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

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

Abstract

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

1. Introduction

Semantic instance segmentation is an important perception task for autonomous driving. It is widely acknowledged that large-scale training data with high-quality annotations is critical to propel the performance of segmentation models. However, manual annotation of pixel-accurate segmentation masks is highly expensive and time-consuming. For example, annotating all instances in a single street scene image requires up to 1.5 hours [31]. Recently, a few urban datasets propose to annotate in 3D space using coarse bounding primitives (e.g., cuboids and ellipsoids) and transfer 3D labels to 2D [59, 31, 22], significantly reducing the annotation time to 0.75 minutes per image [31]. Besides, it is often easier to separate instances in 3D rather than in 2D image space (e.g., pedestrian in front of building). Thus, there is an increasing demand for accurately transferring coarse 3D annotations to 2D semantic and instance labels.

There are only a few existing attempts in this direction. Huang *et al*. [22] perform manual post-processing to obtain accurate 2D labels based on 3D coarse annotations. Another line of work additionally leverages 2D image cues (e.g., noisy 2D semantic predictions) to avoid human intervention, achieved by combining 3D annotations and 2D image cues based on conditional random fields (CRF) [59, 31]. This CRF-based approach relies on intermediate 3D reconstructions for projecting non-occluded 3D points to 2D, and then performs inference in *2D image space*. The 3D reconstruction cannot be jointly optimized in the CRF model and thus erroneous reconstruction leads to inaccurate label transfer results. To alleviate this problem, we propose Panoptic NeRF, a novel label transfer method built on NeRF [38] that infers geometry and semantic jointly in *3D space* to render dense 2D semantic and instance labels, i.e., panoptic segmentation labels [27] (see Fig. 1).

A naïve solution is to first train a vanilla NeRF model, and then render segmentation maps based on semantic/instance labels determined by the 3D bounding primitives. However, as illustrated in Fig. 1, inaccurate geometric reconstruction leads to wrong semantic/instance maps, yet it is hard to obtain accurate geometry using the vanilla NeRF in the driving scenario where input views are sparse. Furthermore, label ambiguity at overlapping regions of the 3D bounding primitives also yields inaccurate 2D labels.

In this work, we aim to tackle both challenges. Inspired by [59, 31], we combine 3D annotations and noisy 2D semantic predictions transferred from existing datasets to fully automate the label transfer process. Specifically, our method consists of a radiance field and *dual semantic fields*, supervised by 3D and 2D weak semantic information as well as posed 2D RGB images. To improve the underlying geometry, we propose a semantically-guided geometry opti-

*Equal contribution. †Corresponding author.

