PanopticNeRF-360: Panoramic 3D-to-2D Label Transfer in Urban Scenes

Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Xiaowei Zhou, Andreas Geiger, Yiyi Liao ✉

Abstract—Training perception systems for self-driving cars requires substantial 2D annotations that are labor-intensive to manual label. While existing datasets provide rich annotations on pre-recorded sequences, they fall short in labeling rarely encountered viewpoints, potentially hampering the generalization ability for perception models. In this paper, we present PanopticNeRF-360, a novel approach that combines coarse 3D annotations with noisy 2D semantic cues to generate high-quality panoptic labels and images from any viewpoint. Our key insight lies in exploiting the complementarity of 3D and 2D priors to mutually enhance geometry and semantics. Specifically, we propose to leverage coarse 3D bounding primitives and noisy 2D semantic and instance predictions to guide geometry optimization, by encouraging predicted labels to match panoptic pseudo ground truth. Simultaneously, the improved geometry assists in filtering 3D&2D annotation noise by fusing semantics in 3D space via a learned semantic field. To further enhance appearance, we combine MLP and hash grids to yield hybrid scene features, striking a balance between high-frequency appearance and contiguous semantics. Our experiments demonstrate PanopticNeRF-360's state-of-the-art performance over label transfer methods on the challenging urban scenes of the KITTI-360 dataset. Moreover, PanopticNeRF-360 enables omnidirectional rendering of high-fidelity, multi-view and spatiotemporally consistent appearance, semantic and instance labels.

Index Terms—3D-to-2D Label Transfer, Panoptic Labeling, Semantic Labeling, Neural Rendering, Urban Scene Understanding

1 Introduction

In autonomous driving, large-scale semantic instance annotations of real-world scenes are foundational for bootstrapping perception models [17], [27], [72]. Manually annotation of semantic and instance masks at the pixel level is recognized to be labor-intensive and costly. For example, a single street scene image demands approximately 1.5 hours for comprehensive instance annotation [15]. Furthermore, going beyond the traditional task of annotating pre-recorded data from fixed viewpoints, offering RGB images and annotations from various new viewpoints carries significant value. This augmentation of viewpoint diversity holds the potential to enhance the generalization capabilities of perception models. Therefore, in this paper, we aim to devise a semi-automated framework that involves relatively low-cost human labor for generating high-fidelity labels and RGB images from both pre-recorded and novel viewpoints.

3D-to-2D label transfer has great potential in this area [32], [43], [77]. By annotating coarse bounding primitives in 3D space and propagating these manually annotated coarse 3D primitives to dense and multi-view consistent 2D semantic and instance annotations, it significantly reduces the labeling time to 0.75 minutes per image [43], yielding a ∼120x speedup compared to 2D per-pixel labeling [15]. Existing methods [43], [77] automatically infer dense 2D labels leveraging 3D annotations, 2D pre-trained models (e.g., noisy 2D semantic predictions) and 2D image cues via conditional random fields (CRF). This CRF-based methodology necessitates intermediary 3D reconstructions for the projection of non-occluded 3D points. A limitation lies in the fact that the 3D reconstruction cannot be collaboratively and jointly optimized within the CRF framework, and as a consequence, any inaccuracies in the reconstruction phase propagate to the label transfer outputs. Besides, these methods are incapable of transferring labels to novel viewpoints.

In this paper, we introduce PanopticNeRF-360, a novel method that utilizes a 360° Neural Radiance Field (NeRF) [52] to estimate geometry and semantics in a joint and differentiable manner. PanopticNeRF-360 takes as input a set of sparse forward-facing stereo images and two side-facing fisheye images, as well as coarse 3D annotations and noisy 2D semantic predictions. By inferring semantic and instance labels in 3D space, the model renders dense 2D semantic and instance labels, i.e., panoptic segmentation labels [36], at novel viewpoints (see Fig. 1). Note that our model enables rendering images from diverse viewpoints and even panoramic images, by incorporating the fisheye images of a large field of view. However, obtaining accurate geometry and semantics is non-trivial in urban scenes. In driving scenarios, where we have sparse input views with frequent over-exposure (particularly common for fisheye views due to directly facing of the sun), reconstructing high-quality geometry using NeRF is challenging. Moreover, inferring precise semantics in 3D space is also difficult given imprecise geometry and coarse 3D annotations with many overlapping regions. Errors in geometric reconstruction (e.g., 3D floaters enclosed by a 3D bounding primitive) and label ambiguity of the 3D coarse annotations (e.g., overlapping regions of car and road) can negatively impact the label transfer step, leading to incorrect 2D semantic and instance labels. While several prior works have modeled 3D semantic fields [73], [78], [83], we make a surprising key ob-

- *X. Fu, S. Zhang, T. Chen, Y. Lu, X. Zhou and Y. Liao are affiliated with Zhejiang University, China.*
- *A. Geiger is affiliated with the Autonomous Vision Group (AVG) at the University of Tübingen and Tübingen AI Center, Germany.*
- ✉*Corresponding author.*

- *Project page: https://fuxiao0719.github.io/projects/panopticnerf360/*
- *Code and data: https://github.com/fuxiao0719/panopticnerf*

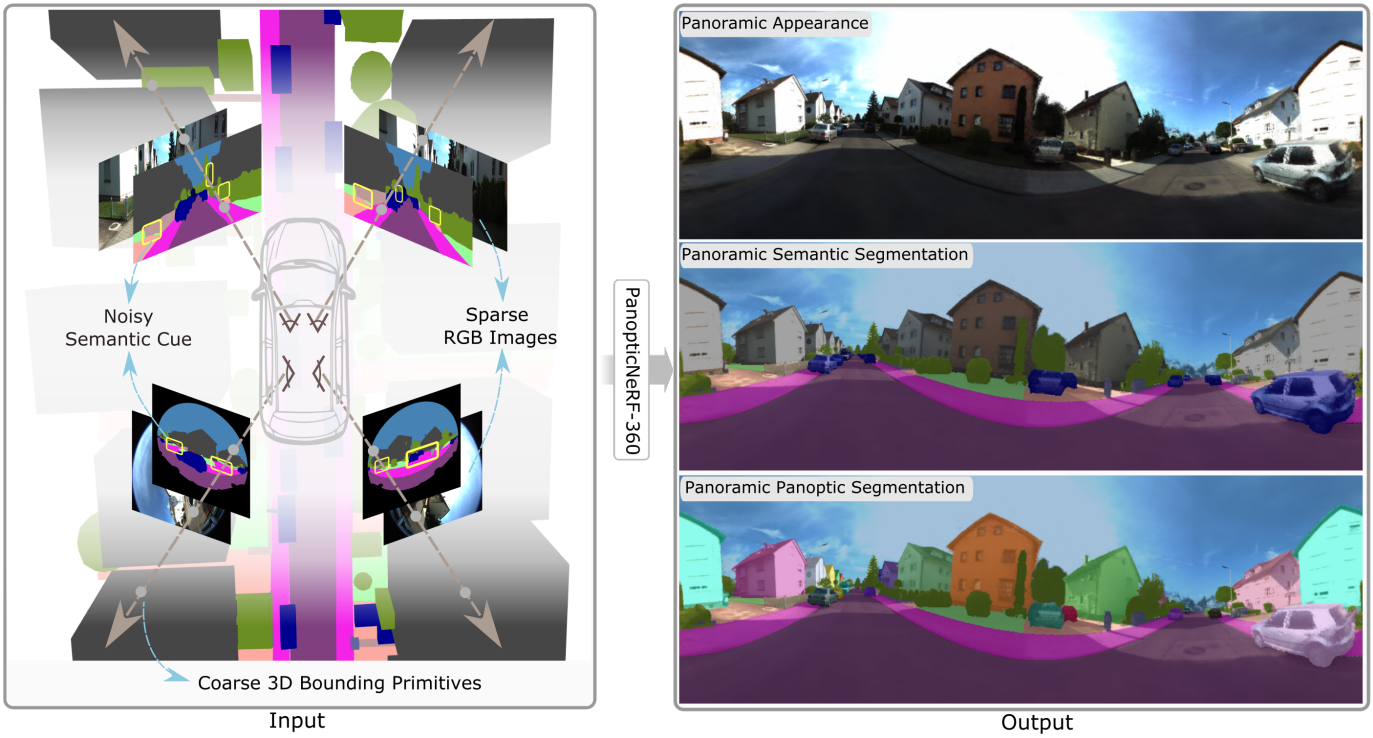


Fig. 1: **PanopticNeRF-360** takes as input a set of sparse forward-facing stereo images and two side-facing fisheye images, coarse 3D bounding primitives and noisy 2D semantic predictions (yellow boxes highlight inaccurate predictions). By inferring in 3D space, our model generates multi-view consistent semantic and instance labels in 2D image space via volume rendering. Our holistic formulation allows for rendering panoramic appearance and label maps (output).

servation: As illustrated in Fig. 2, naïve joint optimization of geometry and semantics does not necessarily yield mutual improvements in our setting.

To tackle these challenges, we aim for *mutual enhancement of geometry and semantics* by introducing two auxiliary parameter-free semantic and instance fields to regularize density estimation, followed by joint optimization of geometry and semantics to resolve label ambiguity, as illustrated in Fig. 3. The results of our model are shown at the bottom of Fig. 2, demonstrating higher accuracy and consistency than the baseline. Firstly, we propose label-guided geometry optimization, utilizing coarse 3D bounding primitives and 2D pseudo labels to guide geometry optimization. In particular, we render panoptic labels utilizing the deterministic 3D bounding primitives (which we refer to as a fixed semantic/instance field as it provides non-trainable semantic and instance logits) and learned density fields, and encourage the rendered labels to match a 2D panoptic pseudo ground truth. As evidenced by our experiments, this leads to significant improvements of the density field despite the pseudo ground truth being noisy. This pseudo ground truth consists of *semantic* predictions from pre-trained 2D segmentation models and *instance* labels derived from a simple geometric prior[1]. Secondly, we propose a joint geometry and semantic optimization strategy to improve semantics. Specifically, conditioned on the improved geometry, we learn to predict semantic categorical 3D logits to match semantic points in 3D bounding primitives and its corresponding 2D distribution via volume rendering to match the 2D noisy predictions. This enables resolving label ambiguity of the 3D bounding primitives between different semantic classes and substantially mitigates noise in the 2D predictions. Note that depsite that 2D semantic predictions of fisheye images may be of low quality due to the lack of training data, our method resolves the noise thanks to the holistic design of utilizing the weak 3D labels and the 2D noisy predictions, enabling rendering improved panoptic labels at arbitrary viewpoints. Despite rendering satisfactory panoptic labels, attaining high-quality appearance remains a problem, as the semantic label is contiguous across the same object/stuff while the corresponding appearance can contain high-frequency details. To tackle this problem, PanopticNeRF-360 combines features of a deep MLP and multi-resolution hash grids to model semantics and appearance. This allows us to leverage the smooth inductive bias of MLPs for semantics and the expressive local hash features for appearance, enabling rendering panoptic labels and photorealistic RGB images from arbitrary novel viewpoints.

We conduct extensive experiments on the KITTI-360 dataset and showcase that our generalization ability on the Waymo dataset. As evidenced by our experimental results, PanopticNeRF-360 showcases state-of-the-art performance and outperforms existing 3D-to-2D and 2D-to-2D label transfer methods and demonstrates a promising path toward the efficient generation of densely annotated datasets that are pivotal for the advancement of autonomous driv-

1. For label-guided geometry optimization, we only consider instance labels of buildings, as buildings are the most frequently connected class that may yield wrong geometry at the adjacent boundary.

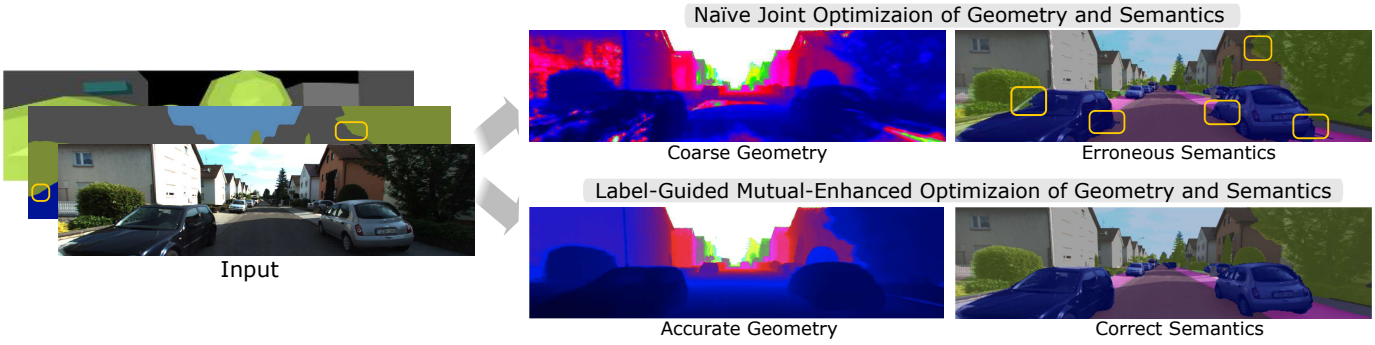


Fig. 2: **Comparison of Different Geometry and Semantic Optimization Methods.** Compared to the naïve joint optimization of geometry and semantics (top), ours leads to mutual enhancement of geometry and semantics (bottom).

ing systems. In summary, our contributions are as follows: (1) We present the first model that tackles 3D-to-2D label transfer via neural rendering. Towards this goal, we unify easy-to-obtain 3D bounding primitives and noisy 2D semantic predictions in a single model, yielding high-quality panoptic labels and high-frequency imagery. (2) We propose novel optimization strategies to enable mutual improvement of geometry and semantics. Our label-guided geometry optimization shows that the underlying geometry can be effectively improved by leveraging noisy 3D and 2D panoptic labels. (3) PanopticNeRF-360 achieves state-of-the-art performance compared to existing label transfer methods in terms of both semantic and instance predictions in challenging urban scenes. Further, PanopticNeRF-360 enables omnidirectional rendering of high-fidelity appearance and spatio-temporally consistent panoptic labels, providing labeled data from novel viewpoints to potentially enhance the generalization ability of perception models.

Relation to PanopticNeRF [19]: This paper is an extension of our earlier conference paper PanopticNeRF [19]. We improve upon [19] in several aspects: We (1) extend perspective label transfer to omnidirectional 360° label transfer, (2) incorporate instance labels into label-guided geometry optimization, thus achieving panoptic label-guided geometry optimization, (3) achieve higher quality semantics (0.8 on forward-facing mIoU) and instances (2.3 on forward-facing PQ), (4) improve scene features from pure MLPs to a hybrid of MLPs and grids for improved appearance (∼4dB) in less training time (∼2.5x speedup), and (5) extensively enrich the experimental section by including new label transfer results on fisheye views and comparison to more recent baselines.

2 Related Work

Urban Scene Segmentation: Semantic instance segmentation is a critical task for autonomous vehicles [58], [84]. Learning-based algorithms have achieved compelling performance [12], [42], [82], but rely on large-scale training data. Unfortunately, annotating images at the pixel level is extremely time-consuming and labor-intensive, especially for instance-level annotation. SAM [37] demonstrates extraordinary generalizable ability in zero-shot object boundary segmentation, but it lacks an understanding of high-level semantics. While most urban datasets provide labels in 2D image space [4], [15], [34], [56], [75], autonomous vehicles are usually equipped with 3D sensors [6], [21], [22], [32], [43], allowing to exploit 3D information for labeling. KITTI-360 [43] demonstrates that annotating the scene in 3D can significantly reduce 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: Several prior works have investigated how to label individual frames more efficiently [1], [2], [8], [25], [45], [46]. In this paper we focus on efficient labeling of video sequences. Existing works in this area fall 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 [20], [29], [59], [60], [61] or transferring labels from frontal views to Bird's-Eye-View (BEV) maps [23], [24], whereas 3D-to-2D methods exploit additional information in 3D for efficient labeling [5], [32], [49], [50], [55], [76], [85]. To obtain dense labels in 2D image space, one line of 3D-to-2D methods requires tedious preprocessing in the 3D space [31]. Another line of methods instead projects coarse 3D labels to the 2D image space and manually refines the labels in 2D [32], [70]. The state-of-the-art works [43], [77] perform per-frame inference jointly over the 3D point clouds and 2D pixels using a non-local multi-field CRF model, avoiding manual pre- or post-processing. However, these methods require reconstructing a 3D mesh to project 3D point clouds to 2D. The mesh reconstruction is not jointly optimized in the CRF model as it is treated as a pre-processing step, leading to inaccurate reconstruction that hinders label transfer performance. In contrast, PanopticNeRF-360 provides a novel end-to-end method for 3D-to-2D label transfer where geometry and semantics are jointly optimized.

Semantic-informed NeRFs: Recently, NeRF [52] emerged as a novel powerful representation for novel view synthesis. Semantic NeRF [83] initially augments NeRF [52] with a semantic branch to encode multi-view consistent semantics from noisy 2D semantic segmentations. However, Semantic NeRF takes as input ground truth 2D labels or synthetic noisy labels (test denoising ability), and faces difficulties generating accurate labels given real-world predictions from pre-trained 2D models. Additionally, Semantic NeRF operates in indoor scenes with dense RGB inputs and degen-

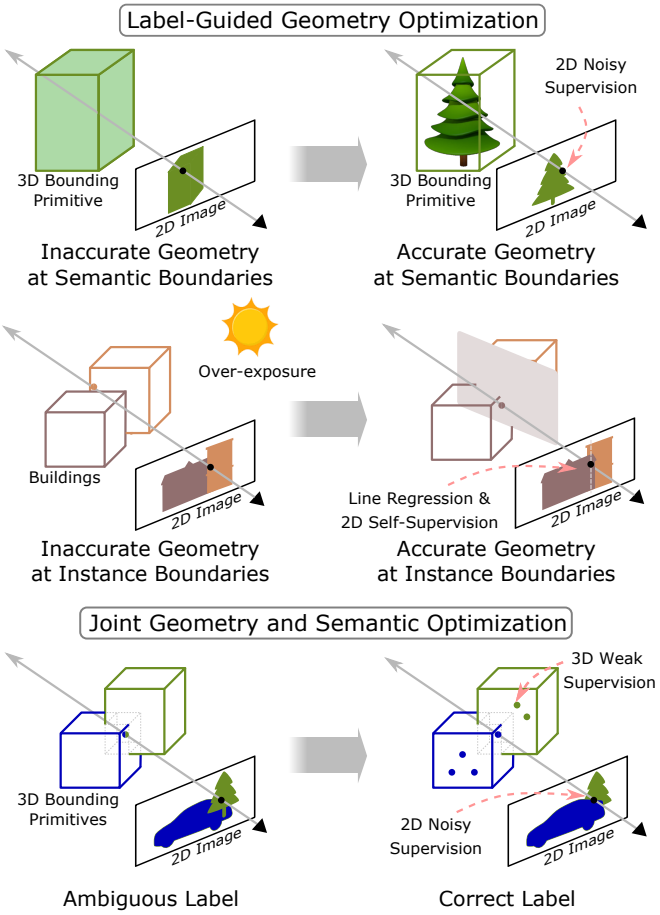


Fig. 3: **Challenges and Solutions.** We propose semantic and instance label-guided geometry optimization to improve the underlying geometry. This allows for rendering accurate semantic and instance labels when coarse 3D bounding primitives correctly enclose the corresponding target without ambiguity. We further resolve label ambiguity at intersection regions of the 3D bounding primitives via joint geometry and semantic optimization.

erates in challenging outdoor driving scenarios with sparse input views [40], as also shown in our experiments. Jacobian NeRF [78] further enhances semantic synergies of correlated entities via contrastive learning on self-supervised visual features for more effective label propagation. In contrast, NeSF [73] emphasizes generalizable semantic field learning via a feed-forward 3D U-Net supervised by 2D GT labels, whereas we concentrate on the 3D-to-2D label transfer task without access to 2D GT. Another line of work distills [38], [68], [69] abstract visual features from zero-shot vision encoders (e.g., [7], [41]) into 3D space for scene editing but does not render precise semantic labels. While semantic-enabled NeRFs [53], [79], [83] are limited to the semantic domain, a natural extension is to explore fine-grained instance information. PNF [40] extracts things with the aid of an off-the-shelf object detector. DM-NeRF [74] requires ground truth instance annotations for 3D geometry decomposition. Panoptic Lifting [65], Nerflets [81], PCFF [14], and Instance NeRF [30] leverage predicted instance/panoptic labels for a compositional panoptic scene representation. SUDS [69] uses geometric clustering to label instances and assign semantic labels based on DINO features. However, this simple clustering framework leads to unsatisfying instance results. While these works mainly focus on scene parsing in close-domain classes (e.g., Cityscapes [15] encompasses only 30 classes), our goal is to transfer 3D annotations of arbitrary classes to 2D image space to foster the development of new datasets, e.g., providing instance labels for buildings that are not available in Cityscapes [15]. Furthermore, we are the first to study 360° outward panoptic urban scene understanding through the powerful lens of neural rendering.

