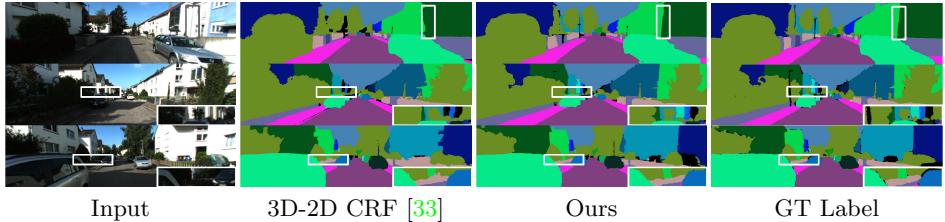


Fig. 5: **Qualitative Comparison of Panoptic Label Transfer.** Our method is capable of distinguishing instances based on inferring in the 3D space. In contrast, 3D-2D CRF struggles in at far and overexposed areas.

**Panoptic Label Transfer:** We split the instance labels into things and stuff classes in KITTI-360. As the class “building” is classified as thing in KITTI-360 but stuff in Cityscapes, our 2D baselines [10] are not suitable for testing performance on KITTI-360. Therefore, we ignore pre-trained 2D SOTA baselines. As shown in Table 2, our proposed method outperforms the 3D-2D CRF in both things and stuff classes. A visual comparison is shown in Fig. 5. As can be seen, we can deal well with overexposure at buildings, which is a challenge for the 3D-2D CRF, as it projects LiDAR points to reconstruct the intermediate meshes whose quality suffers on building class. More details can be found in the supplementary material.

Table 3: **Ablation Study.** Metrics are averaged over 1 test scene on KITTI-360.

	Depth(0-100m) RMSE↓ $\delta_{1.25} \uparrow$	Evaluated Label	Semantic														
			Road	Park	Sdwlk	Bldg	Vegt	Car	Trlr	Gate	Wall	Fence	Box	Sky	mIoU	Acc	
3D-2D CRF	-	-	-	<b>97.1</b>	70.6	85.0	93.0	84.8	88.8	67.4	69.3	67.7	72.1	61.3	95.5	79.4	93.7
NeRF*	10.71	78.4	<b>S</b> (fixed $s_\beta$ )	90.1	35.1	68.4	89.6	72.4	88.0	66.5	26.9	58.5	52.8	65.3	92.9	67.2	88.6
w/o $\mathcal{L}_S^{2D}$	6.28	93.1	<b>S</b> (learned $s_\phi$ )	95.4	55.6	83.8	93.2	83.0	91.0	64.9	50.1	68.0	70.8	<b>72.6</b>	94.7	76.9	93.1
w/o $\mathcal{L}_d$	6.15	94.0	<b>S</b> (learned $s_\phi$ )	96.4	70.3	84.6	93.9	<b>85.7</b>	90.6	65.0	69.9	61.5	76.4	58.9	95.5	79.1	94.1
Uniform S.	6.28	91.1	<b>S</b> (learned $s_\phi$ )	92.8	53.2	75.1	92.6	83.3	83.5	56.7	<b>70.9</b>	55.9	70.0	52.7	94.8	73.5	92.2
w/o $\mathcal{L}_S^{3D}$	6.01	95.0	<b>S</b> (learned $s_\phi$ )	95.6	69.7	83.4	93.7	85.4	91.8	<b>70.0</b>	68.7	69.7	77.0	69.9	95.5	80.8	94.2
w/o $\mathcal{L}_S^*$	6.01	95.0	$\hat{S}$ (fixed $s_\beta$ )	95.3	65.7	83.0	93.8	85.3	91.7	63.0	65.3	67.2	76.4	70.1	95.5	79.4	94.0
w/o $\mathcal{L}_S^{3D}$	5.74	95.0	<b>S</b> (learned $s_\phi$ )	96.1	53.7	75.1	93.3	85.1	87.3	60.9	70.7	58.5	74.0	0.0	93.5	70.7	92.8
w/o $\mathbf{u(r)}$	5.84	94.8	<b>S</b> (learned $s_\phi$ )	96.2	70.1	<b>87.0</b>	93.7	85.6	<b>92.6</b>	64.3	69.9	70.6	78.1	66.2	95.6	80.8	94.4
Complete	<b>5.23</b>	<b>95.1</b>	<b>S</b> (learned $s_\phi$ )	96.5	<b>72.9</b>	86.0	<b>93.9</b>	85.3	91.7	65.6	70.2	<b>71.7</b>	<b>78.1</b>	68.8	<b>95.6</b>	<b>81.4</b>	<b>94.5</b>

