Multi-view Consistent 3D Panoptic Scene Understanding

Xianzhu Liu[1], Xin Sun[1], Haozhe Xie[2], Zonglin Li[1]*, Ru Li[1], Shengping Zhang[1]

[1] Harbin Institute of Technology, Weihai, China
[2] Nanyang Technological University, Singapore

Abstract

3D panoptic scene understanding seeks to create novel view images with 3D-consistent panoptic segmentation, which is crucial for many vision and robotics applications. Mainstream methods (*e.g.,* Panoptic Lifting) directly use machine-generated 2D panoptic segmentation masks as training labels. However, these generated masks often exhibit multi-view inconsistencies, leading to ambiguities during the optimization process. To address this, we present Multi-view Consistent 3D Panoptic Scene Understanding (MVC-PSU), featuring two key components: 1) Probabilistic Semantic Aligner, which associates semantic information of corresponding pixels across multiple views by probabilistic alignment to ensure that the predicted panoptic segmentation masks are consistent across different views. 2) Geometric Consistency Enforcer, which uses multi-view projection and monocular depth consistency to ensure that the geometry of the reconstructed scene is accurate and consistent across different views. Experimental results demonstrate that the proposed MVC-PSU surpasses state-of-the-art methods on the ScanNet, Replica, and HyperSim datasets.

Introduction

3D panoptic scene understanding refers to the ability of computer systems to recognize both categorical "stuff" regions and individual "thing" instances within 3D visual scenes. This capability supports a range of applications (Siddiqui et al. 2023; Hui et al. 2023; Xie et al. 2024; Hui et al. 2024), such as augmented reality, virtual reality, robot navigation, and self-driving.

Over the past few years, there has been extensive research on understanding 3D scenes. Early approaches (Xie et al. 2021; Miao et al. 2022; Mei et al. 2022) address this challenge by converting it into image or video instance segmentation, typically using pre-trained models for 2D scene understanding to predict panoptic segmentation throughout entire videos. However, image and video segmentation often face multi-view inconsistency issues, where the instance labels for the same object differ across various views. Mainstream approaches (Zhou et al. 2021; Xu et al. 2022; Su et al. 2023) have shifted towards point cloud panoptic segmentation, but these methods depend on manually labeled datasets, which are costly and limited in scope, or they require accurately scanned point clouds as inputs. Recent advancements (Liu et al. 2023; Siddiqui et al. 2023; Bhalgat et al. 2023) focus on performing semantic or panoptic segmentation by extending 2D estimated masks by off-the-shelf segmentation models to 3D in a neural field manner without

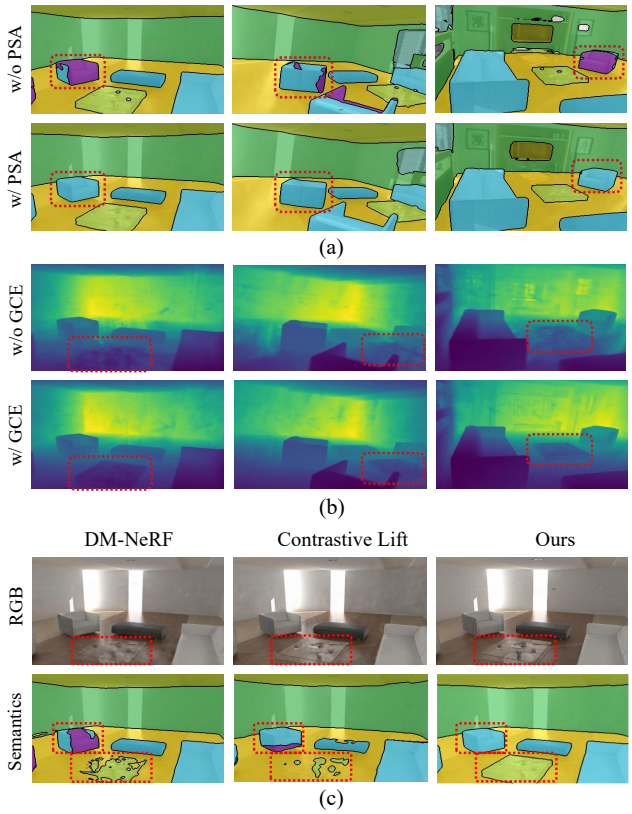

Figure 1: (a) The rendered semantics are consistent and accurate across multiple novel views with Probabilistic Semantic Aligner (PSA). (b) The rendered depth maps across multiple novel views have more accurate geometric details with Geometric Consistency Enforcer (GCE). (c) Qualitative comparison on novel views with DM-NeRF (Bing et al. 2023) and Contrastive Lift (Bhalgat et al. 2023).

*Corresponding Author

manual 3D annotation. However, these methods are trained directly using machine-generated 2D panoptic segmentation masks as labels, which are inconsistent across multiple views, leading to ambiguities during NeRF optimization.

In this paper, we propose Multi-view Consistent 3D Panoptic Scene Understanding (MVC-PSU) to address the challenge of ambiguity and enable rendering 3D-consistent semantics, instance, color, and depth information for novel views. Unlike existing methods (Siddiqui et al. 2023; Bhalgat et al. 2023) that directly fit the 3D panoptic radiance field from 2D posed images and machine-generated panoptic segmentation masks, MVC-PSU further incorporates Probabilistic Semantic Aligner (PSA) and Geometric Consistency Enforcer (GCE) to enforce multi-view consistency, ensuring that semantic and geometric information remains coherent and aligned across different views. Since machine-generated panoptic segmentation masks often exhibit inconsistencies and errors across different views, PSA aligns the semantic information of corresponding pixels from multiple views by probabilistic alignment. This process helps ensure consistency between the input masks and the predicted panoptic segmentation masks. As a result, the optimized scene maintains global semantic consistency, removing ambiguities and errors across various novel views, as illustrated in Figure 1(a). In addition, we observe that more precise geometry facilitates both color and panoptic segmentation prediction in the panoptic radiance field. To achieve this, GCE uses multi-view projection and monocular depth consistency to ensure that the geometry of the reconstructed scene is both accurate and consistent across various views. This improves the quality of the geometric structures, as demonstrated in Figure 1(b). Overall, as demonstrated in Figure 1 (c), the multi-view consistency enforced by jointly PSA and GCE significantly improves the accuracy of the rendered color and segmentation, leading to more accurate and consistent 3D panoptic scene understanding.

