# Aerial Lifting: Neural Urban Semantic and Building Instance Lifting from Aerial Imagery

Yuqi Zhang[1,2] Guanying Chen[1,3*] Jiaxing Chen[1,3] Shuguang Cui[2,1]
[1]FNii, CUHKSZ [2]SSE, CUHKSZ [3]Sun Yat-sen University

## Abstract

*We present a neural radiance field method for urban-scale semantic and building-level instance segmentation from aerial images by lifting noisy 2D labels to 3D. This is a challenging problem due to two primary reasons. Firstly, objects in urban aerial images exhibit substantial variations in size, including buildings, cars, and roads, which pose a significant challenge for accurate 2D segmentation. Secondly, the 2D labels generated by existing segmentation methods suffer from the multi-view inconsistency problem, especially in the case of aerial images, where each image captures only a small portion of the entire scene. To overcome these limitations, we first introduce a scale-adaptive semantic label fusion strategy that enhances the segmentation of objects of varying sizes by combining labels predicted from different altitudes, harnessing the novel-view synthesis capabilities of NeRF. We then introduce a novel cross-view instance label grouping based on the 3D scene representation to mitigate the multi-view inconsistency problem in the 2D instance labels. Furthermore, we exploit multi-view reconstructed depth priors to improve the geometric quality of the reconstructed radiance field, resulting in enhanced segmentation results. Experiments on multiple real-world urban-scale datasets demonstrate that our approach outperforms existing methods, highlighting its effectiveness. The source code is available at https://github.com/zyqz97/Aerial_lifting.*

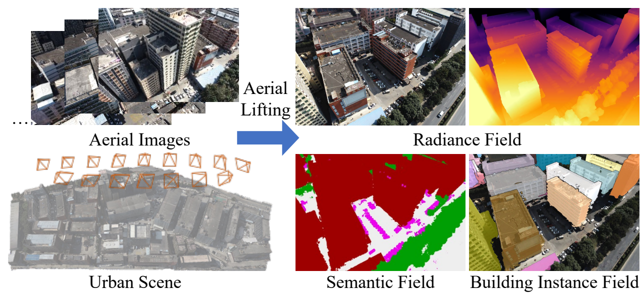



Figure 1. Given multi-view aerial images, our method lifts 2D labels to optimize the radiance, semantic, and instance fields for urban-scale semantic and building-level instance understanding.

## 1. Introduction

3D urban-scale semantic understanding plays a crucial role in various applications, from urban planning to autonomous driving systems. Accurate semantic and instance-level segmentation of objects in 3D scenes is essential for a wide range of tasks.

Existing 3D urban semantic understanding methods primarily rely on point cloud representation [34, 64]. They typically train a point cloud segmentation method on labeled 3D datasets [32]. However, annotating 3D data is labor-intensive, posing challenges in creating a comprehensive training dataset with diverse scenes.

Recently, neural radiance fields (NeRF) [57] have emerged as an effective 3D scene representation, enabling photorealistic rendering of fine details. Several methods are proposed to perform semantic segmentation or panoptic segmentation on NeRF by lifting 2D estimation to 3D [71, 101]. However, these methods mainly validate on the room-scale indoor scenes or street-view outdoor scenes. In this work, we aim to perform urban-scale semantic and building-level instance segmentation from multi-view aerial images. Our method leverages neural radiance fields to lift noisy 2D labels to a 3D representation without manual 3D annotations, effectively bridging the gap between 2D imagery and the complex 3D urban environment (see Fig. 1).

This is inherently challenging due to several factors. On one hand, urban aerial images capture scenes that encompass a wide range of object sizes, including buildings, vehicles, and roads [51]. Existing segmentation methods often struggle to handle these variations effectively as their training data distribution is different from that of aerial images [13], or lack large-scale labeled aerial images for fine-tuning [85]. On the other hand, 2D instance labels generated by existing segmentation methods often suffer from the multi-view inconsistency problem (*e.g.*, an object is segmented as one instance in a view might be segmented

*Corresponding author

into multiple independent instances in another view). This problem becomes particularly pronounced in the context of aerial images, where each image captures only a small portion of the entire scene. Furthermore, the geometry reconstruction quality of the large-scale scene will largely affect the semantic segmentation.

To address these problems, we introduce three key strategies to enhance the accuracy and robustness of our segmentation approach. First, we propose a *scale-adaptive semantic label fusion* strategy, enabling the segmentation of objects of varying sizes by fusing labels predicted from different altitudes. This leverages the novel-view synthesis capabilities of NeRF [57] to render photorealistic images at different altitudes. Second, we introduce a *cross-view instance label grouping* strategy to group instance labels in a view utilizing information from other views. It is achieved by performing cross-view label projection based on the relative camera poses and geometry of the 3D scene representation. This strategy effectively mitigates the multi-view inconsistency problem in the 2D instance labels, providing a more coherent and accurate segmentation of urban objects. Furthermore, we exploit *depth priors obtained from multi-view stereo* to improve the geometric quality of the reconstructed radiance field, ultimately leading to enhanced segmentation results. Our approach has been extensively evaluated on multiple real-world urban-scale scenes, demonstrating its superior performance compared to existing methods.

In summary, the key contributions are as follows:

- We present a novel radiance field approach for urban-scale semantic and building-level instance segmentation from aerial images by lifting noisy 2D labels to 3D, achieving state-of-the-art results.
- We introduce a scale-adaptive semantic label fusion strategy that combines 2D labels predicted from different altitudes to enhance the segmentation of objects of varying sizes, leveraging NeRF's novel-view synthesis capabilities.
- We present a cross-view instance label grouping approach based on the 3D scene representation to mitigate the multi-view inconsistency problem in 2D instance labels, resulting in more reliable instance segmentation results.

## 2. Related Work

**3D Urban Semantic Learning** Traditional 3D semantic learning methods involve training models on 3D datasets with ground-truth annotations [16, 27, 30, 33, 35, 42, 52, 65, 84, 92, 97]. These methods often operate on explicit representations such as point clouds [34, 64]. For 3D urban scenes, some methods perform 3D building instance segmentation from meshes [1, 6, 10] or point clouds [11, 59, 93]. Recent research shows that implicit representations [60, 62, 95] can effectively represent continuous and detailed surfaces and enable differentiable rendering, making them a promising choice for semantic understanding. In this work, we leverage the radiance field representation [57] and lift the estimated 2D labels to 3D through per-scene optimization.