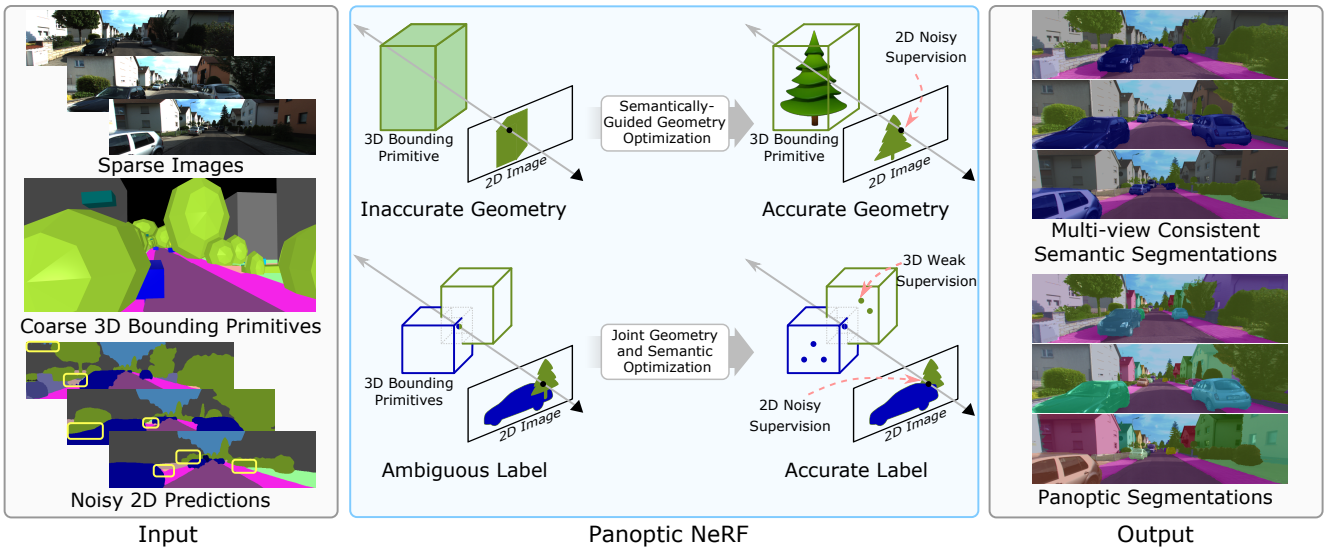

Figure 1: **Panoptic NeRF** takes as input a set of sparse images, coarse 3D bounding primitives and noisy 2D predictions (yellow boxes highlight inaccurate predictions). By inferring in 3D space, it generates semantic and instance labels in the 2D image space via volume rendering. We propose semantically-guided geometry optimization to improve the underlying geometry. We further resolve label ambiguity at intersection regions of the 3D bounding primitives intersect via joint geometry and semantic optimization, enabling rendering consistent and accurate panoptic labels at multiple views.

mization strategy based on a *fixed semantic field* determined by the 3D bounding primitives. With the semantic field fixed, we demonstrate that the geometry can be improved guided by noisy 2D semantic predictions. The semantic rendering is then further refined by joint geometry and semantic optimization, where a *learned semantic field* is adopted to fuse information of the 3D bounding primitives and the 2D noisy predictions. As evidenced by our experiments, this fusion procedure is able to resolve the label ambiguity of the 3D bounding primitives and largely eliminate noise in the 2D predictions. Furthermore, Panoptic NeRF enables rendering globally consistent 2D instance maps across multiple frames, where each object has a unique instance index determined by the 3D bounding primitives. Utilizing 3D bounding primitives of the recently released KITTI-360 dataset, Panoptic NeRF outperforms existing 3D-to-2D and 2D-to-2D label transfer methods, providing a promising approach to efficiently develop large-scale and densely labeled datasets for autonomous driving.

We summarize our contributions as follows: 1) We propose to perform 3D-to-2D label transfer by inferring in the 3D space. This allows us to unify easy-to-obtain 3D bounding primitives and noisy 2D semantic predictions in a single model, yielding high-quality panoptic labels. 2) By leveraging a novel dual formulation of the semantic fields, Panoptic NeRF effectively improves the geometric reconstruction given sparse views, yielding accurate object boundaries. Moreover, it is able to resolve label ambiguities and eliminates label noise based on the improved geometry. 3) Our Panoptic NeRF achieves superior performance compared to existing label transfer methods in terms of both semantic and instance predictions. Furthermore, our 2D semantic and instance labels are multi-view and spatio-temporally consistent by design. Finally, our method enables rendering RGB images and semantic/instance labels at novel viewpoints.

2. Related Work

Urban Scene Segmentation: Semantic instance segmentation is a critical task for autonomous vehicles [45, 64]. Learning-based algorithms have achieved compelling performance [9, 30, 62], but rely on large-scale training data. Unfortunately, annotating images at pixel level is extremely time-consuming and labor-intensive, especially for instance-level annotation. While most urban datasets provides labels in 2D image space [5, 11, 41, 25, 57], autonomous vehicles are usually equipped with 3D sensors [15, 7, 17, 22, 31]. KITTI-360 [31] demonstrates that annotating the scene in 3D can significantly reduce the annotation time. However, transferring coarse 3D labels to 2D remains challenging. In this work, we focus on developing a novel 3D-to-2D label transfer method, exploiting recent advances in neural scene representations.

Label Transfer: There have been several attempts at improving label efficiency for individual frames [33, 20, 8, 1, 32, 2]. In this paper we focus on efficient labeling of video sequences. Existing works in this area can be divided into two categories: 2D-to-2D and 3D-to-2D. 2D-to-2D label

transfer approaches reduce the workload by propagating labels across 2D images [49, 21, 47, 51, 14], whereas 3D-to-2D methods exploit additional information in 3D for efficient labeling [22, 35, 40, 58, 6]. To obtain dense labels in 2D image space, some works [31, 59] perform per-frame inference jointly over the 3D point clouds and 2D pixels using a non-local multi-field CRF model. However, these methods require reconstructing a 3D mesh to project 3D point clouds to 2D. As it is treated as a pre-processing step, the mesh reconstruction is not jointly optimized in the CRF model. Thus, inaccurate reconstruction hinders label transfer performance. In contrast, Panoptic NeRF provides a novel end-to-end method for 3D-to-2D label transfer where geometry and semantic estimations are jointly optimized.