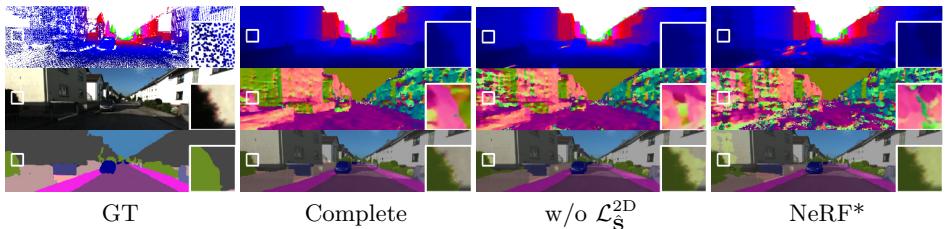


Fig. 6: **Ablation Study.** Top: LiDAR depth map (visually enhanced) and rendered depth maps. Middle: Normal maps obtained from depth maps. Bottom: Semantic GT and predictions. Note that removing  $\mathcal{L}_S^{2D}$  leads to over-smooth object boundaries and inaccurate semantic segmentation.

## 5.2 Ablation Studies

We validate our pipeline’s design modules with an extensive ablation in Table 3 by removing one component at a time. As there is a positive correlation between semantic labels and panoptic labels, we focus on semantic segmentation for this experiment which is conducted on one scene comprising 12 semantic classes.

**Geometric Reconstruction:** We now verify that our method effectively improves the underlying geometry leveraging semantic information. We first remove all the other losses except for  $\mathcal{L}_C$ , leading to a baseline very similar to NeRF except for the different sampling strategy used in our work (NeRF\*). In this case we render a semantic map based on the fixed semantic field  $s_\beta$  for comparison. As can be seen from Table 3 and Fig. 6, the underlying geometry of NeRF\* drops significantly with only  $\mathcal{L}_C$ . More importantly, the depth prediction also degrades considerably when removing  $\mathcal{L}_S^{2D}$  (w/o  $\mathcal{L}_S^{2D}$ ), indicating the importance of the fixed semantic field in improving the underlying geometry. Fig. 6 shows that the full model has sharper edges while removing the fixed semantic fields leads to over-smooth object boundaries. We further show that eliminating  $\mathcal{L}_d$  (w/o  $\mathcal{L}_d$ ) or replacing the sampling strategy with standard uniform sampling (Uniform S.) both impair the geometric reconstruction, and consequently the semantic estimation as well.

**Semantic Segmentation:** When removing  $\mathcal{L}_S$  guided by the pseudo-2D ground truth (w/o  $\mathcal{L}_S$ ), the performance also drops as the learned semantic field  $s_\phi$  is

Table 4: Quantitative Comparison using different 2D Pseudo GTs.

Method	Road	Park	Sdwlk	Bldg	Vegt	Car	Trler	Gate	Wall	Fence	Box	Sky	mIoU	mIoU <sub>sub</sub>	Acc
Deeplab [9]	92.1	-	67.4	89.3	85.0	87.2	-	-	31.7	57.8	-	87.4	-	74.7	88.1
Tao <i>et al.</i> [56]	96.6	-	65.4	94.7	87.7	94.2	-	-	38.3	66.2	-	96.7	-	79.9	91.2
PSPNet [66]	84.2	-	49.7	90.9	83.0	91.5	-	-	22.8	56.0	-	88.7	-	70.9	87.0
PSPNet* [66]	96.5	49.5	76.3	92.6	82.9	87.0	0.0	30.0	40.0	51.1	44.7	93.3	62.0	77.5	90.1
Ours w/ Deeplab [9]	93.1	63.8	80.2	93.5	87.2	90.9	55.7	73.6	73.2	79.5	68.3	93.1	79.3	86.3	94.0
Ours w/ Tao <i>et al.</i> [56]	96.0	63.7	81.6	94.4	88.0	93.6	55.4	79.9	61.4	77.2	69.9	96.7	79.8	86.1	94.5
Ours w/ PSPNet [66]	89.6	55.1	72.2	91.6	78.2	89.3	70.9	58.4	67.2	69.2	72.7	96.1	75.9	81.7	91.6
Ours w/ PSPNet* [66]	96.5	72.9	86.0	93.9	85.3	91.7	65.6	70.2	71.7	78.1	68.8	95.6	81.4	87.3	94.5