**Neural Scene Representations** Traditional multi-view 3D reconstruction methods [2, 21, 44, 47, 72, 103] often apply structure-from-motion (SFM) techniques to estimate camera poses [68], followed by dense multi-view stereo [24, 25] to generate 3D models.

Recently, neural scene representations have achieved significant success in 3D scene modeling [70, 77–80, 90]. Specifically, neural radiance fields (NeRF) [57] achieve photorealistic rendering for diverse scenes. Subsequently, many methods have been proposed to enhance NeRF in various aspects, including surface geometry [61, 86, 94, 98] and optimization speed [9, 20, 36, 58, 74]. To handle large-scale scenes, several methods introduce sophisticated designs to improve rendering quality and reconstruction geometry [28, 45, 50, 53, 56, 87, 89, 91, 99, 100]. For example, Block-NeRF [76] and Mega-NeRF [81] decompose the scene into several partitions, with each partition represented by a different local NeRF. StreetSurf [28] proposed a variant of the hash-grid [58], which allocates the grid space according to the ratios of the three axes, making full use of the grid space. In this work, our goal is to extend NeRF to achieve urban-scale semantic understanding.

**Semantic Understanding with Neural Fields** Recent research has explored the use of NeRF for semantic understanding of 3D scenes. Semantic-NeRF [101] fuses 2D semantic labels into 3D using an additional MLP branch to predict semantic logits [54, 83]. A similar idea of fusing 2D to 3D with NeRF has also been applied to fuse multi-view features [7, 31, 37, 40, 75] to enable open-vocabulary understanding. Moreover, leveraging the powerful segment anything (SAM) model [39], SA3D [8] proposes to segment a single object in NeRF with a user click [12].

For 3D panoptic segmentation, one of the main challenges is obtaining appropriate instance supervision across multiple views [5, 49]. Instance-NeRF [49] and Panoptic NeRF [23] use 3D instance information for training. PNF [41] relies on object tracking to provide instance supervision. Panoptic Lifting [71] adopts linear assignment to match the current predicted 3D instance with the provided 2D labels. Contrastive Lifting [4] utilizes feature contrastive learning, followed by clustering to obtain instance information [15]. However, existing methods are mainly designed for indoor [18, 67, 73] or outdoor street-view [26, 46] scenes. In contrast, our focus is on urban scene understanding from aerial images, which is particularly challenging as aerial images encompass a wide range of object sizes and each image captures only a small portion of the entire scene.

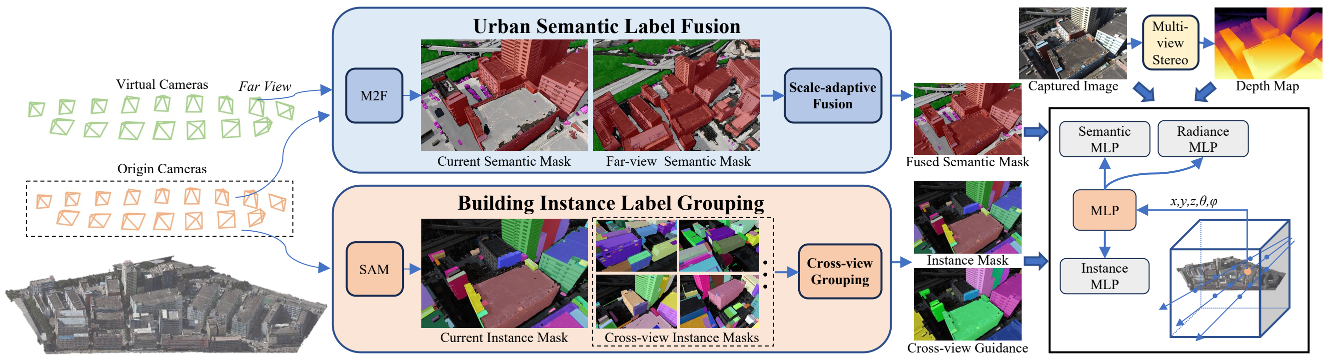



Figure 2. Overview. We present a neural radiance field (NeRF) method for urban-scale semantic and building-level instance segmentation from aerial images by lifting noisy 2D labels to 3D. For semantic segmentation, we adopt a *scale-adaptive semantic label fusion* strategy to fuse the semantic labels from different altitudes using images rendered by NeRF, mitigating the ambiguities of the 2D semantic labels. For instance segmentation, we propose a *cross-view instance label grouping* strategy to guide the training of instance field. In addition, a depth prior from Multi-view Stereo (MVS) is introduced to enhance the geometry reconstruction, leading to more accurate semantic learning.

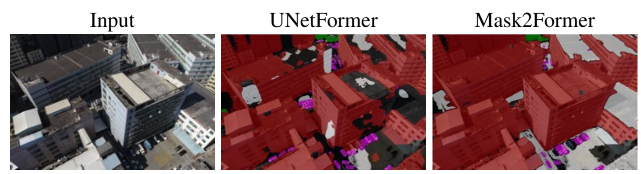
## 3. Method

### 3.1. Overview

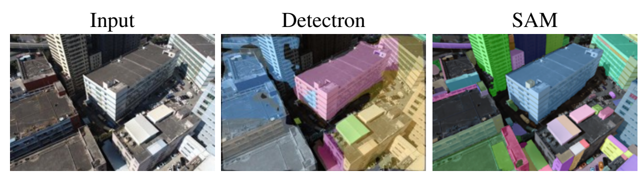
Given multi-view posed aerial images $\{\mathbf{I}\}$ of an urban scene, we perform 3D semantic and building-level instance understanding of the scene based on the neural radiance field (NeRF) [57]. Our method applies off-the-shelf methods [13, 39] to obtain the semantic labels $\{\mathbf{M}\}$ and instance labels $\{\mathbf{H}\}$ for input images, and then lifts the noisy 2D labels to 3D via per-scene optimization (see Fig. 2).