Coordinate-based Neural Representations: Recently, coordinate-based neural representations has received wide attention. in many areas, including 3D reconstruction [48, 23, 16, 37, 43, 60, 18, 44, 50, 54], novel view synthesis [38, 3, 24, 34], and 3D generative modeling [19, 36, 53]. In this paper, we focus on utilizing coordinate-based representations to estimate the semantics of the scene. Towards this goal, NeSF [56] focuses on generalizable semantic field learning from density grids supervised by 2D GT labels, but we concentrate on the 3D-to-2D label transfer task without access to 2D GT. A closely related work Semantic NeRF [63] explores NeRF for semantic fusion. However, Semantic NeRF takes as input ground truth 2D labels or synthetic noise labels, which struggles to produce correct labels given real-world predictions from pre-trained 2D models. Moreover, Semantic NeRF operates in indoor scenes with dense RGB inputs and degenerates in challenging outdoor driving scenarios with sparse input views. Finally, Semantic NeRF is limited to rendering semantic labels, whereas our method can render panoptic labels. A concurrent work PNF [29] allows for rendering papotic labels. While PNF focuses on parsing the scene using known classes of the Cityscapes dataset [11] based on pre-trained segmentation and detection models, we aim to transfer 3D annotations to 2D image space for arbitrary classes to enable the development of new datasets, e.g., providing instance labels for buildings that are not available in Cityscapes.

3. Background

NeRF: NeRF [38] models a 3D scene as a continuous neural radiance field f_θ. Specifically, it maps a 3D coordinate $\mathbf{x}$ and a viewing direction $\mathbf{d}$ to a volume density σ and an RGB color value $\mathbf{c}$:

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

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

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

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

Problem Formulation: As shown in Fig. 1, Panoptic NeRF aims to transfer coarse 3D bounding primitives to dense 2D semantic and instance labels. In addition to a sparse set of posed RGB images, we assume a set of 3D bounding primitives $\beta = \{B_k\}_{k=1}^{K}$ to be available. These 3D bounding primitives cover the full scene in the form of cuboids, ellipsoids and extruded polygons. Each 3D bounding primitive B_k has a *semantic* label, belonging to either "stuff" or "thing". For "thing" classes, B_k is additionally associated with a unique *instance* ID. We further apply a pre-trained semantic segmentation model to the RGB images to obtain a 2D semantic prediction for each image. With this input information, our primary goal is to generate multi-view consistent, semantic and instance labels at the input frames, whereas rendering RGB images and panoptic labels at novel viewpoints is also enabled by inferring in the 3D space.

4. Methodology

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

4.1. Dual Semantic Fields

To jointly improve geometry and semantics, we define dual semantic fields, one is determined by the 3D bounding primitives β and the other is learned by a semantic head ϕ

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

where M_s denotes the number of semantic classes. In combination with the volume density of the radiance field f_θ,

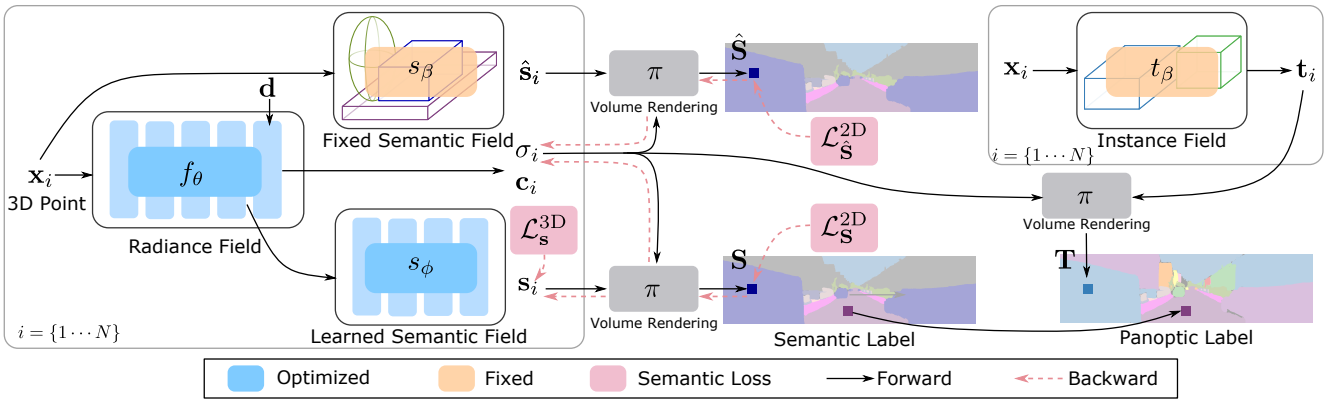

Figure 2: **Method Overview**. *Left* (Semantic Segmentation): At each 3D location $\mathbf{x}_i$, $i = \{1 \cdots N\}$, sampled along a ray, we leverage dual semantic fields to obtain two semantic categorical distributions, $\hat{\mathbf{s}}_i$ and $\mathbf{s}_i$. The 3D semantic distributions are accumulated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

two semantic distributions $\hat{\mathbf{S}}(\mathbf{r})$ and $\mathbf{S}(\mathbf{r})$ can be obtained at each camera ray $\mathbf{r}$ via the volume rendering operation π:

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

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

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

Learned Semantic Field: We add a semantic head parameterized by ϕ to NeRF to learn the semantic distribution $\mathbf{s}$. We apply a softmax operation at each 3D point to ensure that $\mathbf{s}$ is a categorical distribution. The detailed network structure can be found in the supplementary material.

4.2. Loss Functions

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

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

where $\hat{\mathbf{S}}_k(\mathbf{r})$ denotes the probability of the camera ray $\mathbf{r}$ belonging to the class k, and $\mathbf{S}_k^*(\mathbf{r})$ denotes the corresponding pseudo-2D ground truth. As illustrated in Fig. 3, the key to improve density is to directly apply the semantic loss to the fixed semantic field s_β, where $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$ can only be minimized by updating the density σ. Fig. 3 shows that a correct $\mathbf{S}^*$ increases the volume density of 3D points inside the correct 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, it means $\mathbf{S}^*$ corresponds to an occluding/occluded bounding primitive with wrong depth. To compensate, we introduce a weak depth supervision $\mathcal{L}_d$ based on stereo matching to alleviate the misguidance of $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$. 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). Adding our semantically-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.

Joint Geometry and Semantic Optimization: While enabling improved geometry, the 3D label of the overlapping

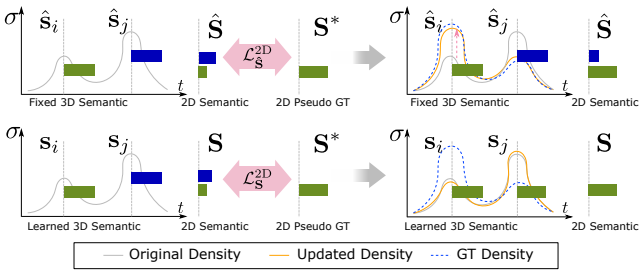

Figure 3: **Semantically-Guided Geometry Optimization**. The top row illustrates a single ray of the fixed semantic field s_β, where $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$ can only update the underlying geometry as the semantic distribution $\hat{\mathbf{s}}$ is fixed. The second row shows a single ray of the learned semantic field s_ϕ. In this case, the network can "cheat" by adjusting the semantic prediction $\mathbf{s}$ to satisfy $\mathcal{L}_{\mathbf{S}}^{2D}$ instead of updating the density σ.

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. We apply a cross-entropy loss $\mathcal{L}_{\mathbf{S}}^{2D}$ to each camera ray based on the filtered 2D pseudo ground truth, where $\mathbf{u}(\mathbf{r})$ is set to 1 if $\mathbf{S}^*(\mathbf{r})$ matches the semantic class of any bounding primitive along the ray and otherwise 0. To further suppress noise in the 2D predictions, we add a per-point semantic loss $\mathcal{L}_{\mathbf{s}}^{3D}$ based on the 3D bounding primitives

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

where $\mathbf{u}_i$ is a per-point binary mask. $\mathbf{u}_i$ is set to 1 if (1) $\mathbf{x}_i$ has a unique 3D semantic label and (2) the density σ is above a threshold σ_{th} to focus on the object surface. As illustrated in Fig. 3, $\mathcal{L}_{\mathbf{S}}^{2D}(\theta,\phi)$ does not necessarily improve the underlying geometry as the network can simply adjust the semantic head s_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 [12, 42].