The main contributions can be summarized as:

- We propose Probabilistic Semantic Aligner (PSA) to ensure that predicted panoptic segmentation masks are consistent across different views, reducing ambiguities and errors caused by inconsistent machine-generated labels.
- We propose Geometric Consistency Enforcer (GCE) to ensure that the geometry of the reconstructed scene is accurate and consistent across different views, ultimately enhancing the quality of panoptic segmentation.
- Experimental results indicate that the proposed MVC-PSU outperforms existing state-of-the-art methods on the ScanNet, Replica, and HyperSim datasets.

Related Works

Semantic Neural Rendering. Initially, NeRF (Mildenhall et al. 2020) provides low-level representations of appearance and 3D geometry but lacks higher-level understanding of scenes, such as semantics and object centers. Semantic-NeRF (Zhi et al. 2021) is a pioneering method that introduces semantic branches into NeRF to predict semantic labels for any 3D locations, enabling new view synthesis of semantic masks. Further, DM-NeRF (Bing et al. 2023) learns object decomposition and manipulation of scenes by using MLPs to decode spatial locations into object identity vectors. Additionally, a series of works (Kundu et al. 2022; Fu et al. 2022, 2023; Zhang et al. 2023; Lin 2024) model each instance by separate MLPs, thus enabling them to handle object perception in dynamic scenes. Note that these methods rely on ground-truth 2D or 3D labels of the target scene, and noisy labels may significantly affect their performance. To circumvent the need for expensive ground-truth labels, Panoptic-Lifting (Siddiqui et al. 2023) uses noisy panoptic segmentation masks predicted by the pre-trained Mask2former (Cheng et al. 2022) for supervision. It adopts segmentation-consistency loss, bounded segmentation fields, and gradient stopping to robustly handle noisy labels. Similarly, Instance-NeRF (Liu et al. 2023) leverages Mask2former (Cheng et al. 2022) and CascadePSP (Cheng et al. 2020) to match the same instances of 2D segmentation across different views and optimize the generated masks, thereby continuously encoding 3D instance information in the form of neural fields.

Panoptic Segmentation. The task of panoptic segmentation is first introduced by Kirillov et al. (2019), which aims to provide a unified understanding of object instances (thing) and semantic regions (stuff) in images. Inspired by it, UPSNet (Xiong et al. 2019) integrates panoptic segmentation into a single network with a novel panoptic head and a parameter-free panoptic merging module, improving the overall performance and efficiency. Subsequent methods (Xiong et al. 2019; Zhang et al. 2021; Ren et al. 2021; Hu et al. 2023) have improved efficiency and performance through innovations in network architectures and multi-scale feature integration. Extending panoptic segmentation to 3D data, such as point clouds or volumetric data, is critical for applications in autonomous driving and robotics. Early efforts (Lahoud et al. 2019; Narita et al. 2019; Dahnert et al. 2021; Xu et al. 2022) mainly focus on volumetric and point cloud data, using volumetric fusion and voxel-based representations to achieve semantic and instance segmentation in indoor environments. Subsequent approaches (Engelmann et al. 2020; Gasperini et al. 2021; Sirohi et al. 2021; Zhou et al. 2021; Chen et al. 2021) continue to improve 3D panoptic segmentation by introducing innovative representations and aggregation mechanisms. Recently, NeRF-based methods (Liu et al. 2023; Siddiqui et al. 2023; Bhalgat et al. 2023; Zhang et al. 2024) focus on performing panoptic segmentation by extending 2D estimated masks by off-the-shelf segmentation models to 3D in a neural field manner.

Our Approach

Given multi-view posed images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ and machine-generated inconsistent panoptic segmentation masks of a scene, our goal is to learn a 3D panoptic radiance field that simultaneously renders 3D-consistent semantics, instance, color, and depth information for novel views. Figure 2 illustrates the proposed MVC-PSU, which incorporates a probabilistic semantic aligner and a geometric consistency enforcer to enforce multi-view consistency, ensuring that semantic and geometric information remains coherent and aligned across different views.

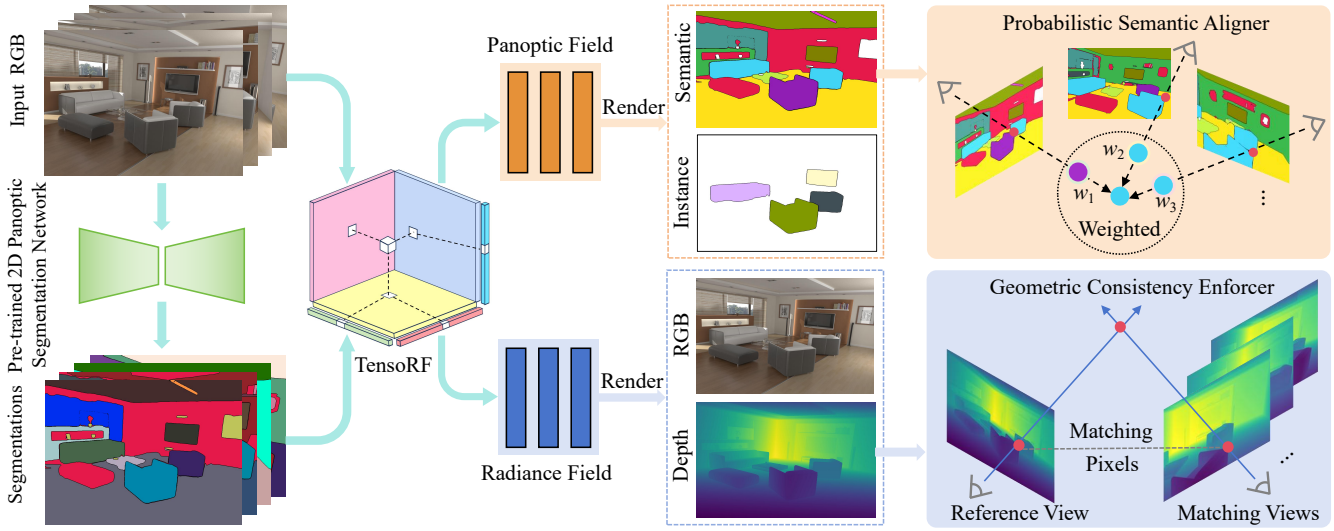

Figure 2: An overview of the proposed MVC-PSU. Taking multi-view posed images and machine-generated 2D panoptic segmentation masks as inputs, it learns a 3D panoptic radiance field that simultaneously renders 3D-consistent semantics, instance, color, and depth information for novel views. To reduce ambiguities and errors caused by inconsistent machine-generated labels, we introduce Probabilistic Semantic Aligner (PSA) to ensure that predicted panoptic segmentation masks are consistent across different views. To improve the quality of the radiance field, we introduce Geometric Consistency Enforcer (GCE) to ensure that the geometry of the reconstructed scene is accurate and consistent across different views.