**Challenges** There are two critical challenges that need to be addressed. First, due to significant variations in object size, state-of-the-art semantic segmentation methods, such as Mask2Former [13, 14] trained on daily images and UNetFormer [85] trained on a small scale of aerial images, struggle to generate reliable semantic labels for aerial images (see Fig. 3 (a)). Secondly, obtaining accurate building instance segmentation is challenging due to the diverse shapes and substantial size of buildings. Figure 3 (b) shows that the leading method [29, 43] for building instance segmentation in aerial images fails to generate robust instance labels, especially for dense cluster of buildings. Recently, SAM [39] demonstrates superior generalization ability in semantic-agnostic instance segmentation with precise mask boundaries. However, SAM produces over-segmented masks and multi-view inconsistent instance segmentation (*e.g.*, a building segmented as one instance in one view might become multiple different instances in other views).

Lifting such inaccurate 2D labels to 3D with NeRF results in inaccurate 3D semantic and instance segmentation. To address these issues, we introduce a *scale-adaptive semantic label fusion* strategy for semantic segmentation and a *cross-view instance label grouping* strategy for instance segmentation to provide more accurate and consistent supervision for the 2D-to-3D lifting process.




(a) Semantic segmentation from 2D methods




(b) Instance segmentation from 2D methods

Figure 3. Problem of existing 2D semantic and instance segmentation methods. (a) We use the red color to highlight the buildings and white for roads. UNetFormer suffers from recognizing road and Mask2Former suffers from misclassification between rooftops and roads. (b) Distinctive colors are assigned to different instances. The instance labels obtained from Detectron appear overly large, while those from SAM seem excessively small.

### 3.2. 3D Scene Representation

**Neural Radiance Field** We represent the geometry and appearance of a 3D scene with NeRF [57], which employs a continuous function to map a 3D point $\boldsymbol{x}_k$ in space and view direction $\boldsymbol{d}$ to density $\sigma_k$ and color $\boldsymbol{c}_k$. The pixel color can be computed by integrating the color of the points sampled along its visual ray $\boldsymbol{r}$ through volume rendering:

$$\tilde{C}(\boldsymbol{r}) = \sum_{k=1}^{K} T_k(1 - \exp(-\sigma_k\delta_k))\boldsymbol{c}_k, \tag{1}$$

where $T_k = \exp\left(-\sum_{j=1}^{k-1}\sigma_j\delta_j\right)$, and $\delta_k = t_{k+1} - t_k$ is the distance between adjacent sampled points. During optimization, a NeRF is fitted to a scene by minimizing the reconstruction error between the rendered color $\tilde{C}$ and the captured color $C$ in the sampled ray set $\boldsymbol{R}$:

$$\mathcal{L}_{\text{color}} = \sum_{\boldsymbol{r}\in\boldsymbol{R}} \|\tilde{C}(\boldsymbol{r}) - C(\boldsymbol{r})\|_2^2. \tag{2}$$

**Semantic and Instance Fields** We follow semantic-NeRF [101] and panoptic lifting [71] to add a semantic branch and an instance branch to represent the 3D semantic and instance fields. The semantic and instance labels $\tilde{S}(\boldsymbol{r})$ of a ray can be rendered by volume rendering as Eq. (1):

$$\tilde{S}(\boldsymbol{r}) = \sum_{k=1}^{K} T_k(1 - \exp(-\sigma_k\delta_k))\boldsymbol{s}_k \tag{3}$$

where $\boldsymbol{s}_k$ is the semantic or instance output of a 3D point.

Given the 2D semantic labels for each image, the semantic field can be optimized by minimizing the multi-class cross-entropy loss $\mathcal{L}_{\text{semantic}}$ between the rendered semantic labels and the 2D labels [101]. The loss function for instance field $\mathcal{L}_{\text{instance}}$ needs special design, as the instance IDs of the same 3D instance predicted from different images are not consistent (*e.g.*, a building instance might have an ID of 1 in one view and an ID of 2 in another view). Existing methods propose to solve a linear assignment problem to match the best 3D and 2D instance pairs [71] or utilize contrastive feature learning to cluster 3D instances [4]. *Note that these methods do not consider the problem of multi-view instance label inconsistency, where an instance might be segmented into multiple different instances in different views, which is common in urban aerial images.*

### 3.3. Urban Semantic Label Fusion

We employ the state-of-the-art Mask2Former [13] to estimate 2D segmentation masks $\{\mathbf{M}\}$ for input views, focusing on the four primary categories in the urban landscape: *Buildings*, *Trees*, *Cars*, and *Roads*. Other methods can also be used, but we found Mask2Former is more robust and accurate [4, 71]. However, segmentation labels generated through 2D methods suffer from ambiguities, *e.g*. building rooftops may erroneously be labeled as road surfaces (see Fig. 3 (a)). This misclassification stems from the scale variability inherent in aerial imagery, where each image captures only a limited portion of the large building.

To circumvent this issue, we propose a scale-adaptive semantic label fusion strategy to improve the semantic label. This idea stems from the observation that the semantic labeling of large object categories (*e.g.*, *building*) is more reliable when viewed from a distance view point, as the object will become smaller in the context (see Fig. 4).

**Scale-adaptive Semantic Label Fusion** NeRF has a great ability for photorealistic novel-view synthesis compared to the explicit representations, *e.g.*, point clouds. We perform novel-view synthesis based on the radiance field to simulate images captured from elevated altitudes. For each original image, we increase its camera altitudes for novel-view rendering, rendering a set of far view images $\{\mathbf{I}^{\mathrm{f}}\}$. We then compute the segmentation mask $\{\mathbf{M}^{\mathrm{f}}\}$ for the far view images. By leveraging the depth information derived from the neural radiance field, the segmentation obtained from the *far view* is then back-projected to the original view for refining the mask of the building category. Specifically, considering a pixel with the coordinate of $p^f$ in a far view image, the projected pixel coordinate $p^o$ in the original image is defined as:

$$p^o \sim K\mathbf{T}_{f\to o}\tilde{D}^f(p^f)K^{-1}p^f, \tag{4}$$

where $K$ is the camera intrinsic, $\mathbf{T}_{f\to o}$ is the relative transformation from the far to the original camera, and $\tilde{D}^f$ represents the rendered depth map of the far view image.

Furthermore, we apply SAM on the original captured images to predict semantic agnostic masks, which will be utilized to refine the masks of small-scale categories, *e.g*. *trees* and *cars*. Specifically, for each semantic mask of the small-scale categories, we match it with the SAM mask that has an intersection of union (IoU) larger than $0.5$. The matched SAM masks will be the refined semantic mask.