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

$$\mathcal{L} = \lambda_{\hat{\mathbf{S}}}\mathcal{L}_{\hat{\mathbf{S}}}^{2D} + \lambda_{\mathbf{S}}\mathcal{L}_{\mathbf{S}}^{2D} + \lambda_{\mathbf{s}}\mathcal{L}_{\mathbf{s}}^{3D} + \lambda_{\mathbf{C}}\mathcal{L}_p + \lambda_d\mathcal{L}_d \tag{7}$$

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

4.3. Rendering of Panoptic Labels

Based on our learned semantic field s_ϕ and the 3D bounding primitives β with instance IDs, we can easily render a panoptic segmentation map. Specifically, for a camera ray $\mathbf{r}$, the panoptic label directly takes the class with maximum probability in $\mathbf{S}(\mathbf{r})$ if it is a "stuff" class. For "thing" classes, we render an instance distribution $\mathbf{T}(\mathbf{r})$ based on the bounding primitives β to replace $\mathbf{S}$ with $\mathbf{T}$. Our instance field is defined as follow

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

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

4.4. Implementation Details

Sampling Strategy and Sky Modeling: With the 3D bounding primitives covering the full scene, we sample points inside the bounding primitives to skip empty space. For each ray, we optionally sample a set of points to model the sky after the furthest bounding primitive. More details regarding the sampling strategy can be found in the supplementary. Our sampling strategy allows the network to focus on the non-empty region. As evidenced by our experiments, this is particularly beneficial in unbounded outdoor environments.

Training: We optimize one Panoptic NeRF model per scene, using a single NVIDIA 3090. For each scene, we set the origin to the center of the scene. We use Adam [26] with a learning rate of 5e-4 to train our models. We set loss weights to $\lambda_{\hat{\mathbf{S}}} = 1, \lambda_{\mathbf{S}} = 1, \lambda_{\mathbf{s}} = 1, \lambda_{\mathbf{C}} = 1, \lambda_d = 0.1$, and the density threshold to $\sigma_{th} = 0.1$. We optimize the total loss $\mathcal{L}$ for 80,000 iterations.

5. Experiments

Dataset: We conduct experiments on the recently released KITTI-360 [31] dataset. KITTI-360 is collected in suburban areas and provides 3D bounding primitives covering the full scene. Following [31], we evaluate Panoptic NeRF on

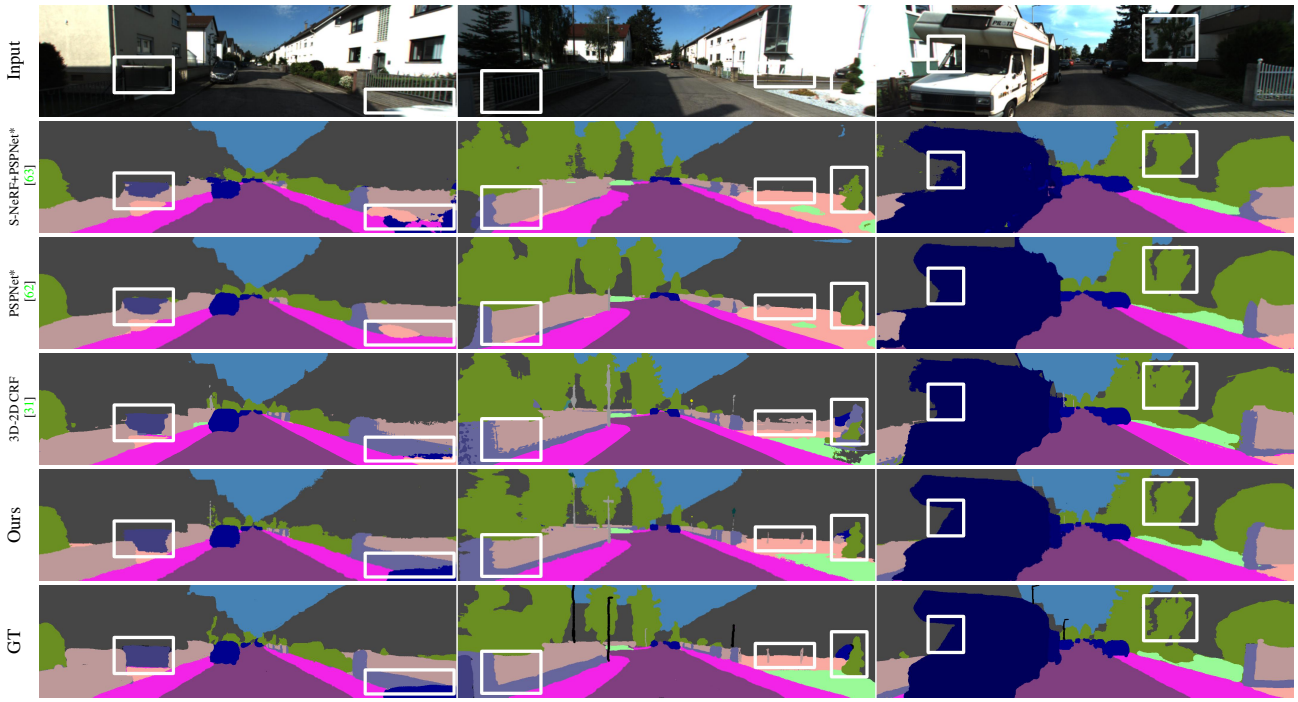

Figure 4: **Qualitative Comparison of Semantic Label Transfer.** Our method achieves superior performance in challenging regions compared to the baselines, e.g. in under- or over-exposed regions, by recovering the underlying 3D geometry, see box next to the building (left) and building above the caravan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