only supervised by weak 3D supervision. Interestingly, this baseline still outperforms the semantic map rendered by the fixed semantic field despite that they share the same geometry. This observation suggests that the weak 3D supervision provided by  $\mathcal{L}_s$  also allows to address the label ambiguity of the overlapping region to a certain extent. Therefore, it is not surprising that removing  $\mathcal{L}_s$  (w/o  $\mathcal{L}_s$ ) worsens the semantic prediction compared to the full model. Note that  $\mathcal{L}_s$  is particularly important for the less-frequently observed class “Box”. Moreover, we observe that performance is also deteriorated without ray masking (w/o  $\mathbf{u}(\mathbf{r})$ ).

**2D Pseudo GT:** Finally, we evaluate how our method is affected by the quality of the pseudo 2D ground truth in Table 4. As there are four classes not considered during training in Cityscapes, we omit these classes and additionally report mIoU<sub>sub</sub> over the remaining classes for all methods. Experimental results shows that the fine-tuned 2D pseudo GT PSPNet\* still yields the best overall performance. However, it is worth to note that using models pre-trained on Cityscapes without any fine-tuning leads to promising results, where the mIoU<sub>sub</sub> of Ours w/ Deeplab and Ours w/ Tao *et al.* is very close to Ours w/ PSPNet\*. More importantly, our method consistently outperforms the corresponding pseudo GT leveraging the 3D bounding primitives.

### 5.3 Limitations

Our method performs per-scene optimization and training takes 4 hours on each scene. Training time needs to be reduced to scale our method to large-scale scenes. Fortunately, training time can be significantly reduced leveraging recent advances in speeding up NeRF training, e.g., Instance NGP [41] and Plenoxels [64]. In addition, we consider label transfer on static scenes in this work. We plan to extend our method to dynamic scenes leveraging recent advances in dynamic radiance field estimation [48].

## 6 Conclusions

We present Panoptic NeRF, a 3D-to-2D label transfer method that infers in 3D space and renders per-pixel semantic and instance labels in 2D. By combining coarse 3D bounding primitives and noisy 2D predictions using our dual semantic fields, Panoptic NeRF is capable of improving the underlying geometry given

sparse input views and resolving label noise. We believe that our method is a step towards more efficient data annotation, while simultaneously providing a 3D consistent continuous panoptic representation of the scene.

## References

1. Acuna, D., Ling, H., Kar, A., Fidler, S.: Efficient interactive annotation of segmentation datasets with polygon-rnn++ (2018) **4**
2. Andriluka, M., Uijlings, J., Ferrari, V.: Fluid annotation: a human-machine collaboration interface for full image annotation (2018) **4**
3. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2021) **4**
4. Bhoi, A.: Monocular depth estimation: A survey. arXiv.org (2019) **10**
5. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters **30**(2), 88–97 (1 2009) **3**
6. Bruls, T., Maddern, W., Morye, A.A., Newman, P.: Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data. In: Proc. IEEE International Conf. on Robotics and Automation (ICRA) (2018) **4**
7. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Lioung, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2020) **1, 3**
8. Castrejon, L., Kundu, K., Urtasun, R., Fidler, S.: Annotating object instances with a polygon-rnn. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2017) **4**
9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) **40**(4), 834–848 (2017) **3, 10, 14**
10. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2020) **12**
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2016) **3, 20**
12. Deng, K., Liu, A., Zhu, J., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. arXiv.org **2107.02791** (2021) **8**
13. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2018) **10**
14. Ganeshan, A., Vallet, A., Kudo, Y., Maeda, S., Kerola, T., Ambrus, R., Park, D., Gaidon, A.: Warp-refine propagation: Semi-supervised auto-labeling via cycle-consistency. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2021) **4**