To verify the effectiveness of our method, we measure the label consistency in 3D based on the UrbanBIS dataset [93]. Given the 3D point cloud, we back-project the predicted 2D semantic label for each view to the 3D space leveraging the camera poses. Each 3D point will receive

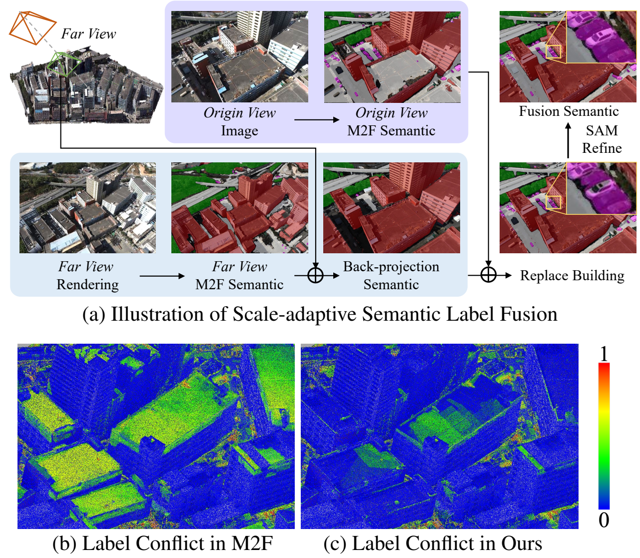



Figure 4. (a) Illustration of the scale-adaptive semantic label fusion process. (b)-(c) Visualization of the conflict level in 3D points using entropy, where higher entropy indicates higher conflict level.

multiple 2D semantic labels from different views, and we compute the entropy to measure the inconsistency across views, reflecting the accuracy of labels. Figure 4 presents the conflict with entropy and shows the visualization results. Compared to Mask2Former, our scale-adaptive integration for semantic labels significantly reduces ambiguity between building rooftops and roadways at the original scale, thereby improving per-view segmentation accuracy and essentially reducing the difficulty of 2D-to-3D lifting.

### 3.4. Building Instance Label Grouping

**Semantic-agnostic Instance Generation** Existing instance segmentation methods [29, 43] struggle with robust instance segmentation for aerial images of diverse urban scenes. Impressed by the superior generalization ability of SAM [39], we utilize SAM to generate semantic-agnostic masks for building instance segmentation. For each image, a grid of $32 \times 32$ points will be utilized as the input prompt for SAM to predict a set of possible instances.

However, despite its generality, the mask generated by SAM has two characteristics that harm the building instance segmentation: 1) The SAM model generates masks at different levels of granularity, which might lead to small masks nested inside larger ones, resulting in redundant masks that belong to the same instance (*e.g.*, window mask on top of the building mask). 2) The generated 2D masks for the same 3D instance are not consistent across multi-view, *e.g.*, a building instance which is accurately segmented in one view might be segmented into multiple different instances.

**Geometry-guided Instance Filtering** The geometry-guided instance filtering is designed to identify and remove smaller masks nested inside larger masks and exhibit limited height variation. Specifically, leveraging the camera parameters and the depth map $\tilde{D}$ of each image computed from the radiance field, we map pixels of each mask to 3D space to determine their physical heights as the difference of the highest and the lowest altitudes. Subsequently, we filter the nested masks with heights smaller than a threshold.

**Cross-view Instance Label Grouping** As an instance might be segmented into different blocks in different views by SAM, directly lifting SAM masks to the 3D instance field is suboptimal as 3D points will receive conflict supervision in different views. To resolve this problem, we introduce a cross-view instance label grouping strategy. The key idea is to synchronize the instance segmentation across different views, thereby consolidating smaller segmented instances into a singular, coherent instance (see Fig. 5).

Consider a scenario with $N$ images. For each image, denoted as the $i$-th view, we have a set of predicted SAM masks, represented as $\mathbf{H}_i$. When examining the instance segmentation from the perspective of the $i$-th view, it is essential to incorporate the segmentation information from other views. To achieve this, we project the SAM masks

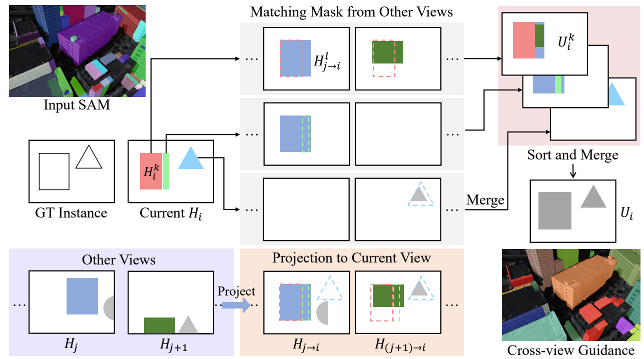



Figure 5. Illustration of cross-view instance label grouping. Given the SAM mask of one view, we can get the cross-view guidance map for the instance field training.

from all other views ($j$) onto the $i$-th view. This set of projected masks is denoted by $\{\mathbf{H}_{j\to i} | j = 1, \dots, N, j \neq i\}$, using camera parameters and depth as specified in Eq. (4).

For each instance mask $\mathbf{H}_i^k$ in the $i$-th view, we seek to identify corresponding masks in $\mathbf{H}_{j\to i}$. A match between a pair of masks, $\mathbf{H}_i^k$ and $\mathbf{H}^l_{j\to i}$, is established if the intersection-over-minimum-area ratio exceeds a predefined threshold $\tau$ as $\frac{|\mathbf{H}_i^k \cap \mathbf{H}^l_{j\to i}|}{\min(|\mathbf{H}_i^k|, |\mathbf{H}^l_{j\to i}|)} > \tau$, where $|\cdot|$ represents the area of a mask, and $\tau$ is set to 0.5. Upon identifying a match, the corresponding masks are merged by uniting their areas, resulting in an expanded mask $\mathbf{H}^k_{i\cup j}$. This process is repeated for all matches, leading to a collection of expanded masks. These expanded masks are then combined to form a comprehensive *cross-view mask* for each instance as $U_i^k = \bigcup_{j\neq i} \mathbf{H}^k_{i\cup j}$.

This procedure is executed for every instance mask in the $i$-th view, resulting into a set of cross-view masks. These masks are then organized in ascending order based on their areas, and the mask value is set to the ID of the instance mask. Subsequently, they are sequentially layered onto a map of dimensions $H \times W$, creating the *cross-view guidance map*, $U_i$. In this map, smaller masks are progressively overwritten by larger ones, which effectively groups the instances more accurately.