3 Methodology

3.1 Problem Formulation

As shown in Fig. 1, PanopticNeRF-360 aims to transfer coarse 3D bounding primitives to dense panoramic 2D semantic and instance labels, utilizing reconstructed geometry for 3D-2D transfer guidance. We follow the hardware and annotation setup of KITTI-360 [43]: In addition to a pair of perspective stereo images and two-sided fisheye images sparsely collected at urban scenes, we assume a set of coarse 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 "stuff" or "thing" label. For "thing" classes (e.g., "building" and "car"), B_k is additionally associated with a unique instance ID. We further apply a pre-trained semantic segmentation model to the RGB images to obtain 2D semantic predictions for each image. Given this input, our goal is to generate multi-view consistent panoptic labels and high-fidelity appearance for all input frames. Moreover, our method allows for rendering RGB images and panoptic labels from a wide range of novel viewpoints, even including 360° omnidirectional views.

PanopticNeRF-360 provides a novel method for label transfer from 3D to 2D. Fig. 4 provides an overview of our method. We first build a 360° scene representation (Section 3.2) using both perspective and fisheye views, mapping a 3D point $\mathbf{x}$ to a density σ and a color value $\mathbf{c}$ along with two semantic categorical logits $\hat{\mathbf{s}}$ and $\mathbf{s}$ based on our dual semantic fields. Here, $\hat{\mathbf{s}}$ is a deterministic semantic logit derived from a *fixed semantic field* s_β defined by the 3D bounding primitives, and $\mathbf{s}$ is learned semantic logit queried from a *learned semantic field* s_ϕ. Accordingly, for each camera ray, two semantic categorical distributions $\hat{\mathbf{S}}$ and $\mathbf{S}$ in 2D image space are obtained via volume rendering π. Using the labeled 3D bounding primitives, we further define a deterministic *fixed instance field* t_β which divides "thing" classes in the fixed semantic field into distinct instances, allowing for rendering panoptic labels $\hat{\mathbf{T}}$ when combined with the learned semantic field. The fixed semantic field s_β and the fixed instance field t_β together serve to improve the underlying scene geometry (Section 3.3). Furthermore, by using semantic losses in 3D and 2D space (Section 3.4), the learned semantic field s_ϕ results in improved semantics in overlapping regions based on refined underlying geometry.

3.2 360° Scene Representation

Hybrid Scene Feature: We seek to represent the 360° scene with both high-fidelity semantics and appearance. We

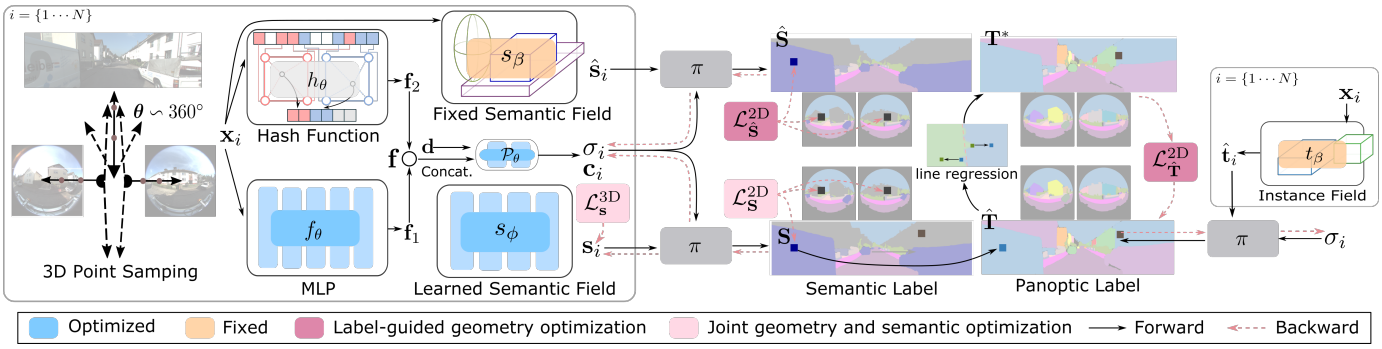


Fig. 4: **Method Overview**. At each 3D location $\mathbf{x}_i$, $i = \{1 \cdots N\}$, we combine scene features obtained from a deep MLP f_θ and multi-resolution hash grids h_θ to jointly model geometry, appearance and semantics. We leverage dual semantic fields to obtain two semantic categorical logits: $\hat{\mathbf{s}}_i$ is obtained from a fixed semantic field s_β determined by the bounding primitives and $\mathbf{s}_i$ is predicted by a learned semantic field s_ϕ. The 3D semantic logits are accumulated along the ray and projected to 2D image space via volume rendering π, resulting in $\hat{\mathbf{S}}$ and $\mathbf{S}$. Our method allows for rendering panoptic labels $\hat{\mathbf{T}}$ by combining the learned semantic field s_ϕ and a fixed instance field t_β determined by the 3D bounding primitives. During training, we leverage label-guided geometry optimization. Here, $\mathcal{L}^{2D}_{\hat{\mathbf{S}}}$ and $\mathcal{L}^{2D}_{\hat{\mathbf{T}}}$ serve as semantic and instance guidance to improve volume density σ_i guided by $\hat{\mathbf{S}}$ and $\mathbf{T}^*$, respectively, where $\mathbf{T}^*$ is obtained by performing line regression on the building class in $\hat{\mathbf{T}}$. Furthermore, we use $\mathcal{L}^{3D}_{\hat{\mathbf{s}}}$ and $\mathcal{L}^{2D}_{\mathbf{S}}$ to train the learned semantic field, yielding joint optimization of geometry and semantics. This allows for resolving label ambiguity at the intersections of the fixed semantic field. The final panoptic segmentation result of our method is denoted by $\hat{\mathbf{T}}$.

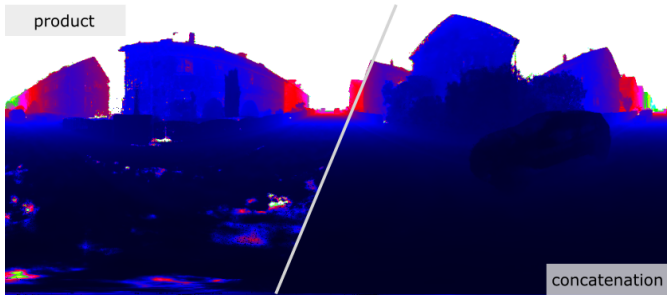


Fig. 5: **Qualitative Geometric Comparison** of feature aggregation operators, "concatenation" and "product" [10]. The latter introduces jagged geometric errors on the road.

empirically observe that the inductive bias of the MLP is a desirable property for smoothing noise in the semantic prediction, yet it is inferior in reconstructing high-fidelity urban scene appearance. On the other hand, existing local feature-based methods like Instant-NGP result in superior appearance in a short time but noisy semantics. Therefore, we fuse features of a deep MLP $f_{\boldsymbol{\theta}}$ [52] and multi-resolution hash grids $h_{\boldsymbol{\theta}}$ [54] to form a joint representation. This simple representation allows us to combine the inductive smoothness bias of the MLP and the expressive representation power of local features, yielding high quality appearance and semantics jointly. Formally, a 3D coordinate $\mathbf{x}$ is mapped to two feature vectors as follows:

$$f_{\boldsymbol{\theta}} : \gamma(\mathbf{x}) \in \mathbb{R}^{L_\mathbf{x}} \mapsto \mathbf{f}_1 \in \mathbb{R}^D, \quad h_{\boldsymbol{\theta}} : \mathbf{x} \in \mathbb{R}^3 \mapsto \mathbf{f}_2 \in \mathbb{R}^D \quad (1)$$

where $\gamma(\cdot)$ denotes the positional encoding, $\mathbf{f}_1$ and $\mathbf{f}_2$ denote MLP-based and grid-based output features respectively. We obtain a unified scene feature $\mathbf{f}$ via concatenation, $\mathbf{f} = [\mathbf{f}_1; \mathbf{f}_2], \mathbf{f} \in \mathbb{R}^{2D}$. DiF [10] introduces the inner product as a general form to unify different basis functions. However, we find that this element-wise "product" results in erroneous geometry, and we instead choose "concatenation" as shown in Fig. 5. We hypothesize this is due to the fact that the MLP and grid branches operate less entangled during the gradient backpropagation, making it simpler for optimization. Along with a viewing direction $\mathbf{d}$ and a frame-based trainable embedding $\mathbf{z}$ for modeling varying exposure, we map $\mathbf{f}$ to a volume density σ and an RGB color value $\mathbf{c}$ with a projection function $\mathcal{P}_\theta$ that consists of shallow MLPs:

$$\mathcal{P}_\theta : (\mathbf{f} \in \mathbb{R}^{2D}, \mathbf{z} \in \mathbb{R}^n, \gamma(\mathbf{d}) \in \mathbb{R}^{L_\mathbf{d}}) \mapsto (\sigma \in \mathbb{R}^+, \mathbf{c} \in \mathbb{R}^3) \quad (2)$$

Let $\mathbf{r}(k) = \mathbf{o} + k\mathbf{d}$ denote a camera ray. The color at the corresponding pixel can be obtained by volume rendering [52]

$$\mathbf{C}(\mathbf{r}) = \sum_{i=1}^{N} \mathcal{T}_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i \ , \ \mathcal{T}_i = \exp\left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right) \quad (3)$$

where σ_i and $\mathbf{c}_i$ are density and color value at point i sampled along the ray, $\mathcal{T}_i$ denotes the transmittance at the sample point, and $\delta_j = k_{i+1} - k_i$ is the distance between adjacent samples. Let π denote the volume rendering process of a ray $\mathbf{r}$, Eq. 3 can be rewritten as $\mathbf{C}(\mathbf{r}|\boldsymbol{\theta}) = \pi(\mathbf{c})$.

While the appearance embedding $\mathbf{z}$ is shared across all points of the same frame during training, we use a different $\tilde{\mathbf{z}}$ for each ray when rendering labels from omnidirectional novel viewpoints. This allows us to find the optimal $\mathbf{z}$ for each ray and avoids stitching artifacts. Specifically, $\tilde{\mathbf{z}}$ is determined based on its relative viewing direction $\mathbf{d}$ with respect to the left and right fisheye cameras :

$$\tilde{\mathbf{z}} = \alpha * \mathbf{z}_l + (1 - \alpha) * \mathbf{z}_r \quad (4)$$

where $\alpha = \cos^{-1}(\mathbf{d} \cdot \mathbf{d}_r) \,/\, (\cos^{-1}(\mathbf{d} \cdot \mathbf{d}_l) + \cos^{-1}(\mathbf{d} \cdot \mathbf{d}_r))$ with $\mathbf{d}_l$ and $\mathbf{d}_r$ denoting unit viewing directions in the center of the left & right fisheye cameras, respectively, and $\mathbf{z}_l$ and $\mathbf{z}_r$ are corresponding latent codes.

Dual Semantic Fields: To jointly optimize the underlying geometry and semantics for mutual improvement, we define dual semantic fields, one is determined by the 3D bounding primitives β and the other is learned by a semantic head $\boldsymbol{\phi}$

$$s_\beta : \mathbf{x} \in \mathbb{R}^3 \mapsto \hat{\mathbf{s}} \in \mathbb{R}^{M_s}, \qquad s_{\boldsymbol{\phi}} : \mathbf{f} \in \mathbb{R}^{2D} \mapsto \mathbf{s} \in \mathbb{R}^{M_s} \tag{5}$$

where M_s denotes the number of semantic classes. In combination with the volume density, two semantic distributions $\hat{\mathbf{S}}(\mathbf{r})$ and $\mathbf{S}(\mathbf{r})$ can be obtained at each camera ray $\mathbf{r}$ via accumulating the pre-softmax logits $\hat{\mathbf{s}}(\mathbf{x})$ and $\mathbf{s}(\mathbf{x})$ through the volume rendering operation π:

$$\hat{\mathbf{S}}(\mathbf{r}|\boldsymbol{\theta}, \beta) = \pi(\hat{\mathbf{s}}), \quad \mathbf{S}(\mathbf{r}|\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\mathbf{s}) \tag{6}$$

Note that both $\hat{\mathbf{S}}(\mathbf{r})$ and $\mathbf{S}(\mathbf{r})$ are multi-class normalized distributions through an extra softmax layer. We apply losses to both semantic distributions for training. During inference, the semantic label is determined as the class of maximum probability in $\hat{\mathbf{S}}(\mathbf{r})$ or $\mathbf{S}(\mathbf{r})$.

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 logit vector 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 all enclosed categories and 0 to the others. As explained in Section 3.4, the fixed semantic field s_β is able to improve geometry but cannot resolve label ambiguity in overlapping regions.

Learned Semantic Field: We add a semantic head parameterized by $\boldsymbol{\phi}$ to learn the semantic logits $\mathbf{s}(\mathbf{x})$. Following [65], we choose to perform softmax on 2D class logits after alpha compositing to obtain the class distribution $\mathbf{S}(\mathbf{r})$. We empirically observe that this leads to better performance than performing softmax on all 3D logits.

Fixed Instance Field: Based on our learned semantic field $s_{\boldsymbol{\phi}}$ and the 3D bounding primitives β with instance IDs, we can easily render a panoptic segmentation mask. 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 $\hat{\mathbf{T}}(\mathbf{r})$ based on the bounding primitives β to replace $\mathbf{S}$ with $\hat{\mathbf{T}}$. Our instance field is defined as follow

$$t_\beta : \mathbf{x} \in \mathbb{R}^3 \mapsto \hat{\mathbf{t}} \in \mathbb{R}^{M_t} \tag{7}$$

where M_t is the number of the things in the scene and $\hat{\mathbf{t}}$ denotes categorical logits indicating which thing it belongs to. Here, $\hat{\mathbf{t}}$ 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. In case $\mathbf{x}$ is enclosed by multiple bounding primitives of different things, equal probabilities are assigned to each of them. 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 $\hat{\mathbf{T}}$.

We observe that optimizing an additional learned instance field is not required. As shown in Fig. 6, overlap often occurs at the intersection regions of stuff and thing regions, and the bounding primitives of things rarely overlap with each other (the number of overlapping regions accounts for only 1.5% of the total number of pixels and their combined volume is only 1.6%). Thus, the deterministic instance field can lead to reliable performance when the underlying geometry is correctly estimated.

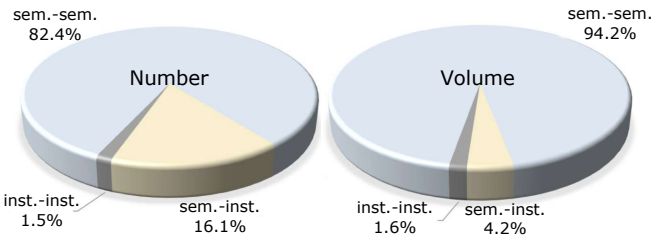


Fig. 6: **Cross-semantics Bounding Primitve Intersection.** We calculate the number of intersected bounding primitives (left) and their corresponding intersected volume (right) in the entire driving sequence. Here, 'sem.-sem.' denotes semantic-semantic intersection, 'sem.-inst.' denotes semantic-instance intersection, and 'inst.-inst.' denotes instance-instance intersection, respectively.

3.3 Label-Guided Geometry Optimization

In the driving scenario considered in our setting, the RGB images are sparsely captured with many textureless and overexposed regions. We observe that a vanilla NeRF model fails to recover reliable geometry in this setting. Therefore, we propose to leverage a fixed semantic field and a fixed instance field to guide the optimization of scene geometry.

Semantic Label-Guided Geometry Optimization: We find that leveraging noisy 2D semantic predictions as pseudo ground truth can substantially boost density prediction when applied to the fixed semantic fields s_β

$$\mathcal{L}^{\text{2D}}_{\hat{\mathbf{S}}}(\boldsymbol{\theta}, \beta) = -\frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \sum_{k=1}^{M_s} \mathbf{S}^*_k(\mathbf{r}) \log \hat{\mathbf{S}}_k(\mathbf{r}) \tag{8}$$

where $\hat{\mathbf{S}}_k(\mathbf{r})$ denotes the probability of the camera ray $\mathbf{r}$ belonging to the class k, and $\mathbf{S}^*_k(\mathbf{r})$ denotes the corresponding 2D pseudo ground truth. As illustrated in Fig. 7, the key to improving density is to directly apply the semantic loss to the fixed semantic field s_β, where $\mathcal{L}^{\text{2D}}_{\hat{\mathbf{S}}}$ can only be minimized by updating the density σ. Fig. 7 shows that a correct $\mathbf{S}^*$ increases the volume density of 3D points inside the respective bounding primitive and suppresses the density of others. When $\mathbf{S}^*$ is wrong, the negative impact can be mitigated: 1) If $\mathbf{S}^*$ does not match any bounding primitive along the ray, it has no impact on the radiance field f_θ. 2) If $\mathbf{S}^*$ exists in one of the bounding primitives along the ray, this indicates that $\mathbf{S}^*$ corresponds to an occluding/occluded bounding primitive with the wrong depth. To compensate for this, we introduce a weak depth loss $\mathcal{L}_d$ based on stereo matching to alleviate the misguidance of $\mathcal{L}^{\text{2D}}_{\hat{\mathbf{S}}}$. Although $\mathcal{L}_d$ improves the overall geometry as shown in our ablation study, it fails to produce accurate object boundaries when used alone (see supplementary). In contrast, adding our semantic label-guided geometry optimization yields more accurate density estimation as pre-trained segmentation models usually perform well on frequently occurring classes, e.g., cars and roads.