Method	Road	Park	Sdwlk	Terr	Bldg	Vegt	Car	Trler	Crvn	Gate	Wall	Fence	Box	Sky	mIoU	Acc	MC
FC CRF + Manual GT [28]	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 [63]	87.0	35.8	64.7	58.2	83.4	76.3	70.3	93.5	76.5	41.4	44.0	52.6	29.0	92.0	64.6	86.8	88.98
S-NeRF + PSPNet* [63]	94.6	52.9	77.7	65.0	88.0	80.8	87.9	58.3	86.0	36.0	44.1	56.8	42.2	90.9	68.6	90.5	93.42
PSPNet* [62]	95.5	49.7	77.5	66.7	88.9	82.4	91.6	46.5	83.1	24.2	43.3	51.3	51.1	89.3	67.2	90.7	91.79
3D 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 [31]	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	**95.6**	**68.4**	**84.1**	**69.5**	**91.0**	**84.4**	**93.0**	**94.8**	**93.7**	**71.6**	63.1	**74.4**	**59.7**	91.7	**81.1**	**93.2**	95.02

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

manually annotated frames from 5 static suburbs. We split these 5 suburbs into 10 scenes, comprising 128 consecutive frames each with an average travel distance of 0.8m between frames. We leverage all 128 pairs of posed stereo images for training. KITTI-360 provides a set of manually labeled frames sampled in equidistant steps of 5 frames. We use half of the manually labeled frames for evaluation and provide the other half as input to 2D-to-2D label transfer baselines. We improve the quality of the manually labeled ground truth which is inaccurate at ambiguous regions, see supplementary for details.

Baselines: We compare against top-performing baselines in two categories: (1) *2D-to-2D label transfer baselines*, including Fully Connected CRF (FC CRF) [28] and Semantic NeRF [63]. For both baselines, 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 [59]. As these 2D annotations are extremely sparse, we further provide the same pseudo-2D labels used in our method to Semantic NeRF. (2) *3D-to-2D label transfer baselines*, including PSPNet*, 3D Primitives/Meshs/Points+GC [31], and 3D-2D CRF [31]. 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: We use PSPNet* to provide pseudo ground truth in our main experiment to supervise our dual seman-

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

Method		PQ	SQ	RQ	PQ†	Method		PQ	SQ	RQ	PQ†
3D-2D CRF	All	62.2	79.1	76.9	64.9	Ours	All	**64.4**	**79.3**	**79.6**	**66.9**
	Things	60.7	79.5	**75.2**	60.7		Things	**61.9**	**80.7**	**75.2**	**61.9**
	Stuff	63.0	**78.9**	77.9	67.3		Stuff	**65.8**	78.6	**82.0**	**69.8**

Table 2: **Quantitative Comparison of Panoptic Label Transfer** over all 10 test scenes on KITTI-360.

tic fields. This ensures fair comparison to the 3D-2D CRF, which takes the predictions of PSPNet* as unary terms. Note that PSPNet* is fine-tuned on KITTI-360. To further simplify the entire process, in the ablation study we investigate the performance of our method using pre-trained models on Cityscape without any fine-tuning, including PSPNet [62], Deeplab [9] and Tao *et al.* [55].

Metrics: We evaluate semantic labels by the mean Intersection over Union (mIoU) and the average pixel accuracy (Acc) metrics. To quantitatively evaluate multi-view consistency (MC), we utilize LiDAR points to retrieve corresponding pixel pairs between two consecutive evaluation frames. The MC metric is then calculated as the ratio of pixels pairs with consistent semantic labels over all pairs. For evaluating panoptic segmentation, we report Panoptic Quality (PQ) [27], which can be decomposed into Segmentation Quality (SQ) and Recognition Quality (RQ). We additionally adopt PQ† [52] as PQ over-penalizes errors of stuff classes. To verify that Panoptic NeRF is able to improve the underlying geometry, we further evaluate the rendered depth compared to sparse depth maps obtained from LiDAR using Root Mean Squared Error (RMSE) and the ratio of accurate predictions ($\delta_{1.25}$) [4, 13].