With the help of the cross-view guidance map, different instances in the current view are considered the same group if more than 50% of their pixels in the cross-view guidance map have the same value. During training, we randomly select a single instance from each group. This approach substantially reduces the occurrence of conflicts in the dense SAM mask annotations, such as when two pixels from the same building instance might be incorrectly classified as belonging to separate instances.

### 3.5. Depth Priors from Multi-view Stereo

In the case of expansive urban environments and sparse observations, optimizing the radiance field solely with the photometric loss can result in imprecise geometry and floating artifacts. To mitigate this, our method integrates depth cues derived from multi-view stereo techniques to enforce geometric consistency [68, 69]. We reconstruct depth map $D$ for each view and incorporate a depth regularization term in our loss function to refine the NeRF's geometry:

$$\mathcal{L}_{\text{depth}} = \sum_{\boldsymbol{r} \in \boldsymbol{R}} \|\tilde{D}(\boldsymbol{r}) - D(\boldsymbol{r})\|_2^2, \tag{5}$$

where $\tilde{D}(\boldsymbol{r})$ denotes the rendered depth obtained by volume rendering the ray distance in a similar way as in Eq. (1).

Previous studies have leveraged monocular depth [98] or sparse point information [19, 22] for similar purposes. However, we found that the monocular depth estimation methods [66] are not robust to aerial images, especially for views orthogonal to the ground, while the supervision from the sparse depth information is not sufficient for urban scenes.

### 3.6. Optimization

The overall loss function can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{color}} + \lambda_d \mathcal{L}_{\text{depth}} + \lambda_s \mathcal{L}_{\text{semantic}} + \lambda_i \mathcal{L}_{\text{instance}}, \tag{6}$$

where $\lambda_d$, $\lambda_s$, and $\lambda_i$ are the loss weights and set to 1 in the experiments.

During optimization, we first optimize the radiance field to recover the scene geometry and appearance, and then optimize the semantic and instance fields. The loss function for semantic is the multi-class cross-entropy loss [101]. For instance field optimization, we integrate our cross-view grouping strategy with loss functions introduced by contrastive lifting [4] and panoptic lifting [71]. During optimization, we filter image rays that do not belong to the building category. Experiments show that our method effectively improves contrastive lifting and panoptic lifting for building instance segmentation in urban scenes.

## 4. Experiments

**Datasets** We evaluate our method on the real-world urban scene dataset, named UrbanBIS [93] dataset. UrbanBIS dataset provides 3D semantic segmentation annotations, including buildings, roads, cars, and trees, as well as 3D building-level instance annotations (see Table 1). We select four regions with a high density of building instances and various architecture styles, namely *Yingrenshi*, *Yuehai-Campus*, *Longhua-1*, and *Longhua-2*. We downsample images by four times for training and uniformly sample around ten images as the testing set for each scene.

**Evaluation Metrics** We measure the quality of the novel-view synthesis and semantic segmentation in terms of PSNR and the mean intersection over union (mIoU), respectively. To evaluate the instance building segmentation, we use a scene-level Panoptic Quality (PQ$^{\text{scene}}$) metric [71], which takes the consistency of the instance across different views into account. As we focus on the segmentation of building instances, we report the PQ$^{\text{scene}}$ of the building.

Table 1. Statistics of the UrbanBIS dataset [93].

| Dataset | Covered area | Number of images | Resolutions | Building instances |
|---|---|---|---|---|
| Yingrenshi | $440 \times 220\ m^2$ | 854 | $3648 \times 5472$ | 41 |
| Yuehai-Campus | $900 \times 280\ m^2$ | 955 | $3648 \times 5472$ | 30 |
| Longhua-1 | $550 \times 530\ m^2$ | 999 | $5460 \times 8192$ | 26 |
| Longhua-2 | $550 \times 300\ m^2$ | 677 | $5460 \times 8192$ | 38 |

Table 2. Comparison with 2D segmentation methods.

| Method | Yingrenshi | | Yuehai-Campus | | Longhua-1 | | Longhua-2 | |
|---|---|---|---|---|---|---|---|---|
| | building | road | building | road | building | road | building | road |
| UNetFormer [85] | 76.0 | 17.5 | 68.7 | 24.5 | 77.5 | 4.5 | **78.4** | 10.8 |
| Mask2Former [13] | 84.8 | 49.7 | 70.9 | 44.7 | 68.5 | 47.5 | 70.0 | 42.5 |
| Scale-adaptive fusion | **93.2** | **56.8** | **90.7** | **52.0** | **77.9** | **48.7** | 77.4 | **43.0** |

### 4.1. Evaluation on Semantic Segmentation

**Choice of 2D Segmentation Method** We discuss two types of 2D segmentation methods for providing the semantic labels for each aerial image, namely the UNetFormer [85], which is designed for the aerial images semantic segmentation, and Mask2Former [13], which is a universal panoptic segmentation method. As shown in Table 2, UNetFormer does not generalize well on the UrbanBIS dataset, which fails to segment the *Road*. Therefore, we take Mask2Former as the foundation to get the initial 2D semantic segmentation.

**2D Semantic Label Fusion** We first demonstrate the effectiveness of the proposed scale-adaptive semantic label fusion by evaluating the accuracy of the 2D semantic labels. We can see from Table 2 that the proposed scale-adaptive fusion can significantly improve the accuracy, especially for the building category.

**3D Semantic Field** To evaluate the performance of semantic segmentation in 3D lifting, we conducted a comparative analysis of our method against the official implementation of Panoptic-Lift [71], and a modified Semantic-NeRF [101]. Panoptic-Lift employs the universal Mask2Former for predicting 2D semantic labels and lifts 2D labels to 3D for room-scale scene. However, Panoptic-Lift performs worse on the urban scene due to the worse geometry reconstruction as it employs the TensoRF [9] as the backbone, which struggles to scale up to a high grid resolution. For a fair comparison, we designed a variant of semantic-NeRF using the same geometry backbone as ours (*i.e*., high-resolution hash-grid). This modified semantic-NeRF is trained with semantic labels from Mask2Former, while our method is trained with fused labels.