Instance Label-Guided Geometry Finetuning: We further develop a simple but effective approach to improve the

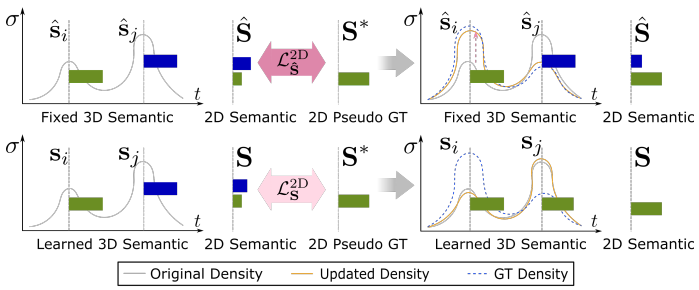


Fig. 7: **Semantic Label Guided Geometry Optimization**. The top row illustrates a single ray of the fixed semantic field s_β, where $\mathcal{L}^{2D}_{\hat{\mathbf{S}}}$ can only update the underlying geometry as the semantic logits $\hat{\mathbf{s}}$ is fixed. The second row shows a single ray of the learned semantic field s_ϕ. In this case, the network can "cheat" by adjusting the semantic prediction $\mathbf{s}$ to satisfy $\mathcal{L}^{2D}_{\mathbf{S}}$ instead of updating the density σ.

geometry of adjacent instances. We observe that the class building is the only "thing" class where two instances of the same semantic class are frequently spatially adjacent, which implies that their boundaries remain unchanged in the semantic label-guided geometry optimization. Further, the geometry of buildings is often negatively impacted by overexposure and low texturedness. Therefore, we propose to use pseudo-GT derived from a simple geometric prior to guide the density field optimization, assuming that neighboring buildings have a line-shaped boundary. Here, we do not use pre-trained segmentation models for two reasons: 1) existing pre-trained instance segmentation models are only applicable to a set of known classes, thus are not suited for unseen thing classes, e.g., buildings; 2) open-set segmentation models such as SAM [37] lead to unsatisfactory boundaries in our challenging setting (see Supplementary).

Let us consider a simple case where two neighboring buildings have a 2D boundary l. As buildings usually have cuboid-like shapes, l can be approximated by *a straight line*. However, the rendered boundary is typically noisy and non-straight when the density field predicts wrong geometry. Thus, we construct pseudo-GTs by applying line regression to rendered 2D boundaries and utilize the pseudo-GTs to optimize the density field. While the pseudo-GTs independently generated across different views may not be multi-view consistent, we fuse them in 3D space to render consistent labels. Specifically, we first render an initial instance boundary based on the fixed instance field and the learned imprecise density. As shown in Fig. 17 (b), the naive boundaries can be jagged. Given the initial noisy boundary, we sample a set of points on the boundary of two adjacent buildings, ignoring the upper and lower regions which may contain unexpected irregular objects (e.g., eave, fence, vegetation) that violate the *line* prior. Next, we apply linear regression on this set of sampled points to obtain a pseudo-GT of the boundary. Based on the regressed boundary formulation, we remap the instance labels on both sides to obtain a pseudo-GT panoptic label of two neighboring buildings. After refining all boundaries, we obtain pseudo panoptic label $\mathbf{T}^*$:

$$\mathbf{T}^* = f_\chi(\hat{\mathbf{T}}) \tag{9}$$

where $f_\chi(\cdot)$ maps the initial instance label to a refined pseudo instance GT and χ denotes all the fitted lines between the buildings.

Then, we leverage $\mathbf{T}^*$ to guide the underlying scene optimization. Similar to Eq. 8, we apply an instance loss based on the fixed instance field t_β on $\mathbf{T}^*$ to improve boundary geometry:

$$\mathcal{L}^{2D}_{\hat{\mathbf{T}}}(\boldsymbol{\theta}, \beta) = -\frac{1}{|\mathcal{R}_k|} \sum_{\mathbf{r}\in\mathcal{R}_k} \mathbf{T}^*_k(\mathbf{r}) \log \hat{\mathbf{T}}_k(\mathbf{r}) \tag{10}$$

where k denotes the "building" class. The instance loss is weighted by $\lambda_{\hat{\mathbf{T}}}$.

3.4 Joint Geometry and Semantic Optimization

While enabling improved geometry, the 3D label of overlapping regions remains ambiguous in the fixed semantic field. We leverage s_ϕ to address this problem by jointly learning the semantic and the radiance fields. Towards this goal, we apply a modified cross-entropy loss $\mathcal{L}^{2D}_{\mathbf{S}}$ to each camera ray based on the filtered 2D pseudo ground truth, where $\mathbb{1}(\mathbf{r})$ is set to 1 if $\mathbf{S}^*(\mathbf{r})$ matches the semantic class of any bounding primitive along the ray and otherwise 0. Due to the imbalanced categorical distribution nature of pseudo semantic labels, we modify the softmax cross-entropy operator by introducing a weight $w(k)$ for each class k, which is distributed between [0,1] determined on the semantic class frequency [51]. We do not use $w(k)$ in $\mathcal{L}^{2D}$ as we experimentally observe that the category distribution prior does not additionally help to improve geometry. To further suppress noise in the 2D predictions, we add a per-point semantic loss $\mathcal{L}^{3D}_{\mathbf{s}}$ based on the 3D bounding primitives

$$\begin{aligned}
\mathcal{L}^{2D}_{\mathbf{S}}(\boldsymbol{\theta}, \boldsymbol{\phi}) &= -\frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \mathbb{1}(\mathbf{r}) \sum_{k=1}^{M_s} w(k) \mathbf{S}^*_k(\mathbf{r}) \log \mathbf{S}_k(\mathbf{r}) \\
\mathcal{L}^{3D}_{\mathbf{s}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \beta) &= -\frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \sum_{i=1}^{N} \mathbb{1}(\mathbf{x}) \sum_{k=1}^{M_s} \hat{\mathbf{s}}^k_i \log \mathbf{s}^k_i
\end{aligned} \tag{11}$$

where $\mathbb{1}(\mathbf{x})$ is a per-point binary mask. $\mathbb{1}(\mathbf{x})$ is set to 1 if (1) $\mathbf{x}_i$ has a unique 3D semantic label and (2) the density σ is above a threshold σ_{th} to focus on the object surface. As illustrated in Fig. 7, $\mathcal{L}^{2D}_{\mathbf{S}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ does not necessarily improve the underlying geometry as the network can learn a simple shortcut and adjust the semantic head s_ϕ 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 [16], [57].

3.5 Implementation Details

Loss: We train our model in two stages. In the first stage without instance finetuning, the total loss takes the following form

$$\mathcal{L} = \lambda_{\hat{\mathbf{S}}} \mathcal{L}^{2D}_{\hat{\mathbf{S}}} + \lambda_{\mathbf{S}} \mathcal{L}^{2D}_{\mathbf{S}} + \lambda_{\mathbf{s}} \mathcal{L}^{3D}_{\mathbf{s}} + \lambda_{\mathbf{C}} \mathcal{L}_p + \lambda_d \mathcal{L}_d \tag{12}$$

where $\mathcal{L}_p = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \|\mathbf{C}^*(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|^2_2$ and $\mathcal{L}_d = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}\in\mathcal{R}} \|\mathbf{D}^*(\mathbf{r}) - \mathbf{D}(\mathbf{r})\|^2_2$ denote the photometric loss

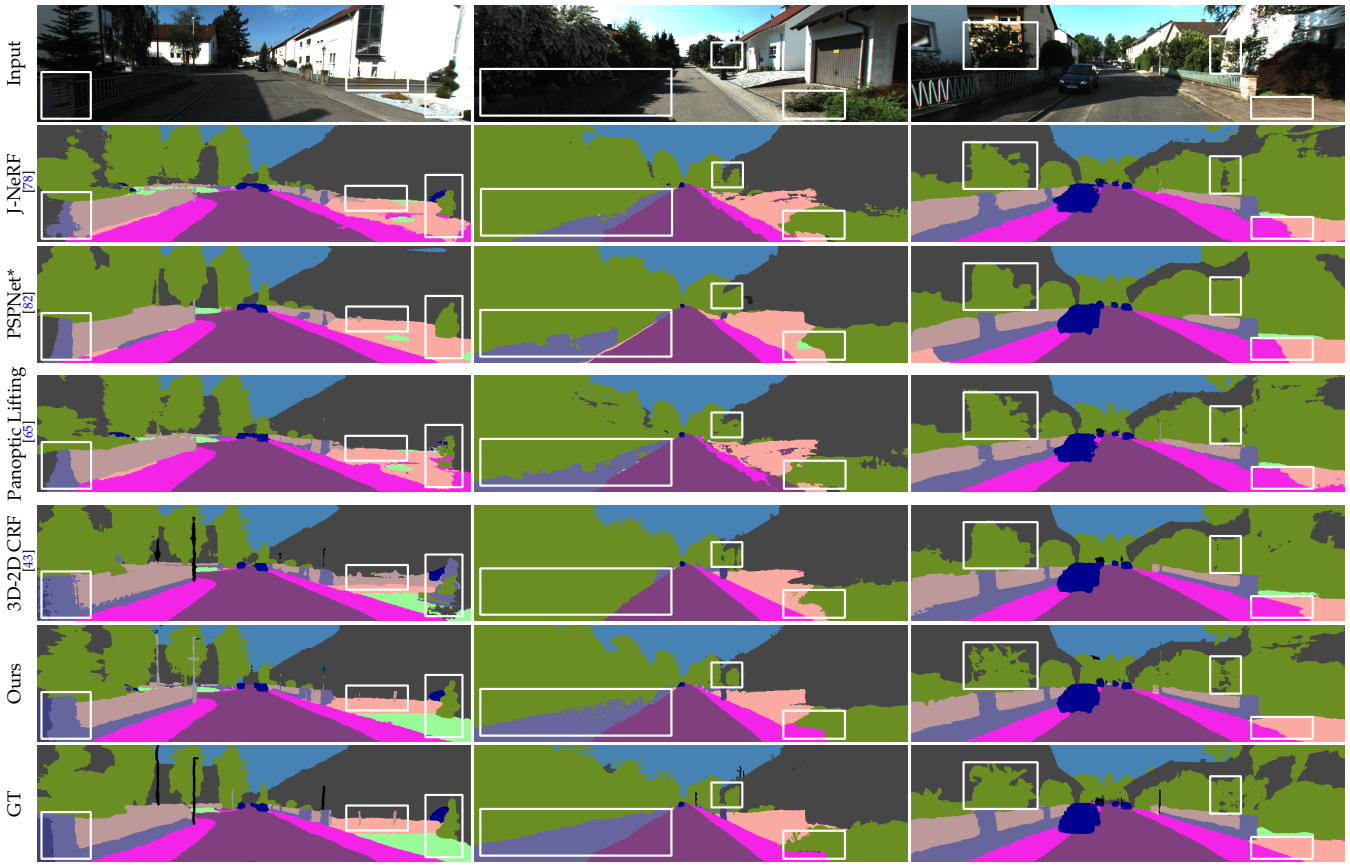


Fig. 8: **Qualitative Comparison of Perspective Semantic Label Transfer.** Our method achieves superior performance in challenging regions compared to the baselines, e.g. in under- or over-exposed regions, by recovering the underlying 3D geometry, see wall (left, middle) and regions where vegetation and building intersect (right).

and the depth loss, respectively. $\lambda_{\hat{\mathbf{S}}}$, $\lambda_{\mathbf{S}}$, $\lambda_{\mathbf{s}}$, $\lambda_{\mathbf{C}}$, and λ_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 about $\mathbf{D}^*$.

We activate instance label-guided geometry optimization in the second stage, leading to the overall loss $\mathcal{L}_f$:

$$\mathcal{L}_f = \mathcal{L} + \lambda_{\hat{\mathbf{T}}} \mathcal{L}^{\text{2D}}_{\hat{\mathbf{T}}} \tag{13}$$

As depicted in Fig. 17, the fine-tuning stage significantly improves the network's ability to produce smoother and more accurate geometry for adjacent buildings.

Training: We optimize one PanopticNeRF-360 model per scene, using a single NVIDIA RTX 3090. For each scene, we set the origin to the center of the scene. We use Adam [35] with a learning rate of 5e-4 to train our models. We set the latent appearance code length to $n = 12$, and loss weights to $\lambda_{\hat{\mathbf{S}}} = 2, \lambda_{\mathbf{S}} = 1, \lambda_{\mathbf{s}} = 1, \lambda_{\hat{\mathbf{T}}} = 2, \lambda_{\mathbf{C}} = 1, \lambda_d = 0.1$, and the density threshold to $\sigma_{th} = 0.1$. We optimize the total loss $\mathcal{L}$ for 30,000 iterations. For the stage of refining instances, we further fine-tune the model for 4,000 iterations.

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. For each ray, we optionally sample a set of points to model the sky after the furthest bounding primitive. More details regarding our sampling strategy can be found in the supplementary. Our sampling strategy allows the network to focus on non-empty regions. As evidenced by our experiments, this is particularly beneficial for unbounded outdoor environments.

4 Experiments

Dataset: We conduct our main experiments on the KITTI-360 [43] dataset, which is collected in suburban areas and provides 3D bounding primitives covering the full scene. Following [43], we evaluate PanopticNeRF-360 on manually annotated frames from 5 static suburbs. We split these 5 suburbs into 10 scenes, each comprising 64 consecutive frames with 4 cameras each with an average travel distance of 0.8m between frames. We leverage these 64 pairs of posed stereo perspective and fisheye images for training. KITTI-360 provides a set of manually labeled frames sampled in equidistant steps of 5 frames on perspective views. Following [43], 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 further improve the quality of the manually labeled ground truth which is inaccurate in ambiguous regions, see supplementary for details. In order to quantitatively evaluate our label transfer performance on side-facing viewpoints, we manually annotate fisheye

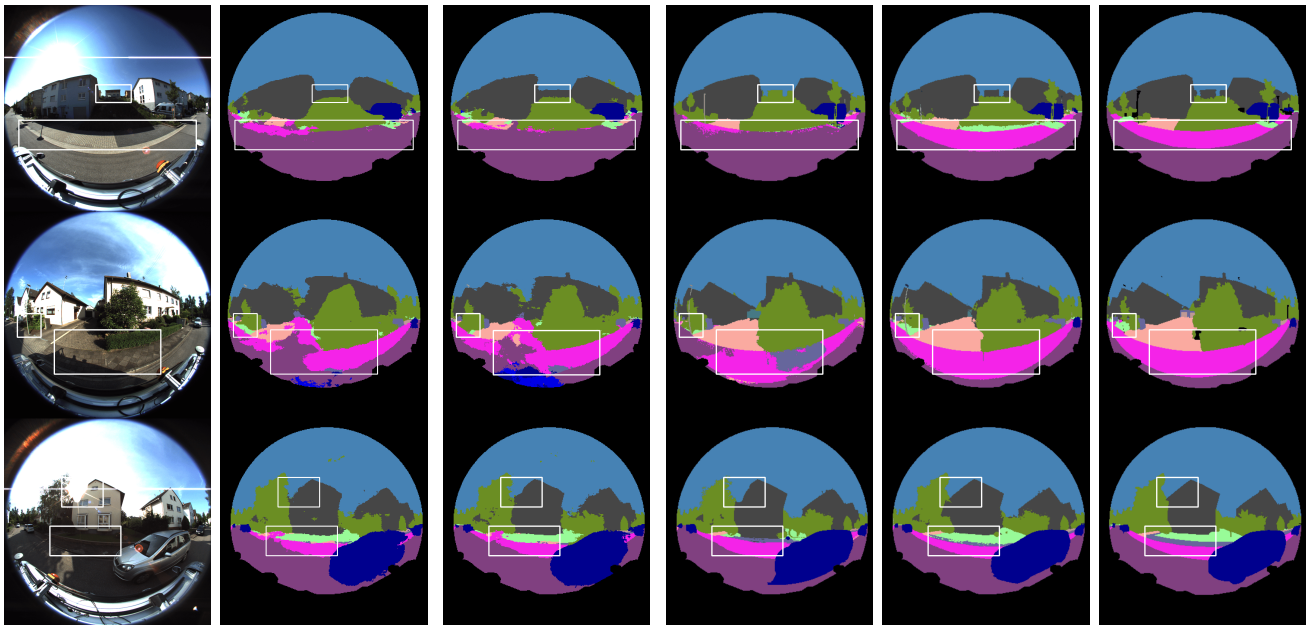


Fig. 9: **Qualitative Comparison of Fisheye Semantic Label Transfer.** Our method consistently achieves the best results. In the third row, ours can generate robust construction semantics under unexpected over-exposure, while the others can not.

Method	Road	Park	Sdwlk	Terr	Bldg	Vegt	Car	Trler	Crvn	Gate	Wall	Fence	Box	Sky	mIoU	Acc	MC
Forward-facing																	
FC CRF + Manual GT [39]	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 + Manual GT [83]	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
Pseudo GT (PSPNet*) [82]	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
S-NeRF + Pseudo GT [83]	94.5	52.7	78.0	64.8	88.5	81.7	89.0	43.9	81.1	35.2	45.6	57.2	43.8	91.0	67.7	90.4	92.76
J-NeRF + Pseudo GT [78]	94.8	49.7	80.2	65.3	86.2	83.1	91.0	45.4	80.6	32.5	48.1	58.1	52.3	91.2	68.5	90.8	92.88
Panoptic Lifting [65]	94.8	57.6	75.0	58.6	83.1	82.9	89.4	68.7	84.1	38.4	54.4	49.4	48.8	88.6	69.6	90.5	93.12
3D Primitives + 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	**96.51**
3D-2D CRF [43]	95.2	64.2	**83.8**	67.9	90.3	84.2	92.2	93.4	90.8	68.2	64.5	70.0	55.8	**92.8**	79.5	92.8	94.98
Ours	**95.4**	**70.0**	83.6	**70.9**	**91.5**	**85.3**	**94.2**	**95.0**	**94.4**	**69.4**	**65.2**	**72.2**	**68.2**	91.7	**81.9**	**93.4**	94.82
Fisheye																	
Pseudo GT (Tao *et al.*) [66]	84.8	0.0	56.2	57.5	89.5	**84.5**	77.7	0.0	0.0	0.0	39.3	60.0	0.0	98.0	46.3	91.0	-
S-NeRF + Pseudo GT [83]	86.2	18.2	59.4	57.6	88.0	81.0	76.5	14.0	33.8	0.0	42.1	53.7	13.4	97.0	51.5	90.7	-
J-NeRF + Pseudo GT [78]	86.3	17.2	61.2	58.4	87.1	81.9	78.1	17.3	30.6	0.0	44.3	55.3	15.2	96.8	52.1	91.1	-
Panoptic Lifting [65]	88.2	20.9	64.3	61.9	85.0	79.1	80.8	13.1	32.6	0.0	46.4	53.3	15.0	95.0	52.5	91.3	-
3D-2D CRF [43]	87.1	56.2	66,4	39.0	88.5	80.8	86.3	77.1	75.8	44.3	41.8	61.1	0.0	97.7	64.4	92.2	-
Ours	**93.7**	**73.7**	**81.9**	**66.0**	**90.0**	83.6	**88.2**	**84.0**	**89.7**	**51.3**	**61.1**	**66.8**	**40.1**	**98.3**	**76.3**	**95.0**	-

TABLE 1: **Quantitative Comparison of Semantic Label Transfer** on 10 scenes of KITTI-360.

images sampled in equidistant steps of 10 frames and use all of them for evaluation. We additionally showcase the generalization ability of our method on the Waymo dataset.