Scene Representation

For a fair comparison, we choose TensoRF (Chen et al. 2022) to model the geometry and appearance of the scene and use two small MLPs to model semantics and instances as in (Bhalgat et al. 2023; Siddiqui et al. 2023).

Neural Radiance Field. TensoRF (Chen et al. 2022) represents the radiation fields as an explicit voxel grid of features. Specifically, it uses a geometry grid $\mathcal{G}_\sigma$ and an appearance grid $\mathcal{G}_c$ with multi-channel features per voxel to model the volume density σ and view-dependent color $\mathbf{c} = (r, g, b)$, respectively. By using the vector-matrix decomposition, $\mathcal{G}_\sigma$ and $\mathcal{G}_c$ are decomposed into compact components. The density tensor can be factorized as

$$\begin{aligned}\mathcal{G}_\sigma &= \sum_k \mathbf{v}^X_{\sigma,k} \circ \mathbf{M}^{YZ}_{\sigma,k} + \mathbf{v}^Y_{\sigma,k} \circ \mathbf{M}^{XZ}_{\sigma,k} + \mathbf{v}^Z_{\sigma,k} \circ \mathbf{M}^{XY}_{\sigma,k} \\ &= \sum_k \sum_{m \in XYZ} \mathbf{v}^m_{\sigma,k} \circ \mathbf{M}^{\tilde{m}}_{\sigma,k}\end{aligned} \tag{1}$$

where $\mathbf{v}^m_{\sigma,k}$ and $\mathbf{M}^{\tilde{m}}_{\sigma,k}$ are the k^{th} vector and matrix factors along the corresponding spatial axes m. Noted, $\tilde{m}$ represents the two axes orthogonal to m (e.g. $\tilde{X} = YZ$). The appearance tensor has an additional feature basis vector $\mathbf{b}$ corresponding to the feature channel dimension

$$\mathcal{G}_c = \sum_k \sum_{m \in XYZ} \mathbf{v}^m_{c,k} \circ \mathbf{M}^{\tilde{m}}_{c,k} \circ \mathbf{b}^m_k \tag{2}$$

The volume density σ can be directly obtained by linear interpolation of $\mathcal{G}_\sigma$. The appearance features $\mathcal{G}_c$ can be converted to color $\mathbf{c}$ through a small MLPs function S. In particular, given a 3D point $\mathbf{x} = (x, y, z)$ and ray direction $\mathbf{d} = (\theta, \phi)$, the corresponding volume density and color are

$$\sigma_\mathbf{x} = \mathcal{G}_\sigma(\mathbf{x}) \tag{3}$$

$$\mathbf{c}_\mathbf{x} = S(\mathcal{G}_c(\mathbf{x}), \mathbf{d}) \tag{4}$$

where $G_\sigma(\mathbf{x})$ and $G_c(\mathbf{x})$ are the density and multi-channel appearance features calculated by linear interpolation.

Volumetric Rendering. To render a pixel, a ray $\mathbf{r}$ is cast from the camera center $\mathbf{o}$ through the pixel along the view direction $\mathbf{d}$. Then, K points $\{\mathbf{x}_k = \mathbf{o} + t_k\mathbf{d}\}_{k=1}^K$ are sampled along the ray and fed into $\mathcal{G}_\sigma$ and $\mathcal{G}_c$ to query density and color $\{(\sigma_k, \mathbf{c}_k)\}_{k=1}^K$. Finally, the pixel color $\hat{\mathbf{c}}$ and depth value $\hat{d}$ for the ray $\mathbf{r}$ are rendered using numerical integration as

$$\hat{\mathbf{c}} = \sum_{k=1}^K T_k(1 - \exp(\sigma_k\delta_k))\mathbf{c}_k \tag{5}$$

$$\hat{d} = \sum_{k=1}^K T_k(1 - \exp(\sigma_k\delta_k))t_k \tag{6}$$

where $T_k = \exp\left(-\sum_{k=1}^{K-1} \sigma_k\delta_k\right)$ is the accumulated transmittance and $\delta_k = t_k - t_{k-1}$ represents the distance between adjacent sampling points.

Following (Siddiqui et al. 2023; Bhalgat et al. 2023), we further design two MLPs for learning the semantic and instance fields, which are formalized as view-invariant functions that map 3D spatial coordinates to semantic and instance distributions. The semantic logit $\hat{\mathbf{s}}$ for the ray $\mathbf{r}$ is calculated by volume rendering as

$$\hat{\mathbf{s}} = \sum_{k=1}^K T_k(1 - \exp(\sigma_k\delta_k))f(\mathbf{x}_k) \tag{7}$$

where $f(\cdot)$ represents the learned MLPs for the semantic distributions. Similarly, we can obtain the instance logit.

Probabilistic Semantic Aligner

To reduce ambiguities and errors caused by inconsistent machine-generated labels, we introduce the probabilistic semantic aligner to enforce the optimized scenes to be semantically globally consistent. Specifically, we first use a pre-trained dense matching model for correspondence generation, which infers pixel correspondences between training views. Based on this explicit connection between views, we further design multi-view semantic consistency to encourage the panoptic radiance field to generate consistent semantic masks for the same object across different views.

Correspondence Generation. Using pixel correspondences between training views as priors is widely applicable and cheap. As long as there are enough regions of texture overlap in the image pairs, any classical or learned matching method can estimate pixel-to-pixel correspondences. In practice, we use a pre-trained dense correspondence regression network PDC-Net (Truong et al. 2021), which predicts a matched point with a confidence score for each pixel based on an input image pair. In particular, we predict M matching relations $\{(\mathbf{p}_m, w_m)\}_{m=1}^{M}$ for each pixel $\mathbf{p}_a$ in the image $\mathbf{I}_i$, where $\mathbf{p}_m$ is the matched pixel in other training views and w_m is the corresponding confidence score.