15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2012) [1](#), [3](#)
16. Genova, K., Cole, F., Sud, A., Sarna, A., Funkhouser, T.: Local deep implicit functions for 3d shape. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2020) [4](#)
17. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., Schubert, P.: A2D2: audi autonomous driving dataset. arXiv.org **2004.06320** (2020) [3](#)
18. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. arXiv.org (2020) [4](#)
19. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenet: A style-based 3d-aware generator for high-resolution image synthesis. arXiv.org (2021) [4](#)
20. Guillumin, M., Küttel, D., Ferrari, V.: Imagenet auto-annotation with segmentation propagation. International Journal of Computer Vision (IJCV) **110**(3), 328–348 (2014) [4](#)
21. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) **30**(2), 328–341 (2008) [22](#)
22. Hong, S., Oh, J., Lee, H., Han, B.: Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2016) [4](#)
23. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (2020) [3](#), [4](#)
24. Jiang, C., Sud, A., Makadia, A., Huang, J., Nießner, M., Funkhouser, T., et al.: Local implicit grid representations for 3d scenes. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2020) [4](#)
25. Kellnhofer, P., Jebe, L.C., Jones, A., Spicer, R., Pulli, K., Wetzstein, G.: Neural lumigraph rendering. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2021) [4](#)
26. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Video panoptic segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2020) [3](#)
27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv.org (2014) [9](#)
28. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) [2](#), [10](#), [21](#)
29. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Advances in Neural Information Processing Systems (NIPS) (2011) [10](#), [11](#)
30. Lee, J., Walsh, S., Harakeh, A., Waslander, S.L.: Leveraging pre-trained 3d object detection models for fast ground truth generation. In: Proc. IEEE Conf. on Intelligent Transportation Systems (ITSC) (2018) [1](#)
31. Lewis, S.E., Searle, S., Harris, N., Gibson, M., Iyer, V., Richter, J., Wiel, C., Bayraktaroglu, L., Birney, E., Crosby, M., et al.: Apollo: a sequence annotation editor. GENOME BIOL **3**(12), 1–14 (2002) [2](#)

32. Li, X., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tan, S., Tong, Y.: Semantic flow for fast and accurate scene parsing. In: European Conference on Computer Vision. pp. 775–793. Springer (2020) [3](#)
33. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. arXiv.org (2021) [1](#), [2](#), [3](#), [4](#), [9](#), [10](#), [11](#), [12](#), [20](#), [23](#), [26](#)
34. Ling, H., Gao, J., Kar, A., Chen, W., Fidler, S.: Fast interactive object annotation with curve-gcn. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) [1](#), [4](#)
35. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing via label transfer. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) **33**(12), 2368–2382 (2011) [4](#)
36. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2021) [4](#)
37. Martinovic, A., Knopp, J., Riemenschneider, H., Van Gool, L.: 3d all the way: Semantic segmentation of urban scenes from start to end in 3d. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2015) [4](#)
38. Meng, Q., Chen, A., Luo, H., Wu, M., Su, H., Xu, L., He, X., Yu, J.: Gnerf: Gan-based neural radiance field without posed camera. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2021) [4](#)
39. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) [4](#)
40. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proc. of the European Conf. on Computer Vision (ECCV) (2020) [2](#), [4](#), [20](#)
41. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. arXiv.org (Jan 2022) [14](#)
42. Mustafa, A., Hilton, A.: Semantically coherent co-segmentation and reconstruction of dynamic scenes. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2017) [4](#)
43. Neuhold, G., Ollmann, T., Rota Bulò, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2017) [3](#)
44. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. arXiv.org (2021) [8](#)
45. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2020) [4](#)
46. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2021) [4](#)
47. Ohn-Bar, E., Prakash, A., Behl, A., Chitta, K., Geiger, A.: Learning situational driving. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2020) [3](#)
48. Ost, J., Mannan, F., Thuerey, N., Knodt, J., Heide, F.: Neural scene graphs for dynamic scenes. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2021) [14](#)

49. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2015) 4
50. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019) 4
51. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2015) 4
52. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: Proc. of the European Conf. on Computer Vision (ECCV) (2020) 4
53. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2015) 4
54. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. In: Advances in Neural Information Processing Systems (NIPS) (2020) 4
55. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: Advances in Neural Information Processing Systems (NIPS) (2020) 4
56. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020) 10, 14
57. Tong, X., Ying, X., Shi, Y., Zhao, H., Wang, R.: Towards cross-view consistency in semantic segmentation while varying view direction. In: Zhou, Z. (ed.) Proc. of the International Joint Conf. on Artificial Intelligence (IJCAI) (2021) 21
58. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity Invariant CNNs. In: Proc. of the International Conf. on 3D Vision (3DV) (2017) 23
59. Vora, S., Radwan, N., Greff, K., Meyer, H., Genova, K., Sajjadi, M.S., Pot, E., Tagliasacchi, A., Duckworth, D.: Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. arXiv.org (2021) 4
60. Weber, M., Xie, J., Collins, M., Zhu, Y., Voigtlaender, P., Adam, H., Green, B., Geiger, A., Leibe, B., Cremers, D., Osep, A., Leal-Taixe, L., Chen, L.C.: STEP: Segmenting and tracking every pixel. In: Proc. of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks (2021) 3
61. Xiao, J., Quan, L.: Multiple view semantic segmentation for street view images. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2009) 4
62. Xie, J., Kiefel, M., Sun, M.T., Geiger, A.: Semantic instance annotation of street scenes by 3d to 2d label transfer. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2016) 2, 4, 10
63. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance (2020) 4
64. Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxtels: Radiance fields without neural networks. arXiv.org (2021) 14
65. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv.org 2(5), 6 (2018) 1

66. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017) [3](#), [10](#), [11](#), [14](#)
67. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2021) [4](#), [10](#), [11](#)
68. Zhou, B., Krähenbühl, P., Koltun, V.: Does computer vision matter for action? Science Robotics **4**(30) (2019) [3](#)

## A Implementation Details

### A.1 Network Architecture

Fig. 7 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 [40], both  $\mathbf{x}$  and  $\mathbf{d}$  are mapped to a higher dimensional space using a positional encoding (PE):

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \dots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)) \quad (9)$$

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.

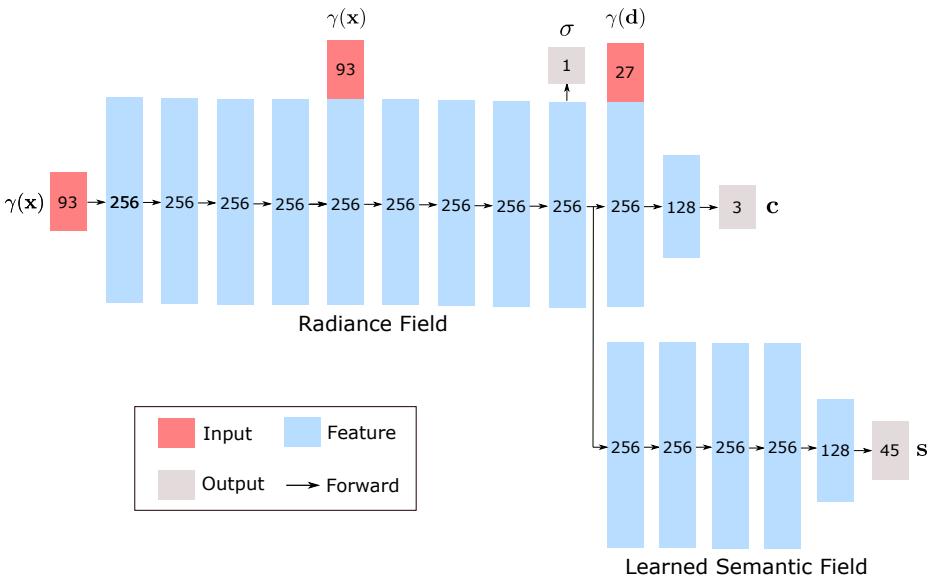


Fig. 7: **Trainable Part of Panoptic NeRF.** For the radiance field, we follow the original implementation of NeRF except for setting  $L = 15$  for  $\gamma(\mathbf{x})$ . The learned semantic field predicts semantic logits independent of the viewing direction. The logits are then transformed into categorical distributions through a softmax layer.