Baselines: We compare against several competitive baselines in two categories: (1) *2D-to-2D label transfer baselines*, including Fully Connected CRF (*FC CRF*) [39], Semantic NeRF (*S-NeRF*) [83], JacobiNeRF (*J-NeRF*) [78], and Panoptic Lifting [65]. We provide manually annotated 2D frames as input, sparsely sampled at equidistant steps of 10 frames. Note that labeling these 2D frames takes similar or longer compared to annotating 3D bounding primitives [77]. As these 2D annotations are extremely sparse, we further provide the same pseudo-2D labels used by our method to Semantic NeRF, JacobiNeRF and Panoptic Lifting. For Panoptic Lifting, We additionally run Mask2Former [13] to generate instance masks, and then fuse them with 2D semantic maps to obtain the final input panoptic masks. For JacobiNeRF, we follow [78] to extract DINO features from the images for its similarity prior. (2) *3D-to-2D label transfer baselines*, including PSPNet*, 3D Primitives/Meshs/Points+GC [43], and 3D-2D CRF [43]. All these baselines leverage the same 3D bounding primitives to transfer labels to 2D. Here, PSPNet* 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: For pseudo ground truth, we use PSPNet* for perspective views and Tao *et al.* [66] that has a stronger

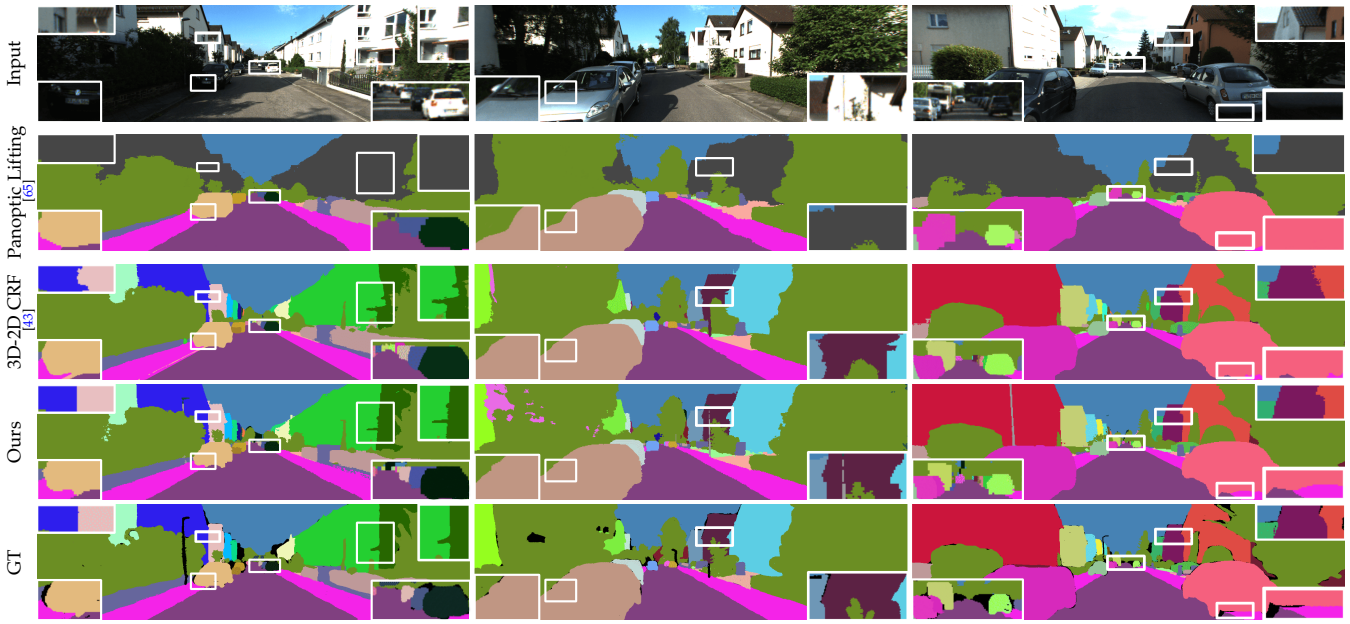


Fig. 10: **Qualitative Comparison of Perspective Panoptic Label Transfer.** Our method is capable of distinguishing instances as it infers in 3D space. In contrast, 3D-2D CRF struggles in far and overexposed regions. Panoptic Lifting falls short in rendering 1) car instances in far regions and 2) building instances.

		Forward-facing	PQ	SQ	RQ	PQ†	Fisheye	PQ	SQ	RQ	PQ†
Building as stuff	Panoptic Lifting [65]	All	52.3	69.4	68.2	54.9	All	42.6	68.5	56.6	44.7
		Things	56.5	76.3	70.4	57.8	Things	37.2	64.7	55.8	38.5
		Stuff	53.9	68.9	75.3	55.7	Stuff	49.3	70.5	61.6	51.6
	Ours*	All	**66.7**	**80.6**	**81.7**	**68.4**	All	**54.2**	**77.6**	**67.6**	**56.5**
		Things	**65.3**	**83.6**	**77.2**	**65.3**	Things	**44.6**	**75.4**	**56.7**	**44.6**
		Stuff	**67.8**	**79.6**	**84.0**	**69.3**	Stuff	**58.9**	**78.2**	**72.8**	**61.9**
Building as thing	3D-2D CRF [43]	All	62.2	79.1	76.9	64.9	All	47.5	72.2	60.1	50.9
		Things	60.7	79.5	75.2	60.7	Things	42.9	68.3	54.3	42.9
		Stuff	63.0	78.9	77.9	67.3	Stuff	50.1	74.6	63.5	55.3
	Ours	All	**66.7**	**80.6**	**81.7**	**69.3**	All	**54.2**	**77.6**	**67.6**	**58.3**
		Things	**64.9**	**83.7**	**76.9**	**64.9**	Things	**46.5**	**77.2**	**59.3**	**46.5**
		Stuff	**67.7**	**78.9**	**84.3**	**71.8**	Stuff	**58.5**	**77.8**	**72.2**	**64.9**

TABLE 2: **Quantitative Comparison of Panoptic Label** over all 10 test scenes on KITTI-360. We consider building as *stuff* when comparing our method (Ours*) to Panoptic Lifting which can not render construction instances, where building is considered as *thing* in our default setting.

generalization ability for fisheye views to supervise our dual semantic fields. This ensures a fair comparison to the 3D-2D CRF, which takes the predictions of PSPNet* and Tao *et al*. as unary terms on perspective views and fisheye views, respectively. Note that PSPNet* is fine-tuned on KITTI-360. In our ablation study we investigate the performance of our method using pre-trained models on Cityscape without any fine-tuning, including PSPNet [82], Deeplab [12] and Tao *et al*. [66]. We also involve SSA [11], a variant of SAM for zero-shot semantic prediction.

Metrics: Following [19], we evaluate semantic labels via 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 pixel pairs with consistent semantic labels over all pairs. For evaluating panoptic segmentation, we report Panoptic Quality (PQ) [36], which can be decomposed into Segmentation Quality (SQ) and Recognition Quality (RQ). We additionally adopt PQ† [62] as PQ over-penalizes errors of stuff classes. To evaluate perspective and fisheye labels in a unified manner, we design $metric^*$ (e.g., mIoU* and PQ*) that reports the average of the perspective metric and fisheye metric. To verify that PanopticNeRF-360 is able to improve the underlying geometry, we further evaluate on perspective views the rendered depth compared to sparse depth maps obtained from LiDAR using Root Mean Squared Error (RMSE) and the ratio of accurate predictions ($\delta_{1.25}$) [3], [18].

4.1 Label Transfer on KITTI-360 Dataset

As most baselines are not designed for panoptic label transfer, we first compare the semantic predictions, and then compare the panoptic predictions to 3D-2D CRF.

Semantic Label Transfer: As shown in Table 1, Fig. 8 and Fig. 9, our method achieves the highest mIoU and Acc quantitatively and qualitatively. Specifically, compared to

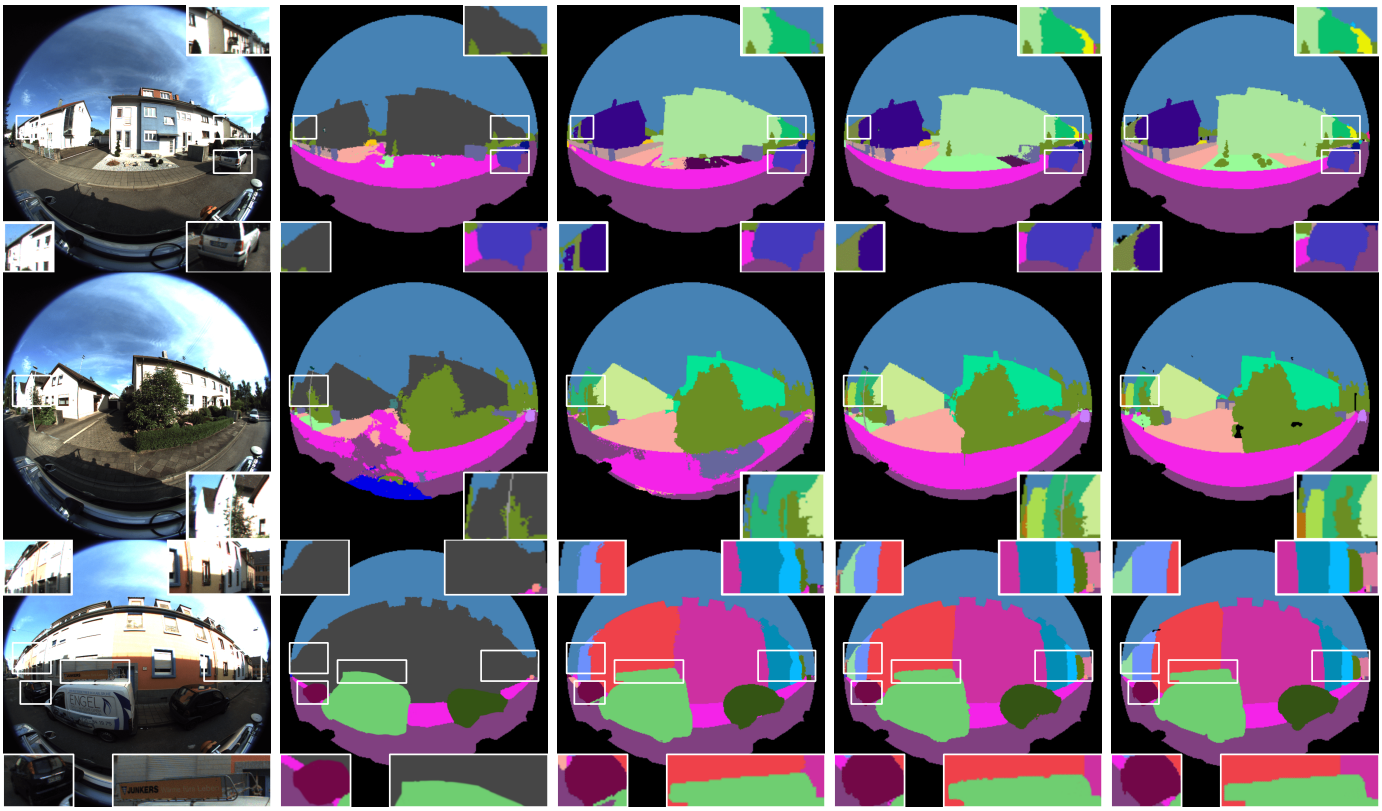


Fig. 11: **Qualitative Comparison of Fisheye Panoptic Label Transfer.** Our method outperforms 3D-2D CRF not only in near-field regions but also in far regions where instances are harder to distinguish. Compared to perspective views, the performance on panoptic labels of 3D-2D CRF and Panoptic Lifting declines more on fisheye views.

3D-2D CRF, we obtain an absolute mIoU improvement of 2.4% on perspective views and 11.9% on fisheye views. The performance gap widens on the fisheye view as the pseudo-GT is less accurate on fisheye imagery. This observation also highlights the value of our method which is able to provide RGB images and semantic labels at novel viewpoints to enhance the generalization ability of 2D perception models. Despite PSPNet* being finetuned on KITTI-360 which reduces the performance gap, our method outperforms PSPNet* by a large margin. Supervised by the extremely sparse manually annotated GT, Semantic NeRF struggles to produce reliable performance. Using pseudo labels of PSPNet* and Tao *et al.*, JacobiNeRF and Panoptic Lifting are capable of denoising and thus improving performance $(67.2\% \rightarrow 68.5\%/69.6\%, 46.3\% \rightarrow 52.1\%/52.5\%)$. However, both JacobiNeRF and Panoptic Lifting are inferior in urban scenarios when the input views are sparse. In terms of perspective MC, ours is comparable to 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, this can be explained as label consistency is evaluated on sparsely projected 3D points which GC takes as input to generate a dense label map.

Panoptic Label Transfer: We compare against Panoptic Lifting and 3D-2D CRF for panoptic label transfer, as the other baselines do not support distinguishing instances. As Panoptic Lifting is limited to segment instance classes known to pre-trained 2D panoptic segmentation models, it is not capable of segmenting the "building" class. Therefore, we consider two settings for quantitatively measuring the panoptic segmentation metrics: "building" as *stuff* when comparing with Panoptic Lifting, and "building" as *thing* when comparing with 3D-2D CRF. Table 2 shows that PanopticNeRF-360 outperforms the baselines in both cases. As shown in Fig. 10, Panoptic Lifting can not reconstruct instances well in far regions where Mask2Former and their instance assignment strategy perform poorly, yielding a larger gap in terms of the panoptic segmentation metrics. In contrast, PanopticNeRF-360 avoids this issue by leveraging weak 3D labels, consistently producing good performance in both near and far regions. Our proposed method also outperforms the 3D-2D CRF in both things and stuff classes, despite that 3D-2D CRF leverages the same weak 3D labels as input. Visual comparisons are shown in Fig. 10 and Fig. 11. As can be seen, our method gracefully handles over- and under-exposure at buildings, which is a challenge for the 3D-2D CRF, as it projects LiDAR points to reconstruct the intermediate meshes whose quality suffers in these scenarios (see Supplementary).

Panoramic Label Synthesis: PanopticNeRF-360 can render RGB images and panoptic labels at omnidirectional novel viewpoints, whereas 3D-2D CRF is not capable of doing so. We therefore show qualitative panoramic results of our methods in Fig. 12.

Panoramic Semantic Maps | Panoramic Panoptic Maps

Fig. 12: **Panoramic Label Rendering.** We enable omnidirectional panoptic label inference in real-world driving scenarios.

4.2 Label Transfer on Waymo Dataset

We additionally conduct experiments on the Waymo dataset [50] to showcase PanopticNeRF-360's generalization ability. We choose two scenes, each consisting of 198 consecutive frames with the front-view camera. During data preparation, we use the KITTI-360 annotation toolkit to label 3D bounding primitives and generate 2D pseudo semantic labels with Mask2Former [13]. This annotation takes around 2 hours for each scene, i.e., the average per-frame labeling time takes only 0.6 minutes. We illustrate the entire set of labeled bounding boxes and camera trajectory in Fig. 13. Ours not only renders high-fidelity appearance (28.28 dB on test views), but also generates high-quality semantic and instance masks on both pre-recorded and novel viewpoints. This demonstrates the applicability of our method.

4.3 Neural Scene Representation

We further compare iNGP [54], Tri-planes [9] (TensoRF-VM), and MLP [52] as scene representation. Towards this goal, we analyze the results over 4 scenes where ambient conditions vary between each other. For a fair comparison, we run each scene for 30,000 iterations. Please refer to the supplementary for implementation details. Table 4 indicates that while the pure MLP demonstrates superior label quality, it falls short in appearance reconstruction. On the other hand, iNGP excels in novel view synthesis (∼3.5dB higher than MLP) and convergence, albeit at the cost of subpar semantic label quality in comparison to the MLP. Our proposed scene feature, a hybrid of MLP and hash grids, achieves a tradeoff between label quality and appearance. Tri-planes fail to represent high-fidelity scene structure both in terms of semantic labels and appearance, especially at boundaries between foreground and background (sky) and in over-exposed regions as grid-based methods are not robust to filter noise and tend to suffer from local minimum. While iNGP is also grid-based, it conditions features on multi-scale levels, thereby potentially aggregating global information, and employs a hash function for non-periodic signal embedding. As depicted in Fig. 15, iNGP contains more noise than ours, and Tri-planes fail to reconstruct the overall label, especially at buildings over-exposed by sunlight.

4.4 Synthesized Labels for Perception Models

To validate that the synthesized novel view labels are important for autonomous driving and robotics applications, we

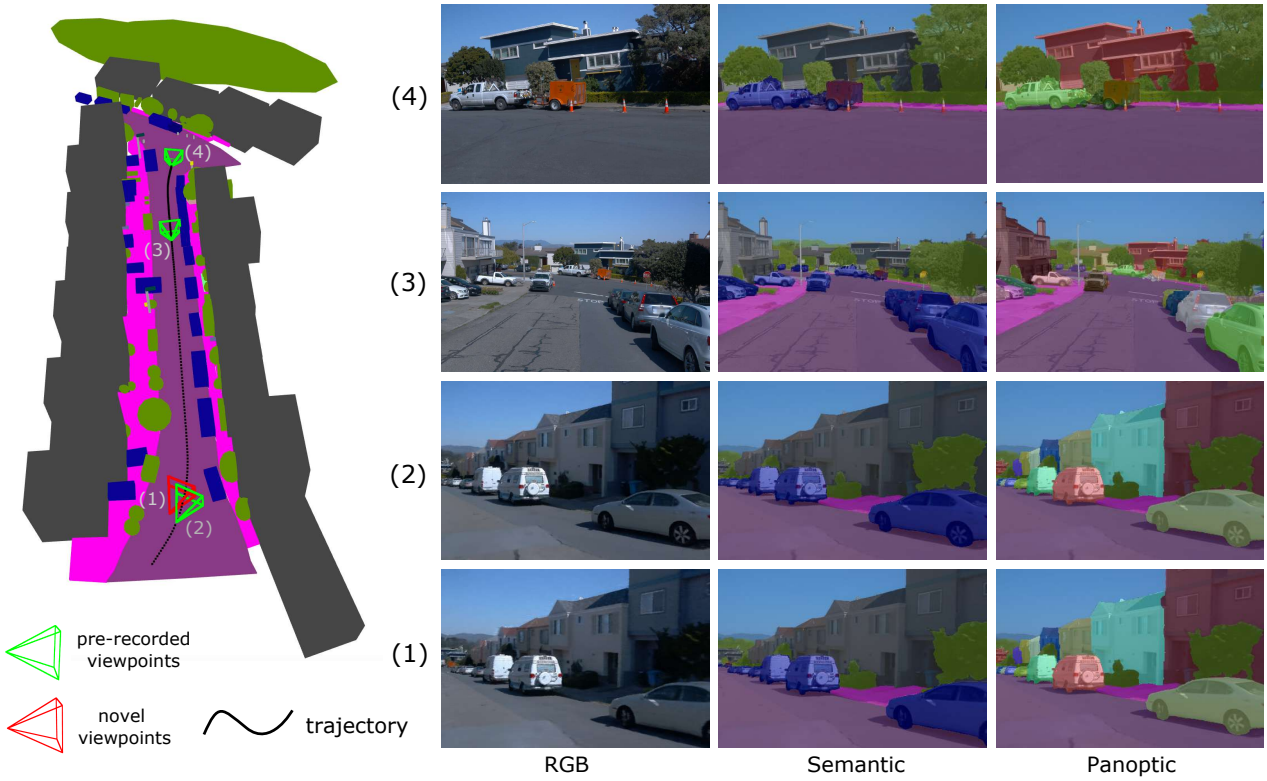


Fig. 13: **Qualitative Label Transfer Results on Waymo.** Our method can render high-quality panoptic label maps as well as high-frequency appearance on the Waymo dataset.