5.1. Label Transfer

We evaluate our model and compare it to our baselines on KITTI-360. As most baselines are not designed for panoptic label transfer, we first compare the semantic predictions of all methods and then compare our panoptic predictions to the 3D-2D CRF.

Semantic Label Transfer: As evidenced by Table 1, our method achieves the highest mIoU and Acc over a line of baselines. Specifically, compared to 3D-2D CRF, we obtain an absolute improvement of 1.6% ($79.5\% \rightarrow 81.1\%$) on mIoU. Despite PSPNet* being finetuned on KITTI-360 which reduces the performance gap, our method outperforms PSPNet* by a significant margin. Supervised by the extremely sparse manually annotated GT, Semantic NeRF struggles to produce reliable performance. Using pseudo labels of PSPNet*, Semantic NeRF is capable of denoising and thus improving performance ($67.2\% \rightarrow 68.6\%$). However, both variants of Semantic NeRF are inferior in the urban scenario when the input views are sparse. Moreover, in terms of MC, our method slightly outperforms the 3D-2D CRF and significantly surpasses 2D-to-2D label transfer methods. While our method slightly lags behind 3D Point + GC in terms of MC, it is reasonable as the label consistency is evaluated on the sparsely projected 3D points which GC takes as input to generate a dense label map.

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

Novel View Label Synthesis: Panoptic NeRF can render RGB images and panoptic labels at novel viewpoints, whereas 3D-2D CRF is not capable of doing so. Please refer to the supplementary material for details.

5.2. Ablation Study

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

Geometric Reconstruction: We now verify that our method effectively improves the underlying geometry lever-

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

	Depth(0-100m)		Evaluated	Semantic	
	RMSE↓	$\delta_{1.25}$ ↑	Label	mIoU	Acc
3D-2D CRF	-	-	-	79.4	93.7
NeRF*	10.71	78.4	$\hat{\mathbf{S}}$ (fixed s_β)	67.2	88.6
w/o $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$	6.28	93.1	$\mathbf{S}$ (learned s_ϕ)	76.9	93.1
w/o $\mathcal{L}_d$	6.15	94.0	$\mathbf{S}$ (learned s_ϕ)	79.1	94.1
Uniform S.	6.28	91.1	$\mathbf{S}$ (learned s_ϕ)	73.5	92.2
w/o $\mathcal{L}_{\mathbf{S}}^{2D}$	6.01	95.0	$\mathbf{S}$ (learned s_ϕ)	80.8	94.2
w/o $\mathcal{L}_{\hat{\mathbf{S}}}^{2D}$	6.01	95.0	$\hat{\mathbf{S}}$ (fixed s_β)	79.4	94.0
w/o $\mathcal{L}_{\mathbf{s}}^{3D}$	5.74	95.0	$\mathbf{S}$ (learned s_ϕ)	70.7	92.8
w/o $\mathbf{u}(\mathbf{r})$	5.84	94.8	$\mathbf{S}$ (learned s_ϕ)	80.8	94.4
Complete	**5.23**	**95.1**	$\mathbf{S}$ (learned s_ϕ)	**81.4**	**94.5**

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

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

Semantic Segmentation: When removing $\mathcal{L}_{\mathbf{S}}^{2D}$ (w/o $\mathcal{L}_{\mathbf{S}}^{2D}$), 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 to address the label ambiguity of the overlapping region 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 semantic prediction compared to the full model. We observe that the performance is also deteriorated without ray masking (w/o $\mathbf{u}(\mathbf{r})$).

Method	mIoU	mIoU_{sub}	Acc	Method	mIoU	mIoU_{sub}	Acc
Deeplab [9]	-	74.7	88.1	Ours w/ [9]	79.3	86.3	94.0
Tao *et al.* [55]	-	79.9	91.2	Ours w/ [55]	79.8	86.1	94.5
PSPNet [62]	-	70.9	87.0	Ours w/ [62]	75.9	81.7	91.6
PSPNet* [62]	62.0	77.5	90.1	Ours	81.4	87.3	94.5

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

2D Pseudo GT: Finally, we evaluate how the quality of the pseudo 2D ground truth affects our method in Table 4. 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 pre-trained on Cityscapes without any fine-tuning leads to promising results, where Ours w/ Deeplab [9] and Ours w/ Tao *et al*. [55] are very close to Ours + PSPNet* in terms of mIoU_{sub}. More importantly, our method consistently outperforms the corresponding pseudo GT leveraging the 3D bounding primitives.