Quantitative results in Table 3 demonstrate that our method outperforms the others in terms of mIoU through the use of scale-adaptive fusion, highlighting its effectiveness. Moreover, Figure 6 shows that our method solves

Table 3. Quanlitative comparison on the novel-view synthesis and the semantic segmentation on the UrbanBIS dataset [93]. Mask2Former is a 2D segmentation method and cannot be evaluated for PSNR, and Semantic-NeRF (M2F) shares the same geometry with ours.

| Method | Yingrenshi [93] | | | | Yuehai-Campus [93] | | | | Longhua-1 [93] | | | | Longhua-2 [93] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mIoU↑ | Building↑ | Car↑ | PSNR↑ | mIoU↑ | Building↑ | Car↑ | PSNR↑ | mIoU↑ | Building↑ | Car↑ | PSNR↑ | mIoU↑ | Building↑ | Car↑ | PSNR↑ |
| Mask2former [13] | 58.4 | 84.8 | 28.4 | – | 57.9 | 70.9 | 42.3 | – | 49.2 | 68.5 | 20.7 | – | 46.2 | 70.0 | 23.5 | – |
| Panoptic-Lift [71] | 32.9 | 83.0 | 1.4 | 21.2 | 33.7 | 69.0 | 0.1 | 21.5 | 29.5 | 60.2 | 0.2 | 21.9 | 20.4 | 59.9 | 0.2 | 20.9 |
| Semantic-NeRF (M2F) [101] | 67.6 | 92.9 | 39.5 | – | 70.9 | 88.1 | 45.3 | – | 61.1 | 86.4 | 34.9 | – | 64.1 | 89.9 | 37.5 | – |
| Ours | **72.0** | **95.5** | **49.3** | **26.8** | **74.9** | **94.1** | **46.2** | **29.8** | **66.1** | **87.3** | **43.8** | **26.7** | **66.7** | **91.0** | **40.9** | **26.4** |

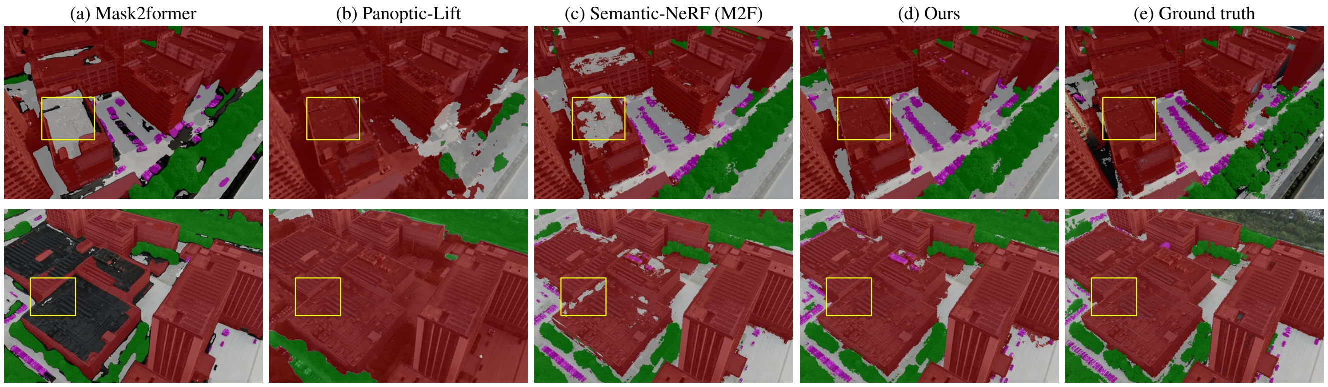



Figure 6. Qualitative comparison of semantic segmentation on *Yingrenshi* and *Longhua-2* from the UrbanBIS dataset (*Building*: Red, *Road*: White, *Car*: Violet, *Tree*: Green, unrecognized areas of Mask2Former: Black). Areas without masks in (e) have no GT annotation.

Table 4. Quantitative comparison of instance segmentation in $PQ^{scene}$ of building category. For brevity, LA and CL denote the linear assignment and contrastive learning, respectively.

| Method | Yingrenshi | Yuehai-Campus | Longhua-1 | Longhua-2 |
|---|---|---|---|---|
| LA + Detectron-Label [43] | 15.8 | 40.8 | 38.2 | 18.2 |
| LA + SAM-Label [39] | 14.4 | 4.0 | 4.5 | 4.0 |
| LA + Ours | 38.7 | 26.0 | 36.3 | 19.0 |
| CL + Detectron-Label [39] | 26.6 | 30.6 | 36.5 | 17.3 |
| CL + SAM-Label [39] | 54.8 | 18.8 | 29.7 | 22.9 |
| CL + Ours | **64.1** | **43.6** | **45.8** | **31.5** |

the issue of misclassification resulting from the ambiguity between building roofs and road surfaces in aerial images, leading to better results.

## 4.2. Evaluation on Instance Building Segmentation

To lift 2D instance labels to 3D, we utilize two different optimization methods: the linear assignment from Panoptic-Lift [71] and the contrastive learning from Contrastive-Lift [4] which is followed by HDBSCAN [55] as post-processing cluster algorithm. Moreover, we did not make a comparison with the Panoptic-Lift official implementation because of its poor semantics results. To mitigate the impact of geometry reconstruction and semantic segmentation, we employ the same geometry and semantic results across experiments in this section.

We establish two baselines, one trained with instance labels obtained from the Detectron [43] and one with labels from SAM [39]. For brevity, we refer to these baselines as Detectron-Label and SAM-Label. Our method builds upon SAM-Label by incorporating the cross-view label grouping.

Table 4 presents the $PQ^{scene}$ metric on four urban scenes. Results on *Yingrenshi* reveal that training with Detectron-Label struggles with dense building instances due to inaccurate instance segmentation. While the SAM-Label model achieves reasonable results in *Yingrenshi*, it struggles to handle large buildings with diverse shapes as shown in the other three scenes. It is because SAM tends to produce over-segmented labels for these buildings, leading to cluttered 3D segmentation, particularly trained with linear assignment. With our proposed cross-view label grouping strategy, we significantly improve the performance compared to that trained with SAM-Label in both linear assignment and contrastive learning. The best results are achieved by integrating the cross-view grouping with the contrastive learning. Figure 7 illustrates the qualitative comparison, where our approach exhibits more accurate segmentation results, further affirming the effectiveness of our method.

## 4.3. Ablation Analysis

To further verify the design of our method, we conduct ablation studies on the *Yingrenshi* dataset.