Method	Test View	F.T. Data	Road	Sdwlk	Terr	Bldg	Vegt	Car	Wall	Fence	Sky	mIoU	Acc
Mask2Former	Persp.	-	93.1	61.3	37.0	88.7	87.7	92.3	31.8	61.4	**94.8**	72.0	89.3
Mask2Former-ft (coarse)		(a)	94.2	66.2	37.5	90.1	**89.6**	92.8	35.7	63.1	94.3	73.7	90.2
Mask2Former-ft (coarse+ours)		(a)+(b)	94.5	71.3	37.9	90.3	89.5	93.3	38.2	64.7	94.5	74.9	90.7
Mask2Former-ft (ours)		(b)	**94.6**	**72.1**	**38.4**	**90.9**	89.3	**93.4**	**40.9**	**66.8**	94.6	**75.7**	**91.1**
Mask2Former	Fisheye	-	84.5	60.7	20.5	80.9	77.1	60.9	26.5	42.3	**98.4**	61.3	89.1
Mask2Former-ft (coarse)		(a)	87.9	64.7	25.3	83.7	**89.1**	**62.3**	33.9	43.7	98.1	65.4	89.8
Mask2Former-ft (coarse+ours)		(a)+(b)	88.3	68.4	26.5	83.5	88.9	61.9	40.3	44.5	97.7	66.7	90.4
Mask2Former-ft (ours)		(b)	**88.4**	**70.3**	**27.2**	**84.9**	88.4	**62.3**	**43.1**	**45.8**	98.3	**67.6**	**90.8**

TABLE 3: **Quantitative Comparison of Semantic Label** on Test Scenes of KITTI-360. (a): 512 coarse pseudo 2D labels (PSPNet*); (b): 1,536 synthesized labels using our method.

	Forward-facing			Fisheye		
Method	mIoU ↑	PQ ↑	PSNR ↑	mIoU ↑	PQ ↑	PSNR↑
MLP [52]	**81.2**	**69.5**	23.78	**68.2**	**55.8**	24.47
iNGP [54]	79.0	67.9	27.22	67.0	54.6	**28.14**
Tri-planes [9]	75.1	62.1	23.25	60.2	50.2	23.32
Ours	<u>80.9</u>	<u>68.7</u>	**27.25**	<u>67.6</u>	<u>55.4</u>	<u>28.05</u>

TABLE 4: **Neural Scene Representation Study** over 4 scenes. We bold and underline the two best performing methods, respectively.

fine-tune Mask2Former [13] pre-trained on CityScape [15] using our *synthesized* semantic labels on the KITTI-360. We fine-tune it using our synthesized semantic labels on the KITTI-360. Specifically, we divide 10 test scenes into 8 for training and 2 for test. For each scene, we render 3 semantic maps at each of 64 trajectory points using the left perspective camera: one in the original direction, one rotated horizontally 30° to the left, and another rotated horizontally 30° to the right, resulting in a total of 1,536 training samples. We further compare Mask2Former-ft (ours) with two additional variants: one fine-tuned solely on coarse semantic labels that we use as pseudo 2D GTs (coarse), and another trained on a combination of coarse and synthesized images (coarse + ours). During fine-tuning, we format the data to match CityScape's and exclude categories not present in CityScape. We fine-tune all the models for 5 epochs with a batch size of 8, using a learning rate of 0.0001 and AdamW optimizer. As illustrated in Table 3 and Fig. 14, Mask2Former-ft (ours) significantly outperforms the original (+3.7 mIoU on perspective views, and +6.3 mIoU on fisheye views), especially on Sidewalk and Wall. It is also reasonable that Mask2Former-ft (ours) outperforms the other two finetuning baselines as our synthesized labels significantly surpass the base 2D pseudo labels. This improvement demonstrates the effectiveness of our synthesized labels in enhancing the perception model's performance. Moreover, the larger improvement on the fish-

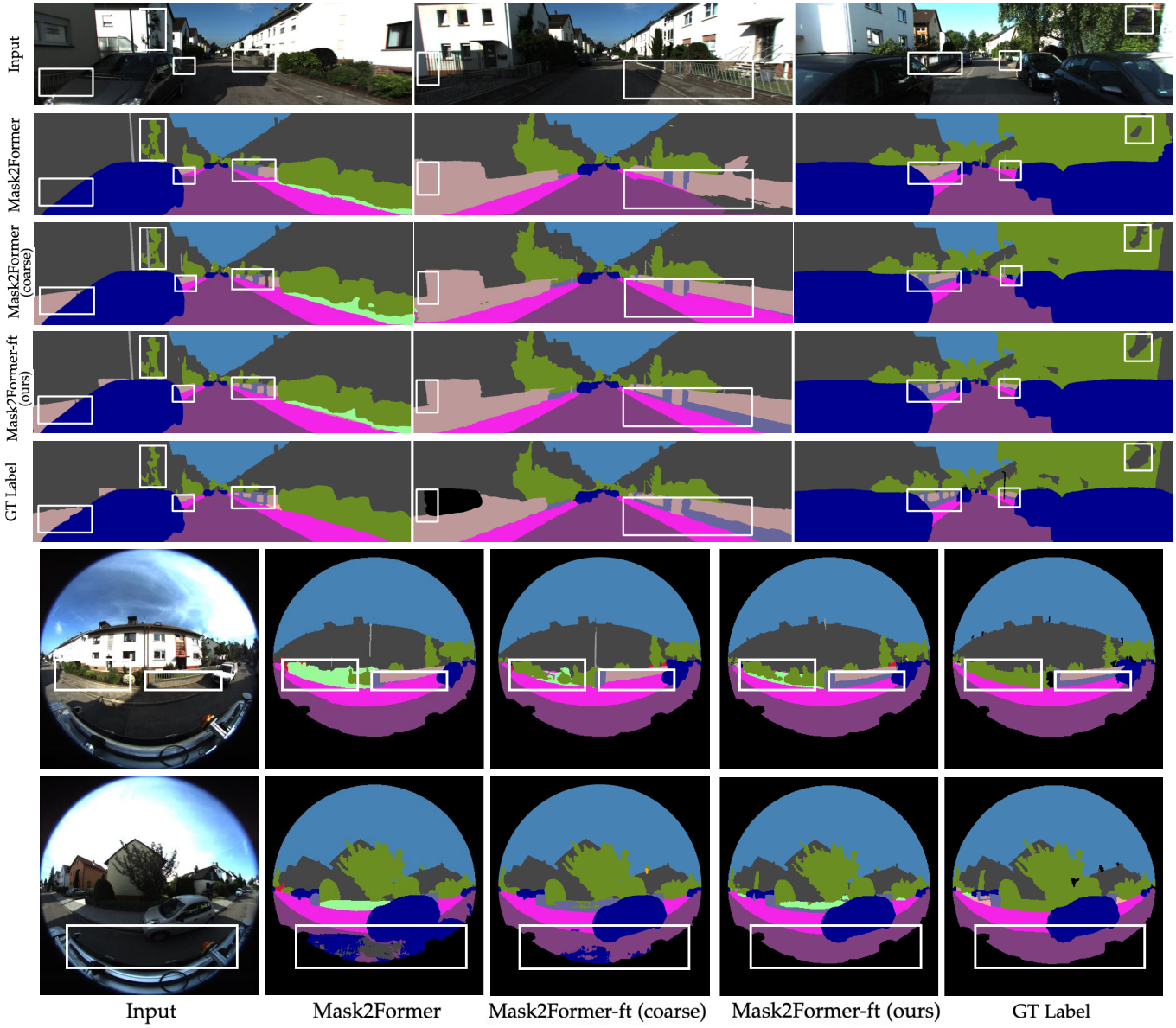


Fig. 14: **Qualitative Comparison of Semantic Prediction on Perspective and Fisheye Views.**

eye views further verifies that existing perception models degenerates on unseen viewpoints and demonstrates the importance of being able to synthesizing labels with large viewpoint changes.

4.5 Ablation Study

We validate our pipeline's design modules with extensive ablations in Table 5 by removing one component at a time. We perform label evaluation on joint perspective & fisheye views and estimate scene geometry on perspective views on one scene.

Geometric Reconstruction: 1) *Semantic Label Guidance*: We now verify that our method effectively improves the underlying geometry by leveraging semantic information. We first remove all the other losses except for $\mathcal{L}_p$, leading to a baseline, NeRF-360*, which still employs a hybrid of MLP and grids and uses our proposed sampling strategy. In this case, we render a semantic map based on the fixed semantic field s_β. As can be seen from Table 5 and Fig. 16, the underlying geometry of NeRF-360* drops significantly

	Depth (0-100m) RMSE↓	$\delta_{1.25}$ ↑	Eval. Label	Perspective & Fisheye mIoU*	Acc*	PQ*
3D-2D CRF	-	-	-	69.7	94.1	57.6
NeRF-360*	19.11	75.5	$\hat{\mathbf{S}}(s_\beta)$	64.6	88.9	54.6
w/o $\mathcal{L}_{\mathbf{s}}^{2D}$	16.90	78.4	$\mathbf{S}(s_\phi)$	71.3	93.8	61.6
w/o $\mathcal{L}_{\mathbf{s}}^{3D}$	7.57	94.7	$\mathbf{S}(s_\phi)$	74.6	94.9	64.0
w/o $\mathcal{L}_{\mathbf{s}}^{2D}$	7.57	94.7	$\hat{\mathbf{S}}(s_\beta)$	73.4	94.8	62.3
w/o $\mathcal{L}_{\mathbf{s}}^{3D}$	7.53	94.7	$\mathbf{S}(s_\phi)$	75.0	95.0	65.3
w/o $\mathcal{L}_d$	8.36	93.4	$\mathbf{S}(s_\phi)$	74.4	94.9	63.8
w/o $\mathbb{1}(\mathbf{r})$	8.29	94.5	$\mathbf{S}(s_\phi)$	75.1	95.0	65.4
Uniform S.	9.52	93.5	$\mathbf{S}(s_\phi)$	67.3	92.2	56.7
S. points↓	7.36	94.6	$\mathbf{S}(s_\phi)$	75.3	95.1	65.4
w/o $w(k)$	**7.12**	**94.8**	$\mathbf{S}(s_\phi)$	75.6	95.1	66.2
w/o fisheye*	7.90	92.5	$\mathbf{S}(s_\phi)$	74.8	94.7	64.0
Complete	7.27	**94.7**	$\mathbf{S}(s_\phi)$	**76.1**	**95.2**	**66.6**

TABLE 5: **Ablation Study** over one scene.

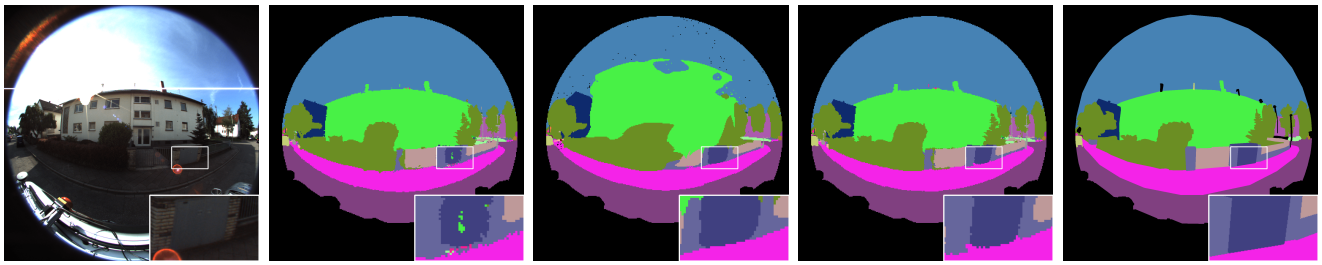


Fig. 15: **Qualitative Comparison of Neural Scene Representations on Fisheye.** Ours generates smoother labels than others.

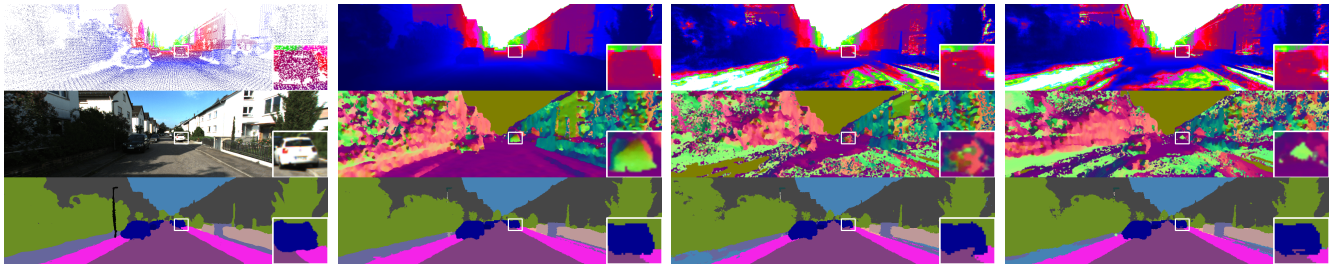


Fig. 16: **Ablation Study.** Top: LiDAR depth map and rendered depth maps. Middle: RGB input and normal maps computed as the gradient of the volume density with respect to 3D position. Bottom: Semantic GT and predictions. Note that removing $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$ leads to severely impaired geometry, and inaccurate boundary and semantic segmentation.

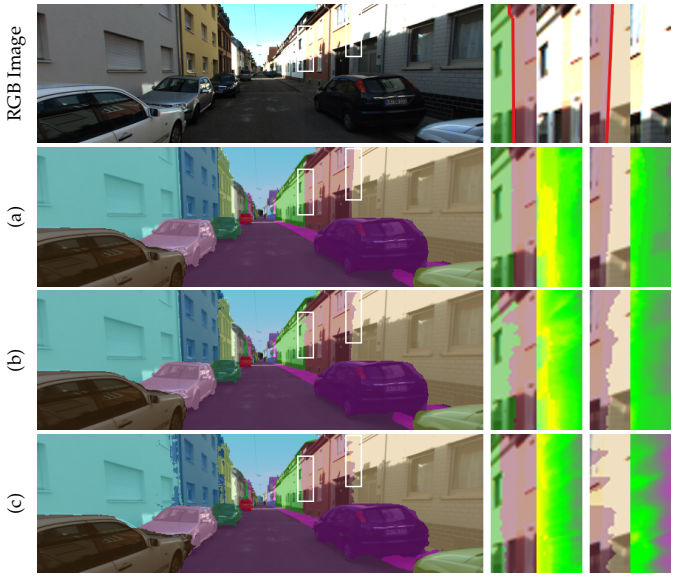


Fig. 17: **Qualitative Ablations of Fine-tuning Stage.** (a) full model (b) w/o finetuning instance (c) w/o fisheye, w/o finetuning instance. Zoom in the visualization of the overexposed area between adjacent "building"s and the corresponding depth. The red lines mark the GT boundaries.

with only using $\mathcal{L}_p$. More importantly, the depth prediction also degrades considerably when removing $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$ (w/o $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$, RMSE: 7.27→16.90), indicating the importance of the fixed semantic field in improving the underlying geometry to render accurate semantic boundaries. This can also be verified by removing fisheye semantic predictions from the input (w/o fisheye*), which results in more inaccurate geometry ($\delta_{1.25}$: 94.7→92.5) and semantics (mIoU*: 76.1→74.8).

2) *Instance Label Guidance*: To illustrate the effectiveness of $\mathcal{L}_{\hat{\mathbf{T}}}^{2D}$ (Eq. 10), we present Fig. 17 for qualitative comparision. In the marked region with overexposure, the full model achieves more accurate boundaries and smoother geometry than the model without the fine-tuning stage (w/o ft. instance).

3) *Others*: For the hybrid scene features, as shown in Fig. 5, "concatenation" performs better than "product" in merging MLP and hash features, leading to smoother and less erroneous geometry. We further show that eliminating $\mathcal{L}_d$ (w/o $\mathcal{L}_d$), replacing the sampling strategy with standard uniform sampling (Uniform S.), or removing ray masking (w/o $\mathbb{1}(\mathbf{r})$) all impair the geometric reconstruction, and consequently the semantic estimation.

Panoptic Segmentation: When removing $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$ (w/o $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$), the performance also drops as the learned semantic field s_ϕ is 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 $\mathcal{L}_{\mathbf{s}}^{3D}$ also allows us to address label ambiguity in overlapping regions to a certain extent. Therefore, it is not surprising that removing $\mathcal{L}_{\mathbf{s}}^{3D}$ (w/o $\mathcal{L}_{\mathbf{s}}^{3D}$) worsens the label prediction compared to the full model. Discarding $w(k)$ (w/o $w(k)$) causes worse label quality by sacrificing geometry (RMSE:7.27→7.12) to model imbalanced label distribution.

2D Pseudo GT: Finally, we evaluate how the quality of the pseudo 2D ground truth affects our method in Table 6. As some classes are not considered during training in Cityscapes, we additionally report mIoU_{sub} over the remaining classes. It is worth noting that using models

Method	mIoU*	mIoU_{sub}*	Acc*	PQ*
SSA [11]	-	67.4	86.0	-
Deeplab [12]	-	73.3	91.0	-
Tao *et al.* [66]	-	75.9	92.5	-
PSPNet [82]	-	71.4	90.4	-
PSPNet* [82]	67.0	74.7	92.0	-
Ours w/ [11]	65.2	71.6	90.9	58.6
Ours w/ [12]	73.3	78.6	94.5	63.6
Ours w/ [66]	73.9	78.4	95.1	64.9
Ours w/ [82]	70.7	74.3	92.3	62.5
Ours	**76.1**	**79.6**	**95.2**	**66.6**

TABLE 6: **Quantitative Comparison** using different 2D Pseudo GTs on one scene. We keep the pesudo labels for the two side-view fisheye cameras and only change the pseudo labels on the forward-facing views.

pre-trained on Cityscapes without any fine-tuning leads to promising results, where Ours w/ Deeplab [12] and Ours w/ Tao *et al.* [66] are very close to Ours + PSPNet* in terms of mIoU_{sub}. More importantly, our method consistently outperforms the corresponding 2D pseudo GT by leveraging the 3D bounding primitives. We additionally incorporate SSA, a SAM-based semantic variant to demonstrate the inefficiency of SAM [37] despite its unprecedented zero-shot mask generation ability.