Multi-view Semantic Consistency. Given the pixel $\mathbf{p}_a \in \mathbf{I}_i$, a pre-trained 2D segmentation network is first applied to generate corresponding probability distribution $\mathbf{s}_a$ over the semantic classes with a prediction confidence score c_a. Similarly, we can obtain the probability distributions and confidence scores $\{(\mathbf{s}_m, c_m)\}_{m=1}^{M}$ for the matched pixels. As shown in Figure 3, we show an example of warping two matched images (rows 2 and 3) towards the target image (view i) based on the predicted matching relations.

Then, we compute the average machine-generated semantic information $\mathbf{s}_{\text{avg}}$ among $\mathbf{s}_a$ and the matched probability distributions $\{\mathbf{s}_m\}_{m=1}^{M}$

$$\mathbf{s}_{\text{avg}} = \frac{\sum_{m=1}^{M} \mathbf{s}_m c_m w_m + \mathbf{s}_a c_a}{\sum_{m=1}^{M} c_m w_m + c_a} \tag{8}$$

$\mathbf{s}_{\text{avg}}$ is a robust semantic label obtained by weighted averaging the semantic information from multiple perspectives, thereby avoiding the influence of a few erroneous perspectives. According to Eq. 7, the semantic logic $\hat{\mathbf{s}}_a$ of pixel $\mathbf{p}_a$ and the semantic logics $\{\hat{\mathbf{s}}_m\}_{m=1}^{M}$ of its matched pixel can be rendered. We further calculate the average rendering semantic information $\hat{\mathbf{s}}_{\text{avg}}$

$$\hat{\mathbf{s}}_{\text{avg}} = \frac{\sum_{m=1}^{M} \hat{\mathbf{s}}_m w_m + \hat{\mathbf{s}}_a}{\sum_{m=1}^{M} w_m + 1} \tag{9}$$

$\hat{\mathbf{s}}_{\text{avg}}$ can be regarded as the most probable predicted semantic category for pixel $\mathbf{p}_a$ and all matched pixels. To optimize the panoptic field, we utilize two Cross Entropy (CE) losses, each assessing the difference between the rendered semantic logits $\{\hat{\mathbf{s}}_m\}_{m=0}^{M}$ and either $\mathbf{s}_{\text{avg}}$ or $\hat{\mathbf{s}}_{\text{avg}}$

$$\mathcal{L}_{\text{semantic}}^{\text{mv}} = \frac{1}{M+1}\Big(\sum_{m=0}^{M} \text{CE}(\hat{\mathbf{s}}_m, \mathbf{s}_{\text{avg}}) + \sum_{m=0}^{M} \text{CE}(\hat{\mathbf{s}}_m, \hat{\mathbf{s}}_{\text{avg}})\Big) \tag{10}$$

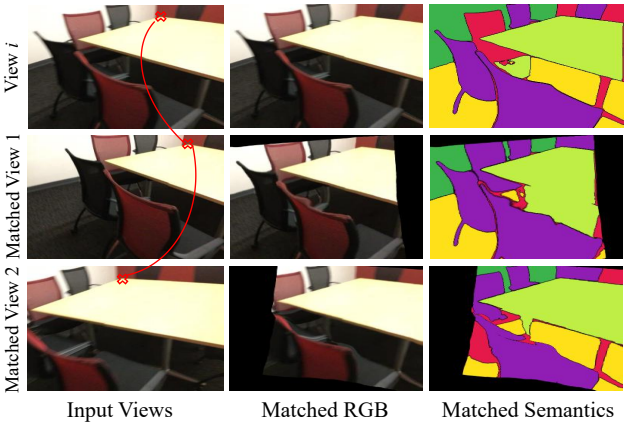

Figure 3: An example of dense matches predicted by PDC-Net (Truong et al. 2021). PDC-Net predicts the dense matching relations between the target (view i) and matched images. In columns 2 and 3, we show two warped RGB and semantics towards the target, respectively.

where $\hat{\mathbf{s}}_0 = \hat{\mathbf{s}}_a$. By aligning semantic information across views, we optimize a globally consistent panoptic field to reduce the ambiguity introduced by machine-generated semantic labels.

Geometric Consistency Enforcer

We observe that more precise geometry facilitates both color and panoptic segmentation prediction in the panoptic radiance field. To achieve this, we further introduce the geometric consistency enforcer, which uses multi-view depth consistency and monocular depth constraint to ensure that the geometry of the reconstructed scene is both accurate and consistent across various views.

Multi-view Depth Consistency. First, we encourage the reconstructed 3D geometry to align well with the depth rendered from multiple views by enforcing the geometric consistency of corresponding points in different views. Specifically, given the pixel $\mathbf{p}_a \in \mathbf{I}_i$, we can compute the depth value $\hat{d}_a$ according to the Eq. 6. Similarly, we can obtain the depth values $\{\hat{d}_m\}_{m=1}^{M}$ for the matched pixels $\{\mathbf{p}_m\}_{m=1}^{M}$. Let $\mathbf{P}_i$ and $\mathbf{K}_i$ be the world-to-camera transform and intrinsic matrix for image $\mathbf{I}_i$, respectively. Using the rendered depth value $\hat{d}_a$ of pixel $\mathbf{p}_a$, the corresponding 3D point $\mathbf{x}_a$ in world coordinates can be derived as $\mathbf{x}_a = \mathbf{P}_i^{-1}\mathbf{K}_i^{-1}(\overline{\mathbf{p}}_a \cdot \hat{d}_a)$, where $\overline{\mathbf{p}}_a$ corresponds to the homogeneous representation of $\mathbf{p}_a$. The 3D point $\mathbf{x}_a$ is then projected into the view of the matched pixel to obtain the corresponding projection depth, i.e., $\overline{\mathbf{p}}_{am} \cdot \hat{d}_{am} = \mathbf{K}_m \mathbf{P}_m \mathbf{x}_a$, where $\overline{\mathbf{p}}_{am}$ and $\hat{d}_{am}$ are the matched pixel coordinate and depth obtained by projection, $\mathbf{K}_m$ and $\mathbf{P}_m$ are the intrinsic matrix and world-to-camera transform of $\mathbf{p}_m$. Theoretically, the projected depth $\hat{d}_{am}$ should be consistent with the rendered depth $\hat{d}_m$. By applying this depth constraint to all the matched pixels, we