### A.2 Evaluation Metric

We evaluate mIoU and pixel accuracy following standard practice [11,33]. 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 [57] 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 B.2. Given LiDAR 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 [28], 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.

### A.3 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  $\mathcal{L}_c$  and the depth loss  $\mathcal{L}_d$ . During per-scene optimization, the photometric loss  $\mathcal{L}_c$  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 C.3 and novel views in Section C.4.

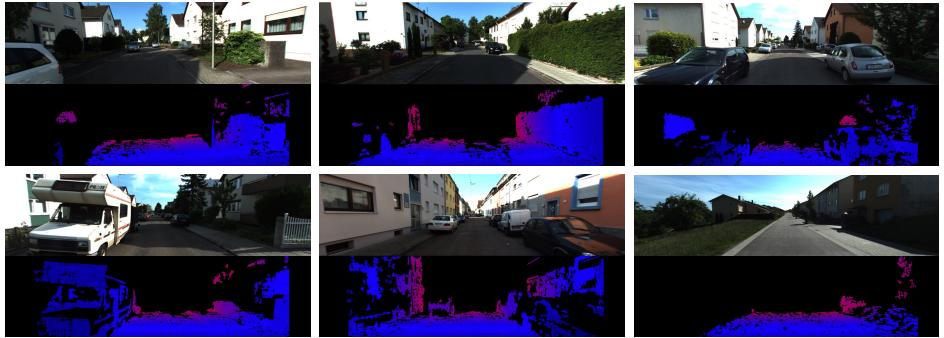


Fig. 8: **Depth Maps for Weak Depth Supervision.** Each group shows the RGB image (top) and the corresponding depth maps (bottom) used for supervision.

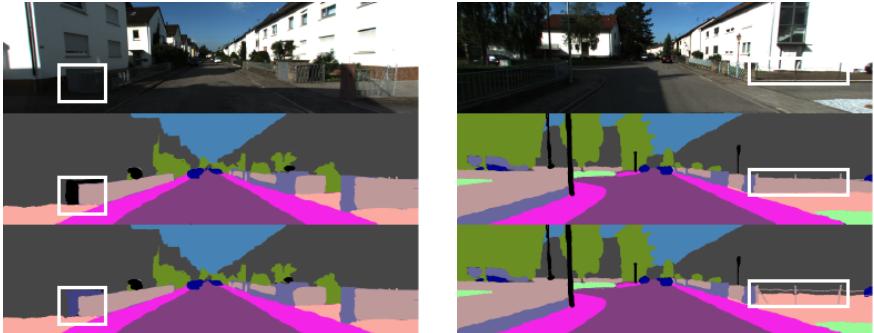


Fig. 9: **Examples of Modified Ground Truth.** We correct some GT pixels that were incorrectly labeled in the KITTI-360 dataset. Top: Input RGB images. Middle: Original ground truth. Bottom: Modified ground truth. In the first column, we add the “box” class. In the second column, we correct the “parking” area.

## B Data Preparation

### B.1 Stereo Depth for Weak Depth Supervision

To provide weak depth supervision to Panoptic NeRF, we use Semi-Global Matching (SGM) [21] 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. 8.