5 Conclusion

We present PanopticNeRF-360 that infers in 3D space and renders omnidirectional per-pixel semantic and instance labels for 3D-to-2D label transfer. By unifying coarse 3D bounding primitives and noisy 2D semantic predictions, PanopticNeRF-360 is capable of performing mutual enhancement of geometry and semantics compared to naïve joint optimization of geometry and semantics. Specifically, it improves the underlying scene geometry given sparse input views leveraging label-guided geometry optimization, while concurrently resolving label noise based on improved geometry. Moreover, it enables label synthesis at a large range of novel viewpoints, including panoramic perspectives. We posit that our method marks a significant step towards improving data annotation efficiency while delivering a consistent, continuous 3D panoptic representation.

5.1 Limitations

Our method has several limitations: (1) We perform per-scene optimization. To further shorten the time for label transfer, future works might explore how to endow the model with training-free inference abilities on novel scenarios. This can be potentially realized by pretaining on a large number of urban scenes of diverse environments. (2) In cases where there are missing bounding boxes in distant regions, our method cannot accurately recover the correct labels and typically categorizes these areas as "sky." However, the impact of this on image-based models is usually negligible since the very distant regions represent only a small portion of the 2D image space. (3) We focus on label transfer of static scenes. It will be interesting to extend our method to dynamic scenes given annotated bounding primitives of the dynamic objects.

Potential Improvement on Backbone: 3D Gaussian Splatting [33] has demonstrated remarkable training and rendering efficiency in comparison to NeRF-like architectures. It is indeed a very interesting future direction to be explored such efficient representations to scaling the proposed method to large-scale scenes [44], [47], [48]. With the 3D bounding primitives, we can seamlessly employ the points within them for initial setup. With the 3D bounding primitives, we can seamlessly employ the points within them for initial setup. Besides, the explicit representation of gaussian splatting is more suitable for modeling and visualizing semantic occupancy compared to NeRF (implicit function).

Acknowledgments

This work was supported by NSFC under grant 62202418, U21B2004, and the Zhejiang University Education Foundation Qizhen Scholar Foundation. Andreas Geiger was supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645.

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *CVPR*, 2018.
[2] Mykhaylo Andriluka, Jasper RR Uijlings, and Vittorio Ferrari. Fluid annotation: a human-machine collaboration interface for full image annotation. In *ACM MM*, 2018.
[3] Amlaan Bhoi. Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*, 2019.
[4] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern recognition letters*, 30(2):88–97, 2009.
[5] Tom Bruls, Will Maddern, Akshay A Morye, and Paul Newman. Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data. In *ICRA*, 2018.
[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.
[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.
[8] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *CVPR*, 2017.
[9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022.
[10] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. Factor fields: A unified framework for neural fields and beyond. *arXiv preprint arXiv:2302.01226*, 2023.
[11] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything Github Repository, 2023.
[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017.
[13] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.
[14] Xinhua Cheng, Yanmin Wu, Mengxi Jia, Qian Wang, and Jian Zhang. Panoptic compositional feature field for editable scene rendering with network-inferred labels via metric learning. In *CVPR*, 2023.
[15] 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 *CVPR*, 2016.
[16] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[18] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

[19] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *3DV*, 2022.

[20] Aditya Ganeshan, Alexis Vallet, Yasunori Kudo, Shin-ichi Maeda, Tommi Kerola, Rares Ambrus, Dennis Park, and Adrien Gaidon. Warp-refine propagation: Semi-supervised auto-labeling via cycle-consistency. In *ICCV*, 2021.

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[22] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020.

[23] Nikhil Gosala, Kürsat Petek, Paulo LJ Drews-Jr, Wolfram Burgard, and Abhinav Valada. Skyeye: Self-supervised bird's-eye-view semantic mapping using monocular frontal view images. In *CVPR*, 2023.

[24] Nikhil Gosala and Abhinav Valada. Bird's-eye-view panoptic segmentation using monocular frontal view images. *RA-L*, 2022.

[25] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *IJCV*, 110:328–348, 2014.

[26] Chenhang He, Ruihuang Li, Yabin Zhang, Shuai Li, and Lei Zhang. Msf: Motion-guided sequential fusion for efficient 3d object detection from point cloud sequences. In *CVPR*, 2023.

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[28] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 30(2):328–341, 2007.

[29] Seunghoon Hong, Junhyuk Oh, Honglak Lee, and Bohyung Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *CVPR*, 2016.

[30] Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai, and Chi-Keung Tang. Instance neural radiance field. *arXiv.org*, 2304.04395, 2023.

[31] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3DV*, 2016.

[32] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *TPAMI*, 2019.

[33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM Trans. on Graphics*, 2023.

[34] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

[38] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *NeurIPS*, 2022.

[39] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011.

[40] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation. In *CVPR*, 2022.

[41] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv.org*, 2201.03546, 2022.

[42] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *ECCV*, 2020.

[43] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *TPAMI*, 45(3):3292–3310, 2022.

[44] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *CVPR*, 2024.

[45] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *CVPR*, 2019.

[46] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *TPAMI*, 33(12):2368–2382, 2011.

[47] Yang Liu, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *ECCV*, 2024.

[48] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *CVPR*, 2024.

[49] Andelo Martinovic, Jan Knopp, Hayko Riemenschneider, and Luc Van Gool. 3d all the way: Semantic segmentation of urban scenes from start to end in 3d. In *CVPR*, 2015.

[50] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *ECCV*, 2022.

[51] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

[52] 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 *ECCV*, 2020.

[53] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *CVPR*, 2023.

[54] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *ACM Trans. on Graphics*, 2022.

[55] Armin Mustafa and Adrian Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *CVPR*, 2017.

[56] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

[57] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022.

[58] Eshed Ohn-Bar, Aditya Prakash, Aseem Behl, Kashyap Chitta, and Andreas Geiger. Learning situational driving. In *CVPR*, 2020.

[59] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[60] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

[61] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[62] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *CVPR*, 2019.

[63] Charles R Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. In *CVPR*, 2021.

[64] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, 2019.

[65] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *CVPR*, 2023.

[66] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[67] Xin Tong, Xianghua Ying, Yongjie Shi, He Zhao, and Ruibin Wang. Towards cross-view consistency in semantic segmentation while varying view direction. In *IJCAI*, 2021.

[68] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *3DV*, 2022.

[69] Haithem Turki, Jason Y. Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *CVPR*, 2023.

[70] Radim Tylecek and Robert B Fisher. Consistent semantic annotation of outdoor datasets via 2d/3d label transfer. In *Sensors*, 2018.

[71] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *3DV*, pages 11–20, 2017.

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[73] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021.

[74] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. In *ICLR*, 2023.

[75] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. In *NeurIPS*, 2021.

[76] Jianxiong Xiao and Long Quan. Multiple view semantic segmentation for street view images. In *ICCV*, 2009.

[77] Jun Xie, Martin Kiefel, Ming-Ting Sun, and Andreas Geiger. Semantic instance annotation of street scenes by 3d to 2d label transfer. In *CVPR*, 2016.

[78] Xiaomeng Xu, Yanchao Yang, Kaichun Mo, Boxiao Pan, Li Yi, and Leonidas Guibas. Jacobinerf: Nerf shaping with mutual information gradients. In *CVPR*, 2023.

[79] Jesus Zarzar, Sara Rojas, Silvio Giancola, and Bernard Ghanem. Segnerf: 3d part segmentation with neural radiance fields. *arXiv.org*, 2211.11215, 2022.

[80] Junming Zhang, Haomeng Zhang, Ram Vasudevan, and Matthew Johnson-Roberson. Hyperspherical embedding for point cloud completion. In *CVPR*, 2023.

[81] Xiaoshuai Zhang, Abhijit Kundu, Thomas Funkhouser, Leonidas Guibas, Hao Su, and Kyle Genova. Nerflets: Local radiance fields for efficient structure-aware 3d scene representation from 2d supervision. In *CVPR*, 2023.

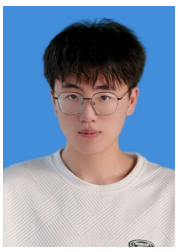
[82] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[83] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021.

[84] Brady Zhou, Philipp Krähenbühl, and Vladlen Koltun. Does computer vision matter for action? *Science Robotics*, 2019.

[85] Walter Zimmer, Akshay Rangesh, and Mohan Trivedi. 3d bat: A semi-automatic, web-based 3d annotation toolbox for full-surround, multi-modal data streams. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1816–1821. IEEE, 2019.

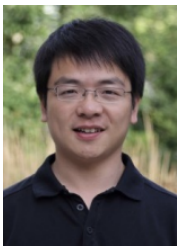
Xiao Fu received bachelor degree in Information Engineering at Zhejiang University in 2022, advised by Prof. Yiyi Liao. His research interest lies in 3D computer vision and machine learning, including neural rendering, 3D/4D reconstruction, generation and scene editing.



Shangzhan Zhang is a highly motivated master's student in Computer Science at Zhejiang University, where he is advised by Professor Xiaowei Zhou. He received his bachelor's degree from the same university in 2022.

Tianrun Chen received the bachelor's degree in College of Information Science and Electronic Engineering, Zhejiang University and is persuing Ph.D. degree in Computer Science and Technology at Zhejiang University. He is also the founder and the technical director of Moxin Technology (KOKONI) Co., LTD. His research interest includes computer vision and its enabling applications.

Yichong Lu is a senior undergraduate student in College of Information Science and Electronic Engineering, Zhejiang University. His research interest includes 3D reconstruction, generation and editing.



Xiaowei Zhou is a Research Professor of Computer Science at Zhejiang University, China. He obtained his Ph.D. degree from The Hong Kong University and Science and Technology, after which he was a postdoctoral researcher at the GRASP Lab, University of Pennsylvania. His research interests include 3D reconstruction and scene understanding.

Andreas Geiger received his Diploma in computer science and his Ph.D. degree from Karlsruhe Institute of Technology in 2008 and 2013. Currently, he is leading the Autonomous Vision Group at the University of Tubingen. He is also a core faculty member of the Tübingen AI Center. His research interests include computer vision, machine learning and scene understanding with a focus on self-driving vehicles.

Yiyi Liao is an assistant professor at Zhejiang University. Before that, she was a Postdoc at Autonomous Vision Group at the University of Tübingen and the MPI for Intelligent Systems. She received her Ph.D. in Control Science and Engineering from Zhejiang University in June 2018 and her B.S. degree from Xi'an Jiaotong University in 2013. Her research interests include 3D vision and scene understanding.

APPENDIX

Overview: In this supplementary document, we first give a detailed overview of our network architecture, sampling strategy, far-class fusing strategy, evaluation metrics, and training and inference procedure in Section A. Next, we describe our data preparation process in Section B, including the stereo depth maps for supervision, LIDAR depth maps, semantic labels for evaluation, and Acquirement of Weak 3D&2D Labels. Then, we provide additional experiments, including quantitative evaluation of bbox intersection, more label transfer results, novel label synthesis at 360° outward rotated viewpoints and panoramic viewpoints, qualitative comparison of label transfer and neural scene representations, weak depth supervision, analysis of 3D-2D CRF, and attempt in generating self-distilled pseudo instance GT using SAM in Section C. Finally, we provide failure cases, including far-region label synthesis and fisheye geometric reconstruction in Section D.

APPENDIX A
IMPLEMENTATION DETAILS

A.1 Network Architecture

Fig. S1 shows the trainable part of our PanopticNeRF-360 model. We adopt the same network architecture in all experiments.

For radiance field, we map $\mathbf{x}$ to a higher dimensional space using positional encoding (PE): $\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \cdots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p))$. Note that we also apply PE in $\mathbf{d}$ to produce a view-dependent effect. To learn high-frequency components in unbounded outdoor environments, we set $L = 15$ for $\gamma(\mathbf{x})$ and $L = 4$ for $\gamma(\mathbf{d})$.

For hash encoding, we use multi-resolution grids [54], with each grid cell vertice mapped to a hash entry: $h(\mathbf{x}) = \left(\bigoplus_{i=1}^{3} x_i\pi_i\right) \mod T$. Each hash entry stores a trainable feature. We set the number of levels 16 and grid resolutions $N_{min} = 16 \sim N_{max} = 524,288$ with hash table size $T = 2^{19}$. We query features via trilinear interpolation and concatenate the features at all spatial resolutions as input to a shallow MLP.

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 logits. The logits are then transformed into categorical distributions through a softmax layer.

Biased Density Initialization: To accelerate convergence when sampling points from coarse annotated bounding primitives, we set the bias of the density layer to 0.2, irrespective of whether the primitives belong to the "thing" or "stuff" categories.

A.2 Sampling Strategy

We sample points within the bounding primitives to skip empty space. As our bounding primitives are convex[2], 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 360° scene and t_{int} is a constant distance interval.

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

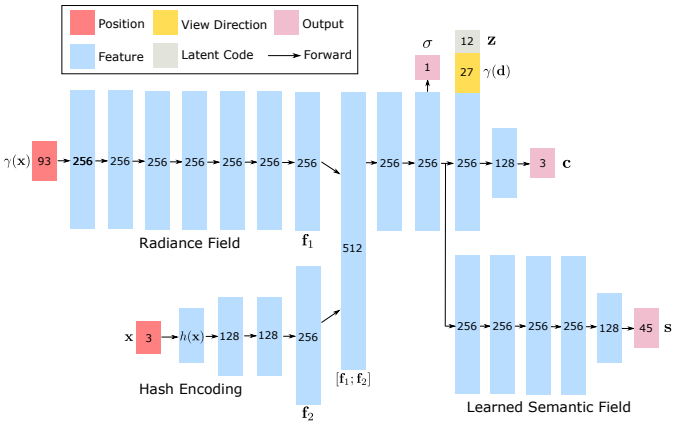


Fig. S1: **Network Architecture.** PanopticNeRF-360 takes as input the 3D location $\mathbf{x}$ (each element normalized to $[-1, 1]$), the viewing direction $\mathbf{d}$, and outputs radiance $\mathbf{c}$ and semantic logits $\mathbf{s}$. Scene feature $\mathbf{f}$ is the concatenation of MLP features $\mathbf{f}_1$ and hash encoding features $\mathbf{f}_2$. $\mathbf{z}$ is the per-frame latent embeddings.

A.3 Fusing Far-region Class

Urban scenes often contain objects that are located in far regions, resulting in missing bounding boxes in the corresponding images and misclassification as "sky" class. For pixels that have been identified as belonging to foreground classes within the 2D pseudo semantic label map, but which do not intersect bounding primitives along the corresponding ray, we replace these with available pseudo labels.

A.4 Training and Inference

As mentioned in Section 4.4 of the main paper, our total loss function comprises six terms, including three semantic losses $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$, $\mathcal{L}_{\mathbf{S}}^{2D}$, $\mathcal{L}_{\mathbf{s}}^{3D}$, the instance loss $\mathcal{L}_{\hat{\mathbf{T}}}^{2D}$, the photometric loss $\mathcal{L}_{\mathbf{c}}$ and the weak depth loss $\mathcal{L}_d$. During per-scene optimization, we resize perspective images to 704×188 pixels and fisheye images to 350×350 pixels. We further mask out invalid regions in the fisheye, including the periphery region and the ego-vehicle. The photometric loss $\mathcal{L}_{\mathbf{c}}$ is defined on the stereo images and two-side fisheye images. We apply the 2D semantic losses $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$, $\mathcal{L}_{\mathbf{S}}^{2D}$ to the left perspective images and fisheye images and the 3D semantic loss $\mathcal{L}_{\mathbf{s}}^{3D}$ directly on 3D points sampled along the camera rays of these images. The instance loss $\mathcal{L}_{\hat{\mathbf{T}}}^{2D}$ is applied to buildings on stereo perspective views as line regression is not applicable on distorted fisheye label maps. The weak depth loss $\mathcal{L}_d$ is defined only on the left images as the information

gain is marginal on the right views and monocular depth estimation on fisheye views is struggling.

For inference, we compare our method to the baselines on the left perspective views and two-side fisheye views of which the manually labeled 2D Ground Truth is defined. Note that our method is not constrained to the fixed viewpoints during inference. We show label transfer results on the 360° outward rotated views in Section C.3 and panoramic views in Section C.4.

A.5 Evaluation Metric

We evaluate mIoU and pixel accuracy following standard practice [15], [43]. 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 [67] 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 [36], we use the PQ metric to evaluate the performance of panoptic segmentation. PQ can be seen as the multiplication of a segmentation quality (SQ) term and a recognition quality (RQ) term. To mitigate the over-penalization errors related to stuff classes in PQ, we further involves PQ† [62] for comprehensive evaluation. 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 contain very small-region objects that are usually less than 100 pixels. To avoid being biased by those extremely small-region 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.

Appendix B
Data Preparation

B.1 Stereo Depth for Weak Depth Supervision

To provide weak depth supervision to PanopticNeRF-360, we use Semi-Global Matching (SGM) [28] to estimate depth given a stereo image pair on perspective views. 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 in each frame as disparity is better estimated in nearby regions, see Fig. S2.

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 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 [71].

B.3 Manually Annotated 2D GT

The manually annotated 2D ground truth of KITTI-360 [43] is inferior at some regions. For a fair comparison, we improve the label quality by manually relabeling ambiguous classes, see Fig. S3 for illustrations.

B.4 Acquirement of Weak 3D&2D Labels

We use the bounding primitives provided by the KITTI-360. As for the acquirement of these 3D labels, Liao et al. [43] first capture 3D point clouds utilizing LiDAR and stereo sensors. Next, using the KITTI-360 annotation toolkit[3], the 3D point clouds are annotated in the form of bounding primitives, *i.e.*, by placing cuboids and ellipsoids to enclose objects in 3D and assigning a semantic label to each of them. The 3D scene is annotated with 37 label classes, including 24 "thing" classes and 13 "stuff" classes. Labels are defined in accordance with the Cityscapes. To obtain more 3D bounding primitives in other scenarios, we may also utilize tools like the KITTI-360 annotation toolkit, as demonstrated in the experiments on Waymo. Besides, it is promising to leverage off-the-shelf 3D understanding algorithms [26], [63], [64], [80] to reduce the cost of labeling. We believe improving the labeling efficiency augmented by these 3D perception methods is an interesting yet orthogonal direction for future work. As for coarse 2D semantic segmentation, state-of-the-art models have intensively investigated semantic segmentation of self-driving scenarios [12], [13], [37], [66], [82]. Applying these models to obtain coarse 2D semantic segmentation masks is cost-effective.