Methods	ScanNet			Replica			HyperSim		
	mIoU↑	PQ^{scene} ↑	PSNR↑	mIoU↑	PQ^{scene} ↑	PSNR↑	mIoU↑	PQ^{scene} ↑	PSNR↑
SemanticNeRF (Zhi et al. 2021)	59.2	–	26.6	58.5	–	24.8	58.9	–	26.6
Mask2Former (Cheng et al. 2022)	46.7	–	–	52.4	–	–	53.9	–	–
PNF (Kundu et al. 2022)	53.9	48.3	26.7	51.5	41.1	29.8	50.3	44.8	27.4
DM-NeRF (Bing et al. 2023)	49.5	41.7	27.5	56.0	44.1	26.9	57.6	51.6	28.1
Panoptic Lifting (Siddiqui et al. 2023)	65.2	58.9	28.5	67.2	57.9	29.6	67.8	60.1	30.1
Contrastive Lift (Bhalgat et al. 2023)	65.2	62.3	28.3	67.0	59.1	29.3	67.9	62.3	30.0
MVC-PSU (Ours)	**68.6**	**63.1**	**28.9**	**68.7**	**59.4**	**31.1**	**69.8**	**62.7**	**30.8**

Table 1: Quantitative comparison on the ScanNet, Replica, and HyperSim datasets, measured by mIoU, PQ^{scene}, and PSNR. The best results are highlighted in bold.

design the multi-view depth consistency as

$$\mathcal{L}_{\text{depth}}^{\text{mv}} = \frac{1}{M}\sum_{m=1}^{M} ||\hat{d}_{am} - \hat{d}_m||_2^2 \tag{11}$$

where $||\cdot||_2^2$ represents the mean square error. By minimizing the differences between the projected depth and rendered depth, we align the reconstructed 3D geometry with depth maps rendered from different views, which helps to accurately reconstruct the scene geometry.

Monocular Depth Constraint. Inspired by (Deng et al. 2022; Chung, Oh, and Lee 2024), we further use the depth maps predicted from a pre-trained monocular depth estimator as additional supervision to encourage geometric consistency within the view. Specifically, the pre-trained Dense Prediction Transformer (DPT) (Ranftl et al. 2021) is first used to generate the monocular depth map $\mathbf{D}$ for each training view $\mathbf{I}$. Subsequently, to alleviate the constrain posed by inconsistencies in absolute depth values, a softened depth constraint based on pearson correlation (Zhu et al. 2023) is introduced, which mitigates the scale ambiguity between the rendered depth map $\hat{\mathbf{D}}$ and the estimated depth map

$$\mathcal{L}_{\text{depth}}^{\text{mo}} = \frac{\text{Cov}(\hat{\mathbf{D}}, \mathbf{D})}{\sqrt{\text{Var}(\hat{\mathbf{D}})\,\text{Var}(\mathbf{D})}} \tag{12}$$

Although the absolute depth scale predicted by the DPT model is inaccurate, this relative relationship contains a relatively accurate 3D consistency that can regulate the optimization of the radiance field.

Optimization

The overall loss can be formulated as

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{color}} + \lambda_{\text{dep}}(\mathcal{L}_{\text{depth}}^{\text{mv}} + \mathcal{L}_{\text{depth}}^{\text{mo}}) + \lambda_{\text{sem}}\mathcal{L}_{\text{semantic}}^{\text{mv}} \\ +\lambda_{\text{int}}\mathcal{L}_{\text{instance}} + \lambda_{\text{seg}}\mathcal{L}_{\text{segment}}\end{aligned} \tag{13}$$

where λ_{dep}, λ_{sem}, λ_{int}, and λ_{seg} are scalar hyperparameters, which control the weights of different parts. In the experiments, we set $\lambda_{\text{dep}} = 1$, while λ_{sem}, λ_{int}, and λ_{seg} are each set to 0.1. $\mathcal{L}_{\text{instance}}$ and $\mathcal{L}_{\text{segment}}$ are the instance segmentation loss and segment consistency loss proposed in Panoptic Lifting (Siddiqui et al. 2023), respectively. $\mathcal{L}_{\text{color}}$ is used to train the radiance field by minimizing the photometric mean square error between the ground truth pixel color $\mathbf{c}$ and the rendered color $\hat{\mathbf{c}}$

$$\mathcal{L}_{\text{color}} = ||\mathbf{c} - \hat{\mathbf{c}}||_2^2 \tag{14}$$

Experiments

Datasets and Metrics

Datasets. Following Panoptic Lifting (Siddiqui et al. 2023), we conduct experiments on three public datasets: ScanNet (Dai et al. 2017), Replica (Straub et al. 2019), and Hypersim (Roberts et al. 2021). Ground truth poses provided by each dataset are used, while ground truth semantic and instance labels are employed solely for evaluation and not for training or model refinement. For a fair comparison, we follow the experimental settings of Panoptic Lifting (Siddiqui et al. 2023) and generate 2D panoptic segmentation masks by the pre-trained Mask2Former (Cheng et al. 2022) as training labels. All three datasets contain 21 categories (9 thing + 12 stuff), where the available posed images in each dataset are divided into 75% for training views and 25% for testing views sampled in between.

Metrics. We use the peak signal-to-noise ratio (PSNR) and the mean intersection over union (mIoU) to evaluate the realism of the synthesized novel views and the accuracy of semantic segmentation, respectively. Panoptic segmentation quality is measured with a scene-level Panoptic Quality (PQ^{scene}) metric (Siddiqui et al. 2023), which considers the consistency of instances across different views.

Main Results

We compare with other state-of-the-art methods on ScanNet, Replica, and Hypersim datasets. Following the data preprocessing steps of Panoptic Lifting (Siddiqui et al. 2023), all methods are trained using the same set of images and machine-produced 2D panoptic segmentation masks.