### B.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

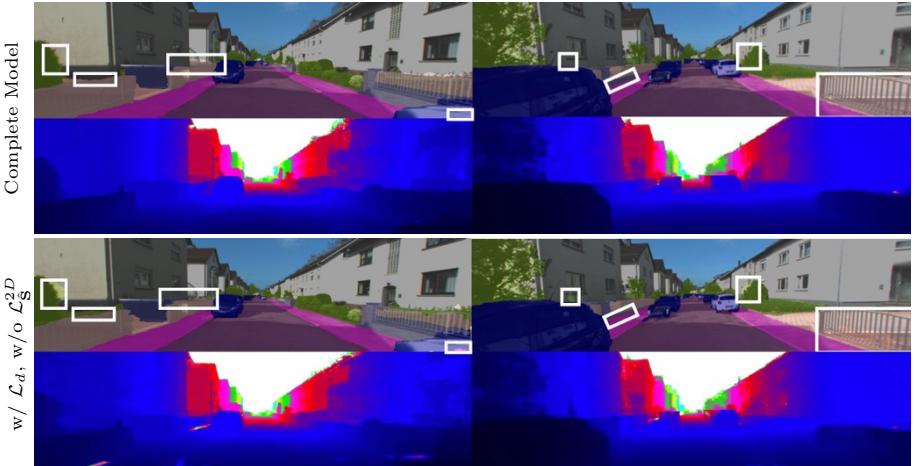


Fig. 10: **Qualitative Comparison of Ablation Study.** We visualize the semantic map and depth map of the complete model (top) and the model without fixed semantic field.

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 LiDAR observations and project the visible points to each frame similar to [58].

### B.3 Manually Annotated 2D GT

The manually annotated 2D ground truth of KITTI-360 [33] is inferior at some regions. For a fair comparison, we improve the label quality by manually relabeling ambiguous classes, see Fig. 9 for illustrations.

## C Additional Experimental Results

### C.1 Depth Loss

We show that using the depth loss  $\mathcal{L}_d$  alone is not able to recover accurate object boundaries in Fig. 10. In contrast, adding the semantic loss  $\mathcal{L}_{\hat{s}}^{2D}$  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  $\mathcal{L}_{\hat{s}}^{2D}$  defined on the fixed semantic field provides a strong penalty as small errors in depth lead to wrong semantics.

## C.2 Qualitative Comparison of Label Transfer

Fig. 11 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. 12 and panoptic label transfer in Fig. 13.

## C.3 Stereo Label Transfer

In Fig. 14, 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.

## C.4 Novel View Label Transfer

Here, we evaluate our performance on novel view label transfer by applying the photometric loss  $\mathcal{L}_c$  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. 15, our method achieves promising results on novel view appearance and label synthesis.