Appendix C
Additional Experimental Results

C.1 Quantitative Evaluation of Bbox Intersection

We provide quantitative bbox intersection evaluation in correspondence to Fig. 5 in the main paper. We evaluate on two sequences ("2013_05_28_$drive$_0000_$sync$" and "2013_05_28_$drive$_0004_$sync$") that contain the 10 test scenes in KITTI-360. As shown in Table R1, the intersection between different semantic class ("sem.-sem." and "sem.-inst.") accounts to 98.5% in number and 98.4% in volume. Thus the learned semantic field is crucial to resolve label

3. https://github.com/autonomousvision/kitti360labeltool

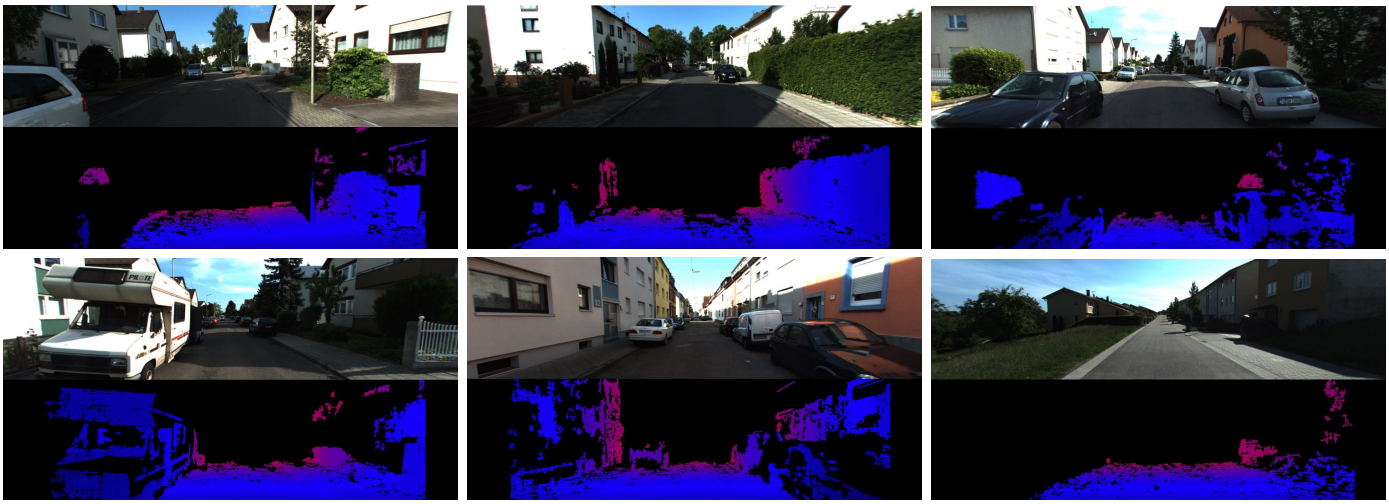


Fig. S2: **Depth Maps for Weak Depth Supervision.** Each group shows the RGB image (top) and the corresponding depth maps (bottom) used for supervision.

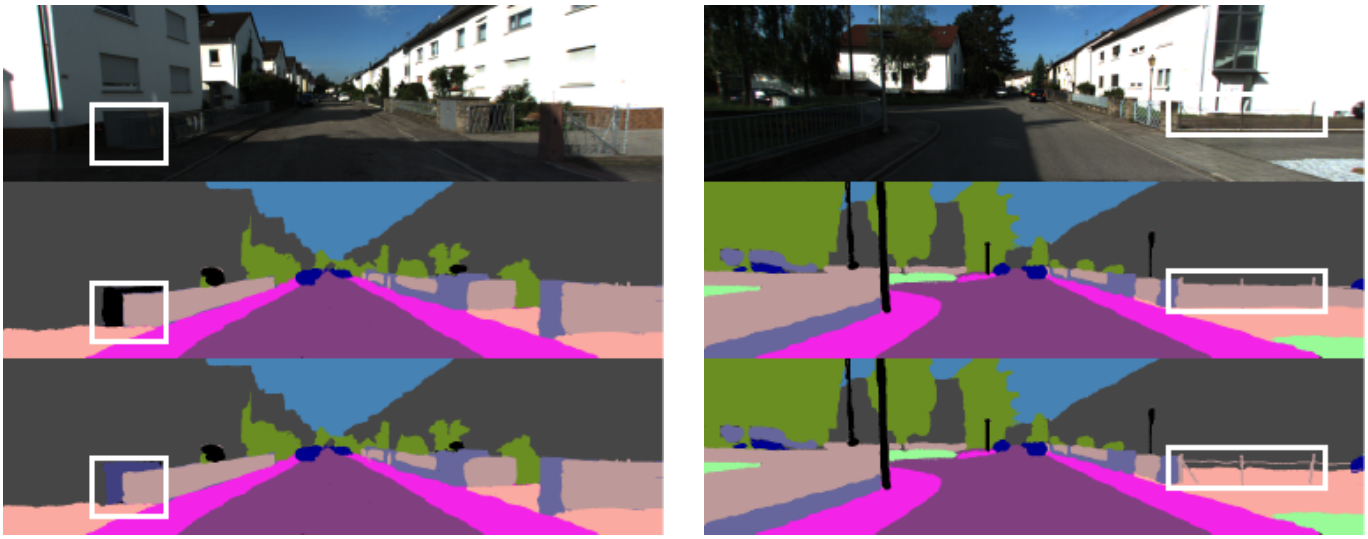


Fig. S3: **Examples of Modified Ground Truth.** We correct some GT pixels that were incorrectly labeled in the KITTI-360 dataset. Top: Input RGB images. Middle: Original ground truth. Bottom: Modified ground truth. In the first column, we add the "box" class. In the second column, we correct the "parking" area.

ambiguity using pre-trained semantic prior transferred from other datasets. As the "inst-inst" intersection is small (1.5% in number and 1.6% in volume), we ignore resolving instance intersection in our experiments. From the distribution of volume, we find that the intersected volume size and quantity of intersection numbers are inversely proportional.

C.2 More Panoptic Label Transfer Results

We provide visualization of more panoptic label transfer results both on perspective views (see Fig. S4) and fisheye views (see Fig. S5).

C.3 360° Outward Rotated Label Synthesis

PanopticNeRF-360 enables omnidirectional rendering of label and appearance. In Fig. S6, we showcase sampling at 360° rotated viewpoints around the z-axis in a scene. The angle between adjacent images is 24°. Please refer to te website for videos with 64 frames.

C.4 Panoramic Synthesis

In Fig. 11 of main paper, we show examples of panoramic semantic/instance labels at the resolution of 960×540 pixels but crop the regions that are too high and too low (50 marginal pixels). In Fig. S7, we visualize panoramic panoptic label and depth map synthesis at full-resolution.

C.5 Qualitative Comparison of Label Transfer

Fig. S8 shows additional qualitative perspective 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 in terms of panoptic label transfer on

Fig. S4: **More Perspective Panoptic Label Transfer Results** overlaid with GT RGBs and predicted panoptic labels.

	Number	Volume (m^3)
Cross-semantics Statistic		
sem.-sem.	140,229 (82.4%)	2,557,030.10 (94.2%)
sem.-inst.	27,345 (16.1%)	112,805.18 (4.2%)
inst.-inst.	2,561 (1.5%)	43,313.79 (1.6%)
Distribution of Volume (m^3)		
0-1	100,864 (59.3%)	23,028.01 (0.8%)
1-5	35,498 (20.9%)	85,393.67 (3.1%)
5-10	11,604 (6.8%)	82,593.65 (3.0%)
10-100	19,341 (11.4%)	557,977.52 (20.6%)
>100	2,828 (1.7%)	1,964,156.23 (72.4%)

TABLE R1: **Evaluation of Bbox Intersection.** 'sem.-sem.' for semantic-semantic intersection, 'sem.-inst.' for semantic-instance intersection, and 'inst.-inst.' for instance-instance intersection, respectively.

a set of unlabeled 2D perspective frames (see Fig. S9) and fisheye frames (see Fig. S10).

C.6 Qualitative Comparison of Scene Representations

We study the influence of different scene representations based on pure MLP [52], iNGP [54], and Tri-planes [9]. MLP conditions on official 8 fully-connected ReLU layers as the scene feature. iNGP adopts the lower branch of Our's network architecture in Fig. S1, and the network parameters also follows the implementation in A.1. Tri-planes takes TensoRF-VM-192 ($R_\sigma = 16, R_c = 48$) architecture with 512^3 grid voxels. We also involve the total variation (TV) loss to avoid high-frequency noise. The comparison is shown in Fig. S11. In label synthesis, iNGP generates more undesired noise and ours is comparable to MLP. The inductive smoothness bias benefits MLP to produce smoother boundaries across frequency-varying regions. For rendering appearance, Ours and iNGP are able to reconstruct high-frequency imagery, while MLP is inferior to them. Tri-planes

Fig. S5: **More Fisheye Panoptic Label Transfer Results** overlaid with GT RGBs and predicted panoptic labels.

struggles when scaled to larger scenes, while it is excellent at small-scale objects.

C.7 Weak Depth Supervision

We show that using the depth loss $\mathcal{L}_d$ alone is not able to recover accurate object boundaries in Fig. S12. 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.

C.8 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. S13 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.

Fig. S6: **Panoptic Label Transfer Results on 360 Outward Rotated Viewpoints.** The frames are overlaid with predicted panoptic labels and rendered appearance. The rotation can be recognized in the image groups from left to right, from top to down.

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.9 Generating Self-distilled Pseudo GT Using SAM

SAM [37] has extraordinary performance on object segmentation and supports multi-forms of input. We try to utilize it to generate general self-distilled pseudo panoptic GT. However, we find its open source code can not support panoptic masks as input and we have to transform them into panoptic 2D bounding boxes before sending them to the network. As shown in Fig. S14, the pseudo GTs generated by SAM have some flaws in buildings. We suppose that SAM tends to segment some components of the buildings (like doors and windows) separately instead of regarding them as part of the buildings and the performance of SAM can be deteriorated by over-exposure.

APPENDIX D
FAILURES

D.1 Far-Region Label Synthesis

As there are missing bounding boxes in far-regions, labels rendered at novel viewpoints at these areas will be classified to "sky" as shown in Fig. S15. Although we can improve the label quality via pseudo label fusion in overfitted views in A.3, the ability to render precise labels in regions with arbitrary distances at omnidirectional viewpoints remains a problem.

D.2 Fisheye Geometric Reconstruction

In our experiments, we find that the geometric reconstruction in two-side fisheye is unstable though perspective information could serve as a complementation. As illustrated in S16, there are irregular holes in the reconstructed geometry especially in low-texture and over-exposed regions.

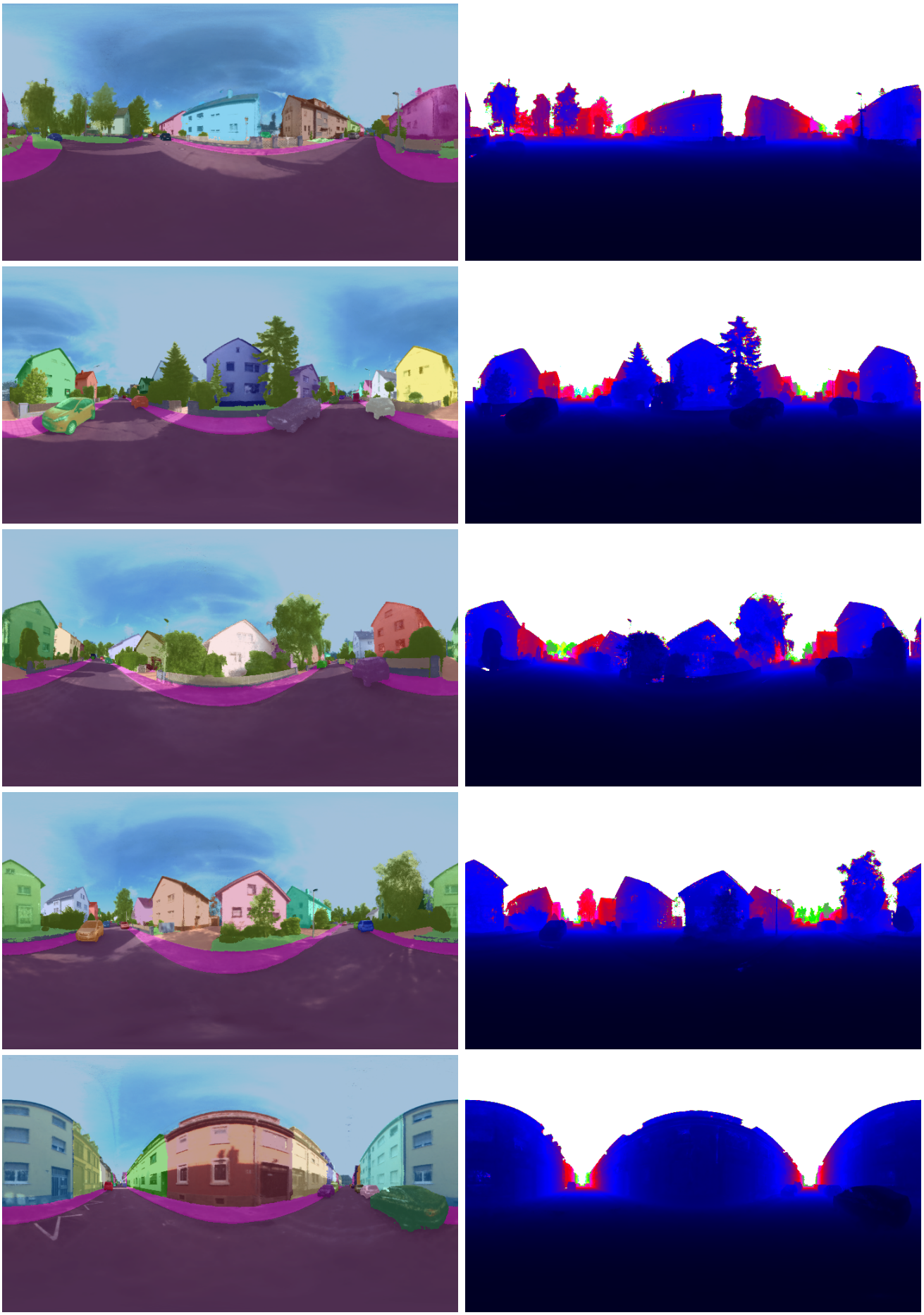


Fig. S7: **Panoramic Label** & **Depth Synthesis at 960×540 pixels**.

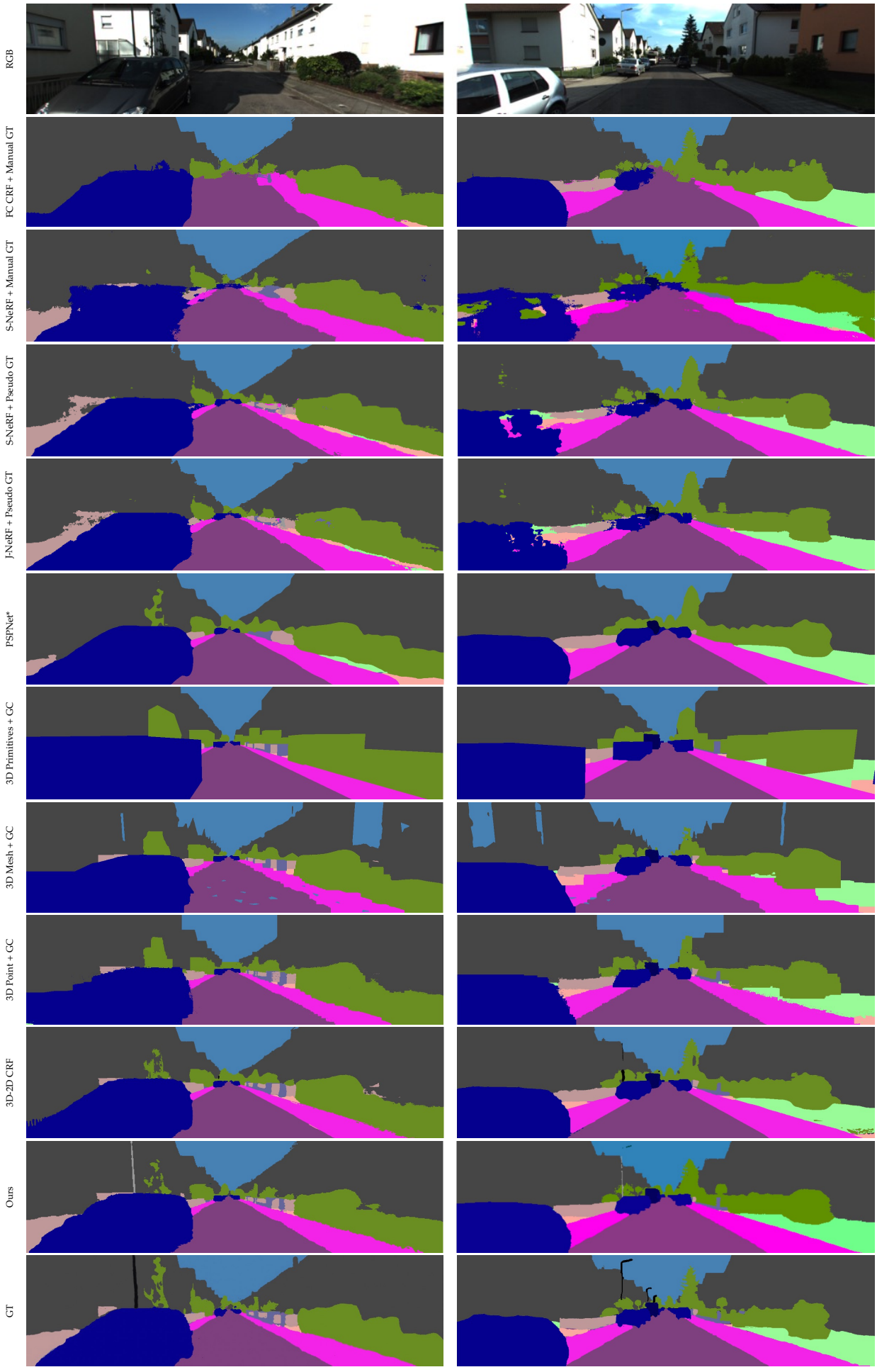


Fig. S8: **Qualitative Comparison of Perspective Semantic Label Transfer** on frames with manually labeled ground truth.