**Effect of Instance Label Grouping** We evaluate the effectiveness of the components employed in the proposed instance segmentation method. As depicted in Table 5, the utilization of geometry-guided instance filtering can improve instance segmentation in some extent, compared to the baseline trained with SAM-Label. More importantly, utilizing the cross-view grouping strategy achieves significant improvement in instance segmentation, demonstrating

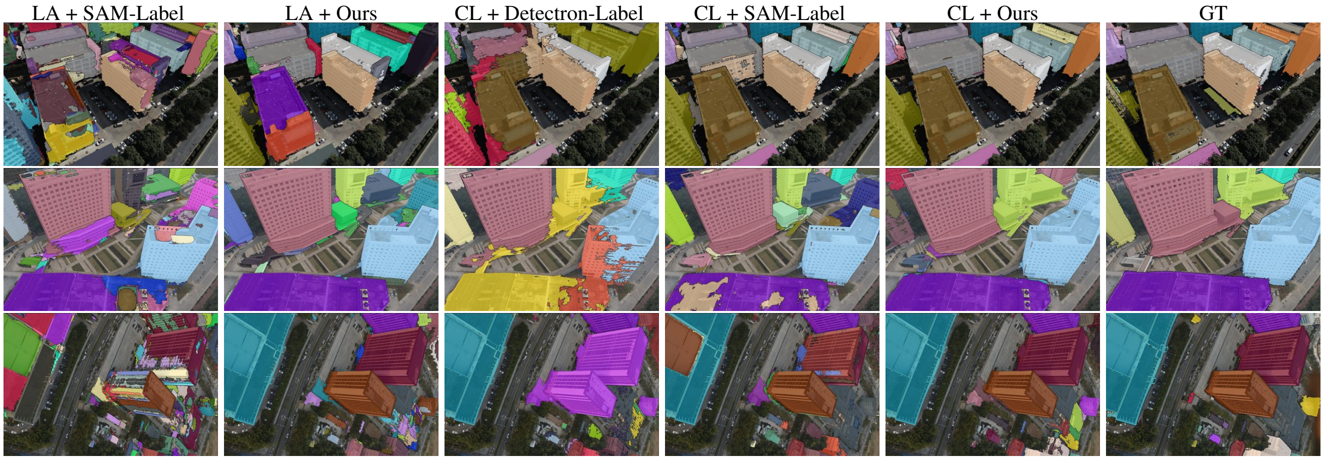



Figure 7. Qualitative comparison on the building instance segmentation. From top to bottom, we show the results of different approaches in three scenarios: *Yingrenshi*, *Yuehai-Campus*, and *Longhua-2*. Different instances are represented in different colors.

Table 5. Effect of cross-view label grouping ($PQ^{scene}$).

| Method | Linear assignment | Contrastive learning |
|---|---|---|
| Baseline (SAM-Label) | 14.4 | 54.8 |
| Baseline + Filter | 19.8 | 55.9 |
| Baseline + Filter + Cross-view | **38.7** | **64.1** |

Table 6. Effect of geometry reconstruction quality.

| Method | PSNR | mIoU | Building | Road | Car | Tree |
|---|---|---|---|---|---|---|
| Panoptic-Lift [71] | 21.24 | 32.9 | 83.0 | 32.6 | 1.4 | 14.5 |
| Ours without depth-prior | 25.01 | 70.4 | 94.9 | 66.6 | 46.7 | 73.3 |
| Ours with depth-prior | **26.79** | **72.0** | **95.5** | **68.9** | **49.3** | **74.5** |

the effectiveness of the cross-view grouping strategy.

**Effect of Geometry Reconstruction** To investigate the effect of the geometry reconstruction, we compare the novel-view synthesis and semantic segmentation results in Table 6. We can see from the table that Panoptic-Lift suffers from low-quality reconstruction, resulting in poor segmentation of small objects (*e.g.* car and tree categories). By incorporating the depth-prior from multi-view stereo, the rendering quality and segmentation quality can be effectively improved. Figure 8 shows an example of the novel-view synthesis.

## 5. Conclusion

In this paper, we have introduced a neural radiance field method for urban-scale semantic segmentation and building-level instance segmentation from aerial images. Our method lifts noisy 2D labels, predicted by off-the-shelf methods, to 3D without manual annotations. We proposed a scale-adaptive semantic label fusion strategy that significantly improves the segmentation results across objects of varying sizes. To achieve multi-view consistent instance supervision for building instance segmentation, we introduced a cross-view instance label grouping strategy based on the 3D scene representation. In addition, we enhanced the reconstructed geometry by incorporating the depth prior from multi-view stereo, leading to more accurate segmentation results. Experiments on multiple real-world scenes demonstrate the effectiveness of our method.


Figure 8. Visualization of novel-view synthesis.

**Future Work** Currently, our method focuses on close-vocabulary scene understanding. Recent methods have shown promising results in open-vocabulary understanding by distilling CLIP features into NeRF [37, 40]. Nonetheless, feature conflicts caused by varying object sizes and multi-view inconsistency can impair the distillation. In the future, we aim to apply the proposed method to enhance the feature distillation process.

**Acknowledgement.** This work was supported in part by NSFC with Grant No. 62293482, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone. It was also partially supported by NSFC with Grant No. 62202409, Shenzhen Science and Technology Program with Grant No. RCBS20221008093241052, the National Key R&D Program of China with grant No.2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No.2017ZT07X152 and No.2019CX01X104, and by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No.2022B1212010001).