## C.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. 16 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.

## C.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. 17.

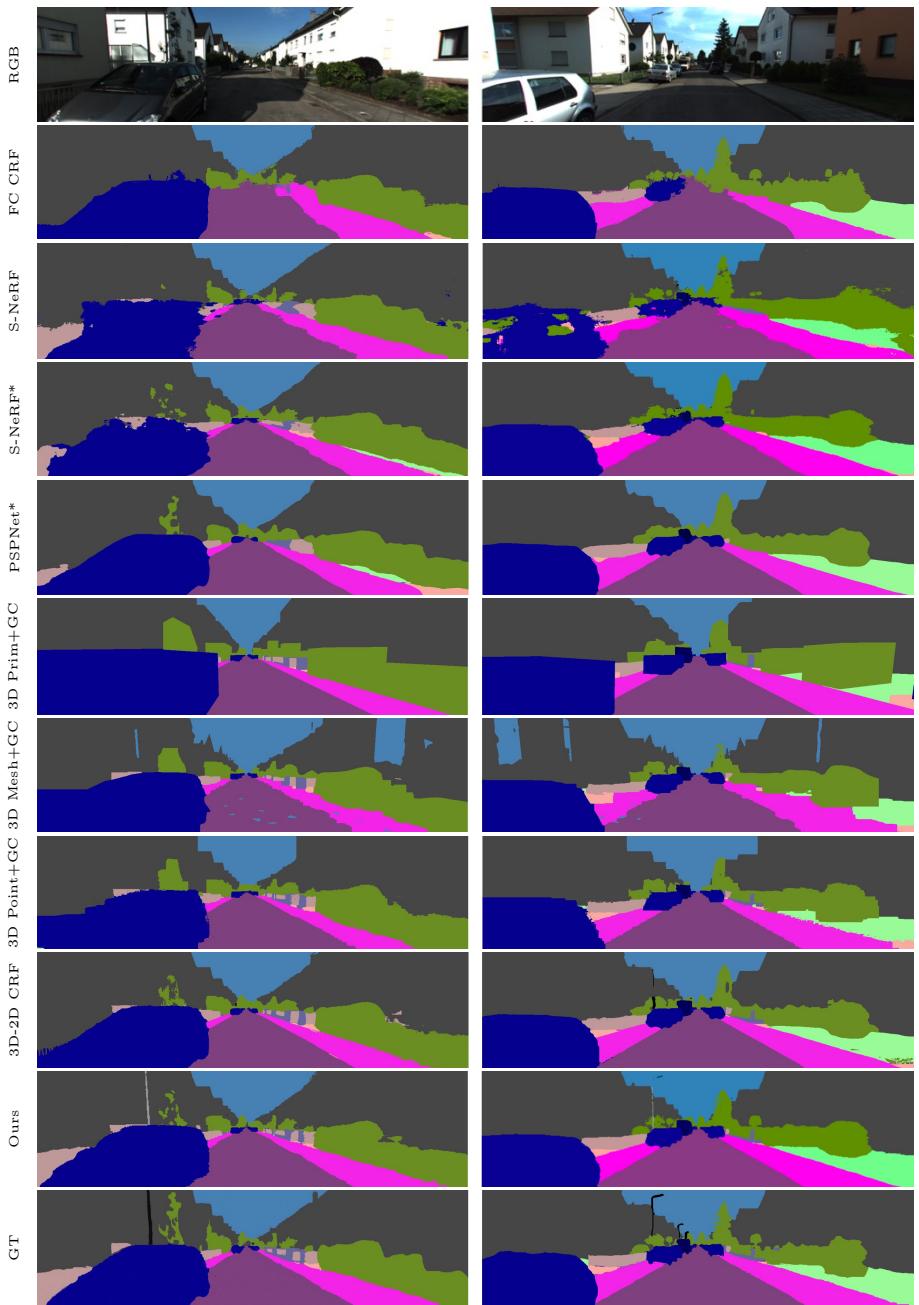


Fig. 11: Qualitative Comparison of Semantic Label Transfer on frames with manually labeled ground truth.



Fig. 12: **Qualitative Comparison of Semantic Label Transfer** on frames without manually labeled ground truth. Each group shows the prediction of 3D-2D CRF [33] (top) and ours (bottom).



Fig. 13: **Qualitative Comparison of Panoptic Label Transfer** on frames without manually labeled ground truth. Each group shows the prediction of 3D-2D CRF [33] (top) and ours (bottom). The colors of the instances do not match due to missing 2D ground truth.



Fig. 14: **Qualitative Results for Stereo Label Transfer.** Top: Blended semantic results of left views. Bottom: Blended semantic results of right views.



Fig. 15: **Qualitative Results for Novel View Transfer.** Top: GT RGB images. Middle: Rendered RGB images. Bottom: Rendered semantic maps.

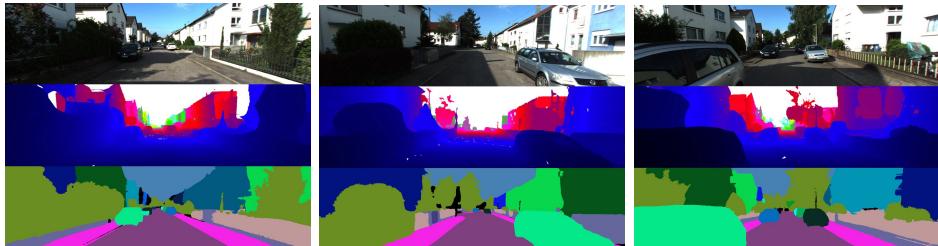


Fig. 16: **Qualitative Results of 3D-2D CRF.** Top: Input RGB images. Middle: 3D-2D CRF mesh depth. Bottom: Panoptic label transfer results of the 3D-2D CRF method.



Fig. 17: **Failure Cases.** Although our semantic map (left) is correct, the boundary of two adjacent buildings is not well segmented in the panoptic map (right).