Quantitative Results. We report quantitative comparison results on the three datasets in Table 1, which show that the proposed MVC-PSU significantly outperforms other state-of-the-art methods. Moreover, although we use the same underlying TensoRF architecture as Panoptic Lifting (Siddiqui et al. 2023), our method improves PSNR by 0.6, 1.5, and 0.7 on ScanNet, Replica, and Hypersim datasets, respectively. We attribute this to the improved geometric quality of the

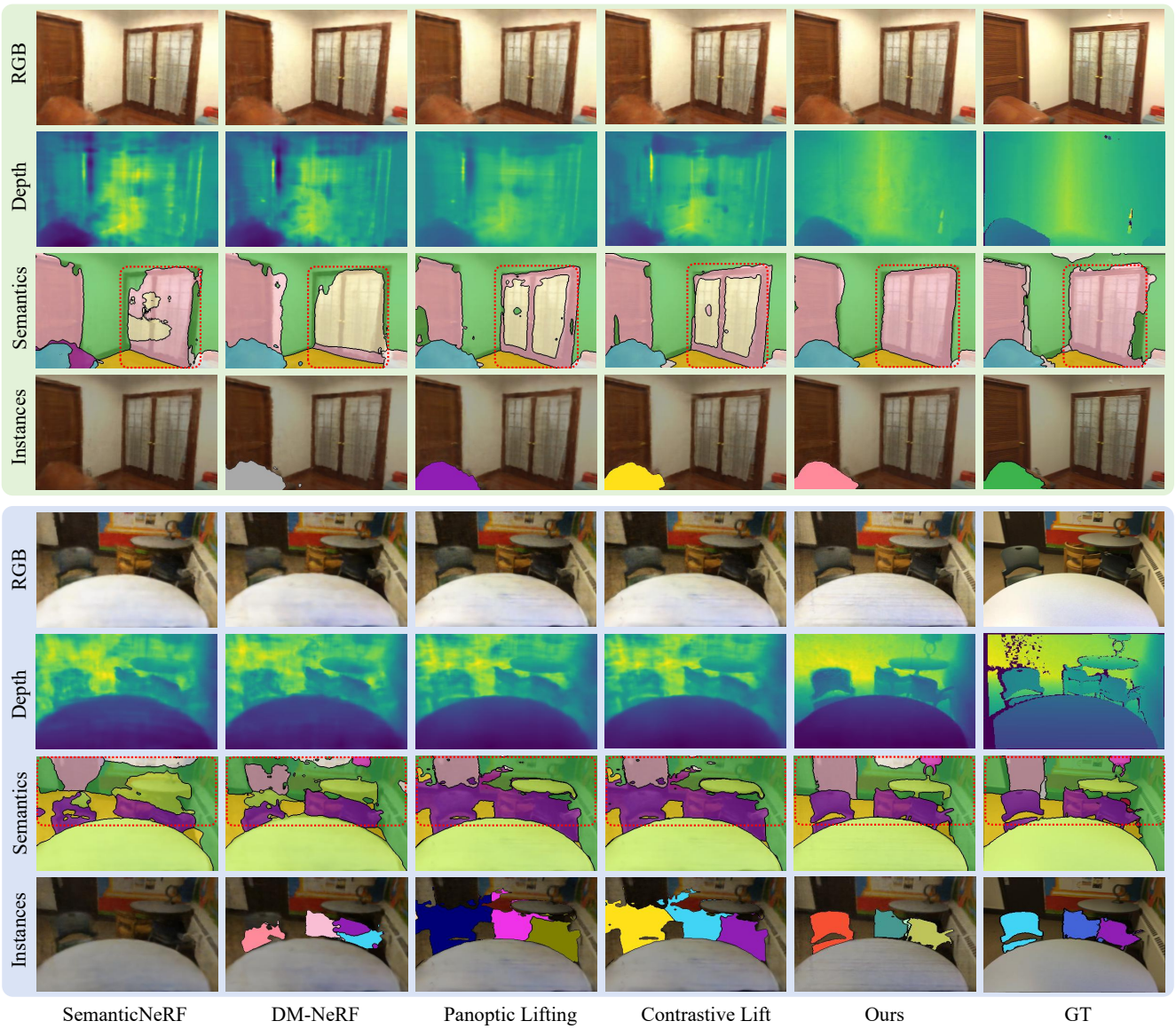

Figure 4: Qualitative comparison on the ScanNet dataset against four baseline methods. Note that SemanticNeRF (Zhi et al. 2021) does not predict 3D instance segmentation. Ground truth depth maps are sourced from depth cameras.

reconstructed radiance field achieved by the proposed geometric consistency, thus improving the performance of novel view synthesis. In addition, Table 1 also shows that compared to the second-ranked method Contrastive Lift (Bhalgat et al. 2023), the proposed method has significant improvements in terms of both mIoU and PQ^{scene} on all three datasets. In summary, our method can effectively improve the quality of panoptic segmentation and reconstruction by enforcing multi-view consistency.

Qualitative Results. Figure 4 shows the qualitative comparison results with SemanticNeRF (Zhi et al. 2021), DM-NeRF (Bing et al. 2023), Panoptic Lifting (Siddiqui et al. 2023), and Contrastive Lift (Bhalgat et al. 2023) on the ScanNet dataset. From the figure, we can see that the proposed method outperforms all compared methods in both semantic and instance segmentation tasks and achieves the best view synthesis quality. As shown in the third row of the figure, all the compared methods mistakenly identified the door as a window due to inaccurate machine-generated panoptic labels. Benefiting from the proposed probabilistic semantic aligner, our method can be more robust to the noise in machine-generated labels. Moreover, when faced with areas with dense objects and similar textures, all compared methods fail to capture the precise geometry of the objects, resulting in segmentation errors, as shown in the red box area in the seventh row of the figure. In contrast, the proposed geometric consistency enforcer brings significant gains in improving the geometric quality of the reconstructed radiance

Model	PSA	MVDC	MDC	mIoU↑	PQ^{scene} ↑	PSNR↑
A	×	×	×	64.9	58.8	28.3
B	×	✓	✓	66.3	60.6	**28.9**
C	✓	×	✓	68.1	62.7	28.6
D	✓	✓	×	67.9	62.5	28.5
E	✓	✓	✓	**68.6**	**63.1**	**28.9**

Table 2: Ablation studies of Probabilistic Semantic Aligner (PSA), Multi-view Depth Consistency (MVDC), and Monocular Depth Constraint (MDC) on the ScanNet dataset.