## References


[1] Jibril Muhammad Adam, Weiquan Liu, Yu Zang, Muhammad Kamran Afzal, Saifullahi Aminu Bello, Abdullahi Uwaisu Muhammad, Cheng Wang, and Jonathan Li. Deep learning-based semantic segmentation of urban-scale 3D meshes in remote sensing: A survey. *International Journal of Applied Earth Observation and Geoinformation*, 2023. 2

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building Rome in a day. *Communications of the ACM*, 2011. 2

[3] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 1975. 16

[4] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. *arXiv:2306.04633*, 2023. 2, 4, 6, 7, 15

[5] Wang Bing, Lu Chen, and Bo Yang. DM-NeRF: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*, 2022. 2

[6] Maros Blaha, Christoph Vogel, Audrey Richard, Jan D Wegner, Thomas Pock, and Konrad Schindler. Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In *CVPR*, 2016. 2

[7] Kenneth Blomqvist, Francesco Milano, Jen Jen Chung, Lionel Ott, and Roland Siegwart. Neural implicit vision-language feature fields. *arXiv preprint arXiv:2303.10962*, 2023. 2

[8] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. *NeurIPS*, 36, 2024. 2

[9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, pages 333–350. Springer, 2022. 2, 6

[10] Jiazhou Chen, Yanghui Xu, Shufang Lu, Ronghua Liang, and Liangliang Nan. 3-D instance segmentation of MVS buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2

[11] Meida Chen, Qingyong Hu, Zifan Yu, Hugues Thomas, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. STPLS3D: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv preprint arXiv:2203.09065*, 2022. 2

[12] Xiaokang Chen, Jiaxiang Tang, Diwen Wan, Jingbo Wang, and Gang Zeng. Interactive segment anything NeRF with feature imitation. *arXiv preprint arXiv:2305.16233*, 2023. 2

[13] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1, 3, 4, 6, 7, 13

[14] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 3

[15] Xinhua Cheng, Yanmin Wu, Mengxi Jia, Qian Wang, and Jian Zhang. Panoptic compositional feature field for editable scene rendering with network-inferred labels via metric learning. In *CVPR*, 2023. 2

[16] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 2

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 13

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2

[19] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022. 6

[20] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, pages 5501–5510, 2022. 2

[21] Christian Früh and Avideh Zakhor. An automated method for large-scale, ground-based city model acquisition. *IJCV*, pages 5–24, 2004. 2

[22] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-NeuS: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *NeurIPS*, 2022. 6

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

[24] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010. 2

[25] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. *TPAMI*, 2010. 2

[26] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[27] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 2

[28] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. StreetSurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 2, 13

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 3, 5

[30] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. In *CVPR*, 2021. 2

[31] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *CVPR*, 2023. 2

[32] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *CVPR*, 2021. 1

[33] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, 130(2):316–343, 2022. 2, 17

[34] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *CVPR*, 2020. 1, 2, 17

[35] Juana Valeria Hurtado, Rohit Mohan, Wolfram Burgard, and Abhinav Valada. MOPT: Multi-object panoptic tracking. *arXiv preprint arXiv:2004.08189*, 2020. 2

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 2023. 2

[37] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *ICCV*, pages 19729–19739, 2023. 2, 8

[38] Diederik Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *ICLR*, 2015. 13

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 3, 5, 7, 14, 15

[40] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing NeRF for editing via feature field distillation. *NeurIPS*, 2022. 2, 8

[41] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *CVPR*, 2022. 2

[42] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3D point cloud segmentation. In *CVPR*, 2022. 2

[43] Russell Land. detectron2-spacenet. `https://github.com/rcland12/detectron2-spacenet`, 2023. 3, 5, 7, 15

[44] Xiaowei Li, Changchang Wu, Christopher Zach, Svetlana Lazebnik, and Jan-Michael Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*, pages 427–440, 2008. 2

[45] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. MatrixCity: A large-scale city dataset for city-scale neural rendering and beyond. In *ICCV*, 2023. 2

[46] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *TPAMI*, 2022. 2

[47] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the UrbanScene3D dataset. In *ECCV*, 2022. 2, 16

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 13

[49] Yichen Liu, Benran Hu, Junkai Huang, Yu-Wing Tai, and Chi-Keung Tang. Instance neural radiance field. In *ICCV*, pages 787–796, 2023. 2

[50] Chongshan Lu, Fukun Yin, Xin Chen, Wen Liu, Tao Chen, Gang Yu, and Jiayuan Fan. A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. In *ICCV*, 2023. 2

[51] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020. 1

[52] Yongqiang Mao, Kaiqiang Chen, Wenhui Diao, Xian Sun, Xiaonan Lu, Kun Fu, and Martin Weinmann. Beyond single receptive field: A receptive field fusion-and-stratification network for airborne laser scanning point cloud classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:45–61, 2022. 2

[53] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. In *CVPR*, pages 1311–1321, 2022. 2

[54] Ruben Mascaro, Lucas Teixeira, and Margarita Chli. Diffuser: Multi-view 2D-to-3D label diffusion for semantic scene segmentation. In *ICRA*, 2021. 2

[55] Leland McInnes, John Healy, and Steve Astels. HDBSCAN: Hierarchical density based clustering. *J. Open Source Softw.*, 2017. 7, 15

[56] Zhenxing Mi and Dan Xu. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In *ICLR*, 2022. 2

[57] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 1, 2, 3

[58] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *TOG*, 41(4):1–15, 2022. 2

[59] William Nguatem and Helmut Mayer. Modeling urban scenes from pointclouds. In *ICCV*, 2017. 2

[60] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, pages 3504–3515, 2020. 2

[61] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, pages 5589–5599, 2021. 2

[62] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 2

[63] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration, 2017. 13

[64] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 1, 2

[65] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 2

[66] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 44(3):1623–1637, 2020. 6

[67] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *CVPR*, 2021. 2

[68] Schönberger, Johannes L, and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 2, 6

[69] Schönberger, Johannes L, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pages 501–518, 2016. 6

[70] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing 2000*, 2000. 2

[71] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *CVPR*, 2023. 1, 2, 4, 6, 7, 8, 13, 15

[72] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. In *SIGGRAPH*, 2006. 2

[73] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[74] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, pages 5459–5469, 2022. 2

[75] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 2

[76] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, pages 8248–8258, 2022. 2

[77] Jiaxiang Tang. Torch-ngp: a pytorch implementation of instant-ngp. https://github.com/ashawkey/torch-ngp, 2022. 2

[78] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Compressible-composable nerf via rank-residual decomposition. *arXiv preprint arXiv:2205.14870*, 2022.

[79] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, 2020.

[80] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. *arXiv preprint arXiv:2111.05849*, 2021. 2

[81] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *CVPR*, pages 12922–12931, 2022. 2

[82] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 15

[83] 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:2111.13260*, 2021. 2

[84] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *CVPR*, 2022. 2

[85] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022. 1, 3, 6

[86] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, volume 34, 2021. 2

[87] Xiuchao Wu, Jiamin Xu, Zihan Zhu, Hujun Bao, Qixing Huang, James Tompkin, and Weiwei Xu. Scalable neural indoor scene rendering. *TOG*, 