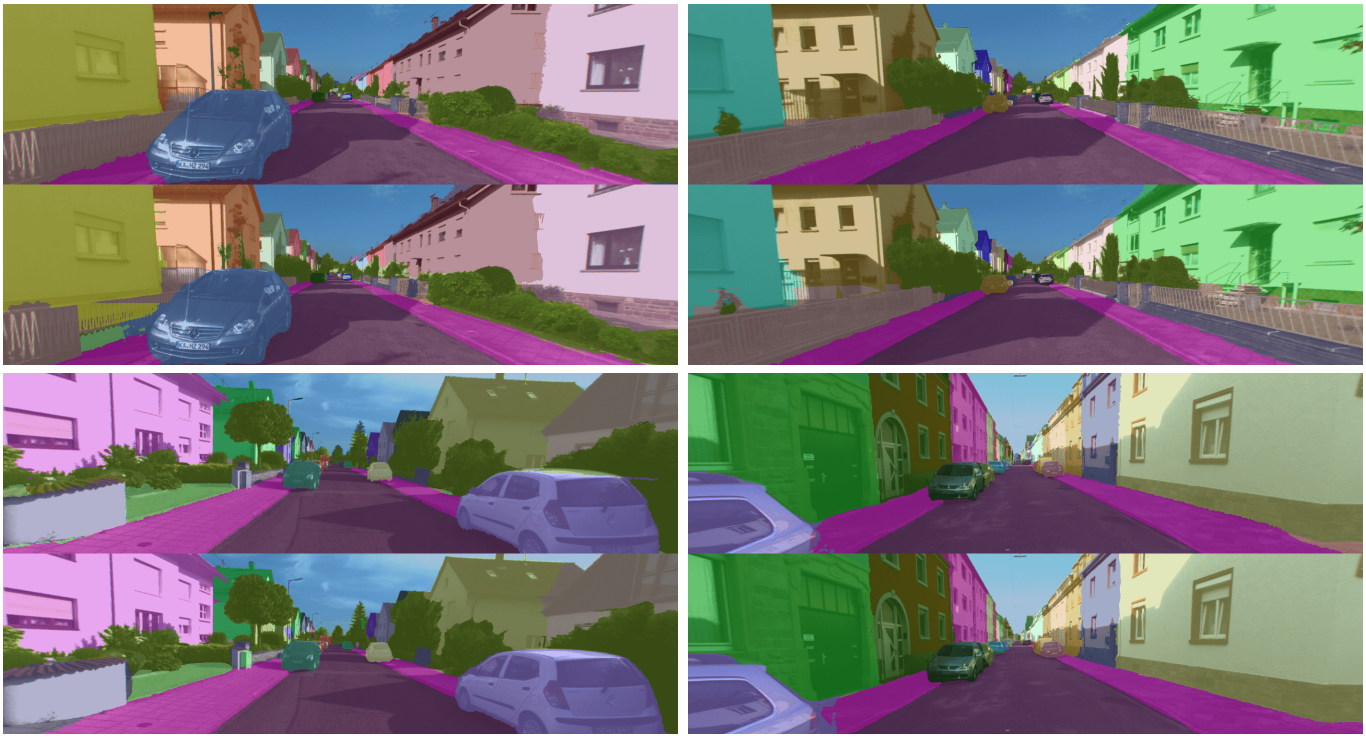


Fig. S9: **Qualitative Comparison of Perspective Panoptic Label Transfer on frames without manually labeled ground truth**. Each group shows the prediction of ours (top) and 3D-2D CRF [43] (bottom).

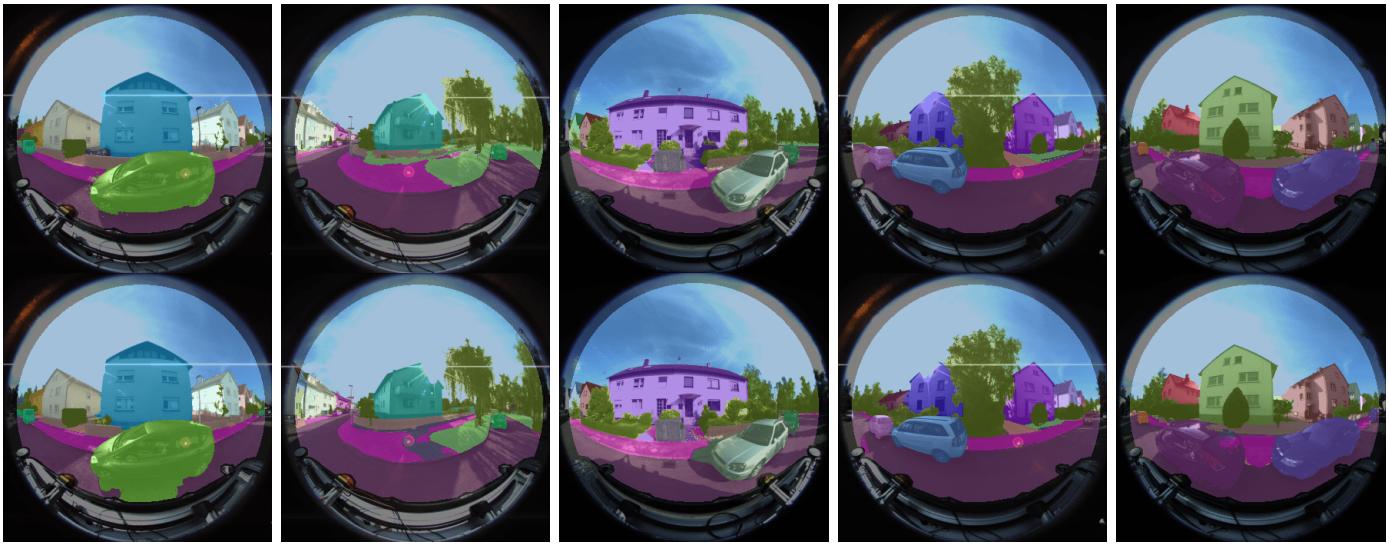


Fig. S10: **Qualitative Comparison of Fisheye Panoptic Label Transfer on frames without manually labeled ground truth**. Each group shows the prediction of ours (top) and 3D-2D CRF [43] (bottom). We can infer from groups 2 and 4 that Ours is superior to 3D-2D CRF on over-exposed areas on buildings.

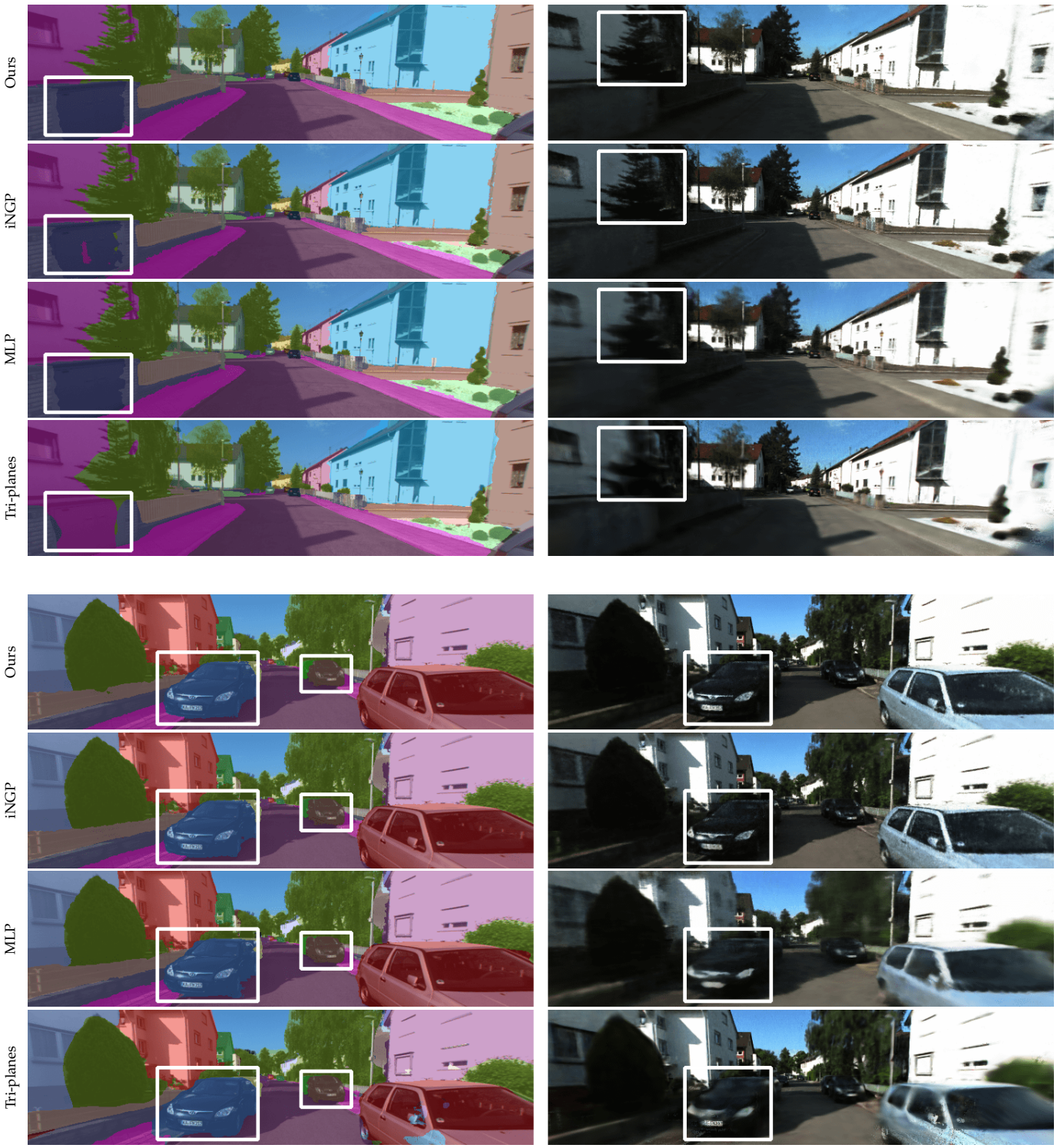


Fig. S11: **Qualitative Comparison of Neural Scene Representations**. We visualize the predicted panoptic labels overlaid with GT images (left) and rendered appearance (right).

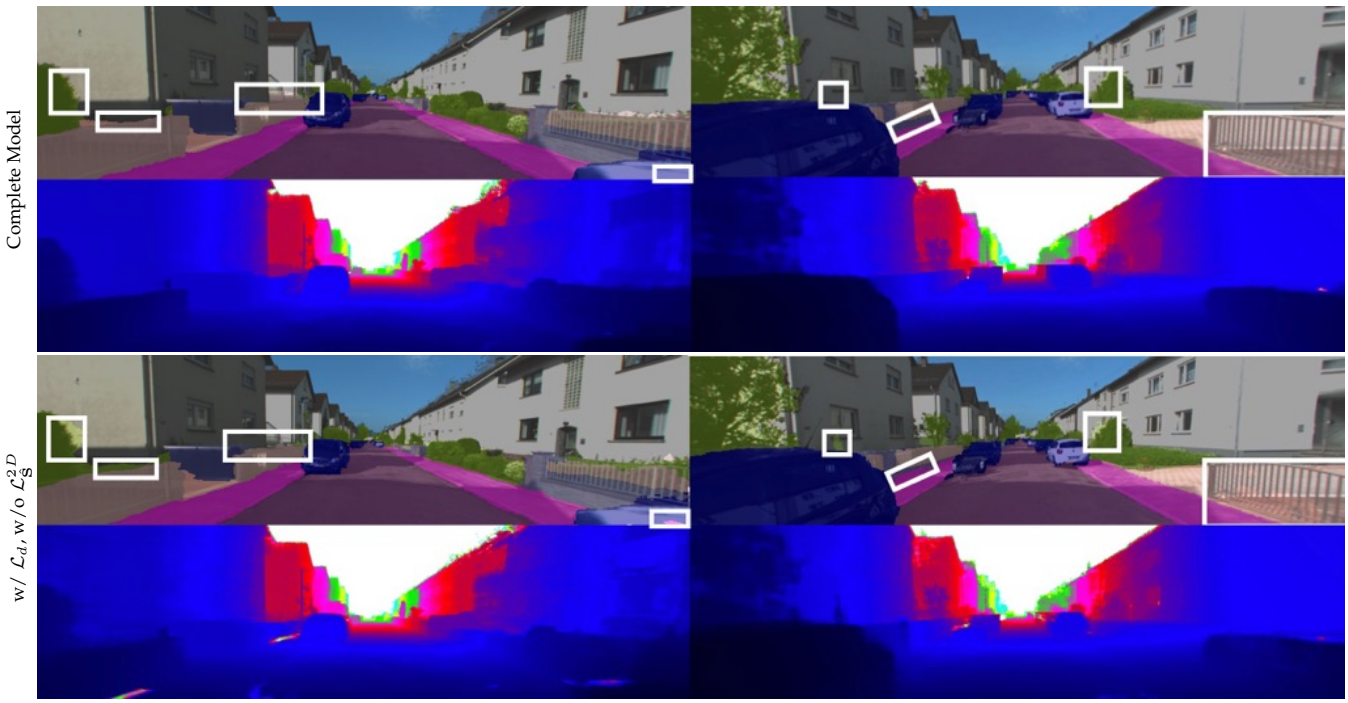


Fig. S12: **Qualitative Comparison of Ablation Study.** We visualize the semantic map and depth map of the complete model (top) and the model without fixed semantic field (bottom).

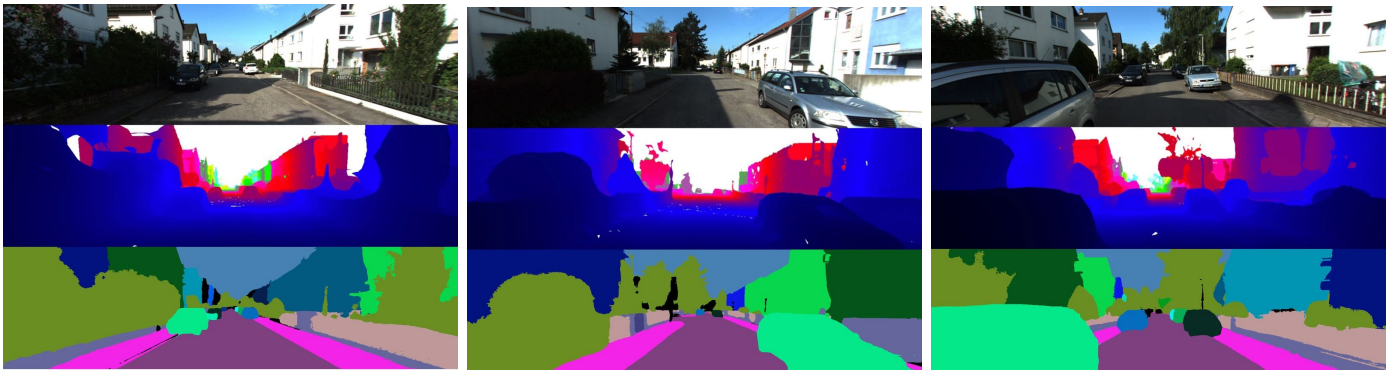


Fig. S13: **Qualitative Results of 3D-2D CRF.** Top: Input RGB images. Middle: 3D-2D CRF mesh depth. Bottom: Panoptic label transfer results of the 3D-2D CRF method.

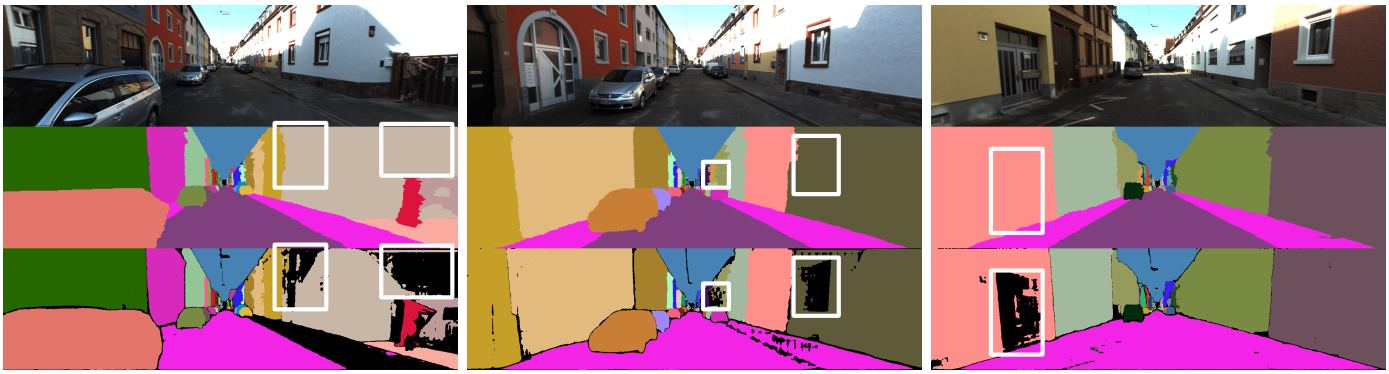


Fig. S14: **Self-distilled Pseudo GT Using SAM.** Top: Input RGB images. Middle: Middle results generated by our method. Bottom: Pseudo GT generated by SAM.

Fig. S15: **Failure in Label Synthesis in Far-Region at Novel Viewpoints**.

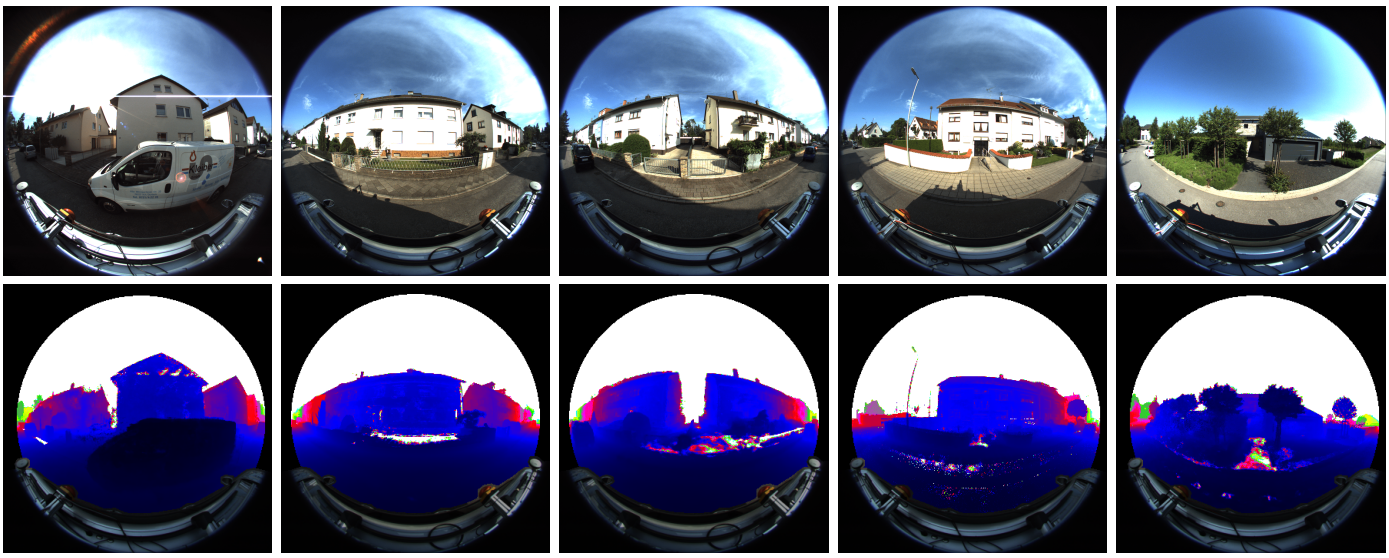


Fig. S16: **Failure in Geometric Reconstruction on Fisheye Views.** Each group contains a fisheye RGB input (upper) and reconstructed depth map (bottom).