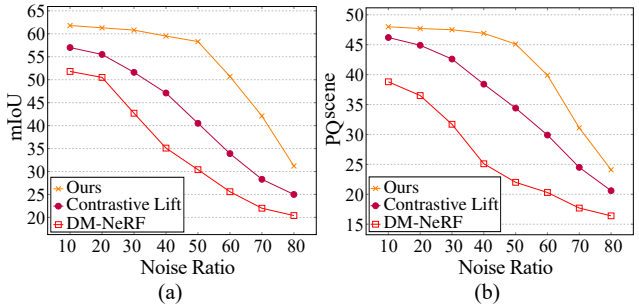

Figure 5: Quantitative comparison on ScanNet with varying label noise ratios: (a) mIoU results and (b) PQ^{scene} results for different methods.

field, thus boosting the performance of panoptic segmentation. As explained in (Yu et al. 2022; Lyu et al. 2023), using only RGB reconstruction loss may lead to under-constraint when faced with larger and more complex indoor scenes, especially in areas with few observations and similar textures.

Ablation Studies

Following Panoptic Lifting (Siddiqui et al. 2023), We conduct ablation studies to confirm the effectiveness of each component on the ScanNet dataset.

Effectiveness of Probabilistic Semantic Aligner. We first perform ablations of removing Probabilistic Semantic Aligner (PSA), as denoted by Model B in Table 2. Removing the PSA component reduces the mIoU and PQ^{scene} by 2.3 and 2.5, which indicates that it can effectively boost the performance of panoptic segmentation. This improvement is attributed to PSA ensuring that the optimized scene is globally semantically consistent, thereby reducing ambiguities and errors caused by inconsistent 2D labels across views generated by off-the-shelf segmentation models.

Effectiveness of Geometric Consistency Enforcer. The geometric consistency enforcer uses Multi-view Depth Consistency (MVDC) and Monocular Depth Constraint (MDC) to improve the geometric quality of the reconstructed radiance field. As denoted by Models C and D in Table 2, removing either the MVDC or MDC component results in a decrease in performance. In particular, as denoted by Model A, removing both components leads to a significant decrease in PSNR, demonstrating that they can improve the performance of panoptic scene understanding by improving the geometric quality of the reconstructed radiance field.

#M.R.	2	4	6	8	10	12	14
mIoU↑	66.4	66.7	67.6	68.3	68.6	**68.8**	**68.8**
PQ^{scene} ↑	60.9	61.5	62.3	62.9	63.1	63.1	**63.2**
PSNR↑	28.6	28.7	28.7	28.8	**28.9**	**28.9**	**28.9**
Time (ms)	**12.3**	22.7	34.8	47.2	60.5	75.4	90.8

Table 3: Impact of the number of matching relations on the ScanNet dataset.

Impact of the Number of Matching Relations. As mentioned above, we predict M matching relations for pixel correspondences in the training views. In this ablation study, we analyze the impact of different M values on the performance. Table 3 shows the corresponding results on the ScanNet dataset, which indicates that mIoU, PQ^{scene}, and PSNR all improve as the number of matches increases. In addition, the time required to train one iteration also increases significantly. To achieve a better trade-off between complexity and performance, we set the number of matches to 10.

Robustness to Label Noise. To further verify the robustness of Probabilistic Semantic Aligner (PSA) to inconsistent 2D machine-generated labels across views, we conduct experiments on the segmentation labels with various noise ratios. Specifically, we manually randomly change the semantic labels of pixels in M matching relations of all training views by p percentage points, where p ranges from 10% to 80% with a step size of 10%. We choose scene0050_02 in the ScanNet dataset as the test scene and the comparison results with other competitive methods are shown in Figure 5. The proposed method achieves the best mIoU and PQ^{scene} under all label noise rates, which indicates that the PSA module is more robust in dealing with labels with different inconsistency levels. In addition, benefiting from aligning semantic information across views, the performance of our method is stable when the label noise rate is less than 50%, while the performance of competing methods drops significantly.

Conclusion

In this paper, we propose Multi-view Consistent 3D Panoptic Scene Understanding (MVC-PSU) to address the challenge of ambiguity and enable rendering 3D-consistent semantics, instance, color, and depth information for novel views. Firstly, we propose Probabilistic Semantic Aligner (PSA) to associate the semantic information of corresponding pixels across multiple views through probabilistic alignment to ensure that the predicted panoptic segmentation masks are consistent across different views. Additionally, we introduce Geometric Consistency Enforcer (GCE) to ensure that the geometry of the reconstructed scene is accurate and consistent across different views by correcting discrepancies in the depth information. Extensive experiments on the ScanNet, Replica, and HyperSim datasets demonstrate that MVC-PSU outperforms existing state-of-the-art methods, validating the effectiveness of PSA and GCE in achieving consistent and accurate 3D scene understanding.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62272134, 62072141, and 62402136, in part by the National Natural Science Foundation of Shandong Province under Grant ZR2024QF064.

References

Bhalgat, Y.; Laina, I.; Henriques, J. F.; Zisserman, A.; and Vedaldi, A. 2023. Contrastive Lift: 3D object instance segmentation by slow-fast contrastive fusion. In *Advances in Neural Information Processing Systems*.

Bing, W.; et al. 2023. DM-NeRF: 3D scene geometry decomposition and manipulation from 2D images. In *International Conference on Learning Representations*.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022. TensoRF: Tensorial radiance fields. In *European Conference on Computer Vision*, 333–350.

Chen, S.; Fang, J.; Zhang, Q.; Liu, W.; and Wang, X. 2021. Hierarchical aggregation for 3D instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, 15467–15476.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Cheng, H. K.; Chung, J.; Tai, Y.-W.; and Tang, C.-K. 2020. CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8890–8899.

Chung, J.; Oh, J.; and Lee, K. M. 2024. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 811–820.

Dahnert, M.; Hou, J.; Nießner, M.; and Dai, A. 2021. Panoptic 3D scene reconstruction from a single RGB image. In *Advances in Neural Information Processing Systems*, volume 34, 8282–8293.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5828–5839.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE Conference on Computer Vision and Pattern Recognition*, 12882–12891.

Engelmann, F.; Bokeloh, M.; Fathi, A.; Leibe, B.; and Nießner, M. 2020. 3D-MPA: Multi-proposal aggregation for 3D semantic instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9031–9040.

Fu, X.; Zhang, S.; Chen, T.; Lu, Y.; Zhou, X.; Geiger, A.; and Liao, Y. 2023. PanopticNeRF-360: Panoramic 3D-to-2D label transfer in urban scenes. *arXiv preprint arXiv:2309.10815*.

Fu, X.; Zhang, S.; Chen, T.; Lu, Y.; Zhu, L.; Zhou, X.; Geiger, A.; and Liao, Y. 2022. Panoptic NeRF: 3D-to-2D label transfer for panoptic urban scene segmentation. In *International Conference on 3D Vision*, 1–11.