5.3. Limitations

Our method performs per-scene optimization and training takes 4 hours on one scene. Training time needs to be reduced to scale our method to large-scale scenes, e.g., by adopting improvements in speeding up NeRF training [39, 61]. In addition, we consider label transfer on static scenes in this work. We plan to extend our method to dynamic scenes leveraging recent advances in dynamic radiance field estimation [46].

6. Conclusion

We present Panoptic NeRF that infers in 3D space and renders per-pixel semantic and instance labels for 3D-to-2D label transfer. By combining coarse 3D bounding primitives and noisy 2D predictions using our dual semantic fields, Panoptic NeRF is capable of improving the underlying geometry given sparse input views and resolving label noise. Moreover, it enables label synthesis at novel view points. We believe that our method is a step towards more efficient data annotation, while simultaneously providing a 3D consistent continuous panoptic representation of the scene.

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–868, 2018. 2

[2] Mykhaylo Andriluka, Jasper RR Uijlings, and Vittorio Ferrari. Fluid annotation: a human-machine collaboration interface for full image annotation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1957–1966, 2018. 2

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3

[4] Amlaan Bhoi. Monocular depth estimation: A survey. *arXiv.org*, 1901.09402, 2019. 7

[5] 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. 2

[6] Tom Bruls, Will Maddern, Akshay A Morye, and Paul Newman. Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1863–1870. IEEE, 2018. 3

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 2

[8] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5230–5238, 2017. 2

[9] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 2, 7, 8

[10] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020. 7

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 2, 3

[12] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv.org*, 2107.02791, 2021. 5

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 7

[14] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15499–15509, 2021. 3

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 2

[16] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020. 3

[17] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: audi autonomous driving dataset. *arXiv.org*, 2004.06320, 2020. 2

[18] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv.org*, 2002.10099, 2020. 3

[19] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv.org*, 2110.08985, 2021. 3

[20] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. 110(3):328–348, 2014. 2

[21] Seunghoon Hong, Junhyuk Oh, Honglak Lee, and Bohyung Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3212, 2016. 3

[22] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2702–2719, 2019. 1, 2, 3

[23] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 3

[24] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph render-

ing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4287–4297, 2021. 3

[25] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9868, 2020. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv.org*, 1412.6980, 2014. 5

[27] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 1, 7

[28] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems*, 24, 2011. 6

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[30] 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 *Proceedings of the European Conference on Computer Vision*, pages 775–793. Springer, 2020. 2

[31] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv.org*, 2109.13410, 2021. 1, 2, 3, 5, 6, 7

[32] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5257–5266, 2019. 2

[33] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2368–2382, 2011. 2

[34] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 3

[35] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2015. 3

[36] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2021. 3

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, pages 405–421. Springer, 2020. 1, 3

[39] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv.org*, 2201.05989, 2022. 8

[40] Armin Mustafa and Adrian Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 422–431, 2017. 3

[41] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4990–4999, 2017. 2

[42] 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. *arXiv.org*, 2112.00724, 2021. 5

[43] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 3

[44] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3

[45] Eshed Ohn-Bar, Aditya Prakash, Aseem Behl, Kashyap Chitta, and Andreas Geiger. Learning situational driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11296–11305, 2020. 2

[46] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021. 8

[47] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1742–1750, 2015. 3

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 3

[49] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1796–1804, 2015. 3

[50] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proceedings of the European Conference on Computer Vision*, pages 523–540. Springer, 2020. 3

[51] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015. 3

[52] Lorenzo Porzi, Samuel Rota Bulo, Aleksander Colovic, and Peter Kontschieder. Seamless scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8277–8286, 2019. 7

[53] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 20154–20166, 2020. 3

[54] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473, 2020. 3

[55] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020. 7, 8

[56] 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.org*, 2111.13260, 2021. 3

[57] 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. *arXiv.org*, 2102.11859, 2021. 2

[58] Jianxiong Xiao and Long Quan. Multiple view semantic segmentation for street view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 686–693, 2009. 3

[59] Jun Xie, Martin Kiefel, Ming-Ting Sun, and Andreas Geiger. Semantic instance annotation of street scenes by 3d to 2d label transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3688–3697, 2016. 1, 3, 6

[60] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems*, volume 33, pages 2492–2502, 2020. 3

[61] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 8

[62] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017. 2, 6, 7, 8

[63] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 3, 6

[64] Brady Zhou, Philipp Krähenbühl, and Vladlen Koltun. Does computer vision matter for action? *Science Robotics*, 4(30), 2019. 2