Gasperini, S.; Mahani, M.-A. N.; Marcos-Ramiro, A.; Navab, N.; and Tombari, F. 2021. Panoster: End-to-end panoptic segmentation of lidar point clouds. *IEEE Robotics and Automation Letters*, 6(2): 3216–3223.

Hu, J.; Huang, L.; Ren, T.; Zhang, S.; Ji, R.; and Cao, L. 2023. You only segment once: Towards real-time panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 17819–17829.

Hui, C.; Zhang, S.; Cui, W.; Liu, S.; Jiang, F.; and Zhao, D. 2023. Rate-Adaptive Neural Network for Image Compressive Sensing. *IEEE Transactions on Multimedia*, 26: 2515–2530.

Hui, C.; Zhu, H.; Yan, S.; Liu, S.; Jiang, F.; and Zhao, D. 2024. S^2-CSNet: Scale-Aware Scalable Sampling Network for Image Compressive Sensing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2515–2524.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019. Panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9404–9413.

Kundu, A.; Genova, K.; Yin, X.; Fathi, A.; Pantofaru, C.; Guibas, L. J.; Tagliasacchi, A.; Dellaert, F.; and Funkhouser, T. 2022. Panoptic neural fields: A semantic object-aware neural scene representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 12871–12881.

Lahoud, J.; Ghanem, B.; Pollefeys, M.; and Oswald, M. R. 2019. 3D instance segmentation via multi-task metric learning. In *IEEE/CVF International Conference on Computer Vision*, 9256–9266.

Lin, Y. 2024. Ced-NeRF: A compact and efficient method for dynamic neural radiance fields. In *AAAI Conference on Artificial Intelligence*, 3504–3512.

Liu, Y.; Hu, B.; Huang, J.; Tai, Y.-W.; and Tang, C.-K. 2023. Instance neural radiance field. In *IEEE/CVF International Conference on Computer Vision*, 787–796.

Lyu, X.; Dai, P.; Li, Z.; Yan, D.; Lin, Y.; Peng, Y.; and Qi, X. 2023. Learning a room with the occ-sdf hybrid: Signed distance function mingled with occupancy aids scene representation. In *IEEE/CVF International Conference on Computer Vision*, 8940–8950.

Mei, J.; Zhu, A. Z.; Yan, X.; Yan, H.; Qiao, S.; Chen, L.-C.; and Kretzschmar, H. 2022. Waymo open dataset: Panoramic video panoptic segmentation. In *European Conference on Computer Vision*, 53–72.

Miao, J.; Wang, X.; Wu, Y.; Li, W.; Zhang, X.; Wei, Y.; and Yang, Y. 2022. Large-scale video panoptic segmentation in the wild: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, 21033–21043.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 405–421.

Narita, G.; Seno, T.; Ishikawa, T.; and Kaji, Y. 2019. PanopticFusion: Online volumetric semantic mapping at the level of stuff and things. In *International Conference on Intelligent Robots and Systems*, 4205–4212.

Ranftl, R.; et al. 2021. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision*, 12179–12188.

Ren, J.; Yu, C.; Cai, Z.; Zhang, M.; Chen, C.; Zhao, H.; Yi, S.; and Li, H. 2021. REFINE: Prediction fusion network for panoptic segmentation. In *AAAI Conference on Artificial Intelligence*, 2477–2485.

Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M. A.; Paczan, N.; Webb, R.; and Susskind, J. M. 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10912–10922.

Siddiqui, Y.; Porzi, L.; Bulò, S. R.; Müller, N.; Nießner, M.; Dai, A.; and Kontschieder, P. 2023. Panoptic lifting for 3D scene understanding with neural fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9043–9052.

Sirohi, K.; Mohan, R.; Büscher, D.; Burgard, W.; and Valada, A. 2021. Efficientlps: Efficient lidar panoptic segmentation. *IEEE Transactions on Robotics*, 38(3): 1894–1914.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

Su, S.; Xu, J.; Wang, H.; Miao, Z.; Zhan, X.; Hao, D.; and Li, X. 2023. PUPS: Point cloud unified panoptic segmentation. In *AAAI Conference on Artificial Intelligence*, 2339–2347.

Truong, P.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2021. Learning accurate dense correspondences and when to trust them. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5714–5724.

Xie, H.; Chen, Z.; Hong, F.; and Liu, Z. 2024. CityDreamer: Compositional generative model of unbounded 3D cities. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9666–9675.

Xie, H.; Yao, H.; Zhou, S.; Zhang, S.; and Sun, W. 2021. Efficient regional memory network for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1286–1295.

Xiong, Y.; Liao, R.; Zhao, H.; Hu, R.; Bai, M.; Yumer, E.; and Urtasun, R. 2019. UPSNet: A unified panoptic segmentation network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8818–8826.

Xu, S.; Wan, R.; Ye, M.; Zou, X.; and Cao, T. 2022. Sparse cross-scale attention network for efficient lidar panoptic segmentation. In *AAAI Conference on Artificial Intelligence*, 2920–2928.

Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. In *Advances in Neural Information Processing Systems*, volume 35, 25018–25032.

Zhang, G.; Gao, Y.; Xu, H.; Zhang, H.; Li, Z.; and Liang, X. 2021. Ada-Segment: Automated multi-loss adaptation for panoptic segmentation. In *AAAI Conference on Artificial Intelligence*, 3333–3341.

Zhang, X.; Kundu, A.; Funkhouser, T.; Guibas, L.; Su, H.; and Genova, K. 2023. Nerflets: Local radiance fields for efficient structure-aware 3D scene representation from 2D supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8274–8284.

Zhang, Y.; Chen, G.; Chen, J.; and Cui, S. 2024. Aerial Lifting: Neural Urban Semantic and Building Instance Lifting from Aerial Imagery. In *CVPR*, 21092–21103.

Zhi, S.; Laidlow, T.; Leutenegger, S.; and Davison, A. J. 2021. In-place scene labelling and understanding with implicit scene representation. In *IEEE/CVF International Conference on Computer Vision*, 15838–15847.

Zhou, Z.; et al. 2021. Panoptic-PolarNet: Proposal-free lidar point cloud panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 13194–13203.

Zhu, Z.; Fan, Z.; Jiang, Y.; and Wang, Z. 2023. FSGS: Real-time few-shot view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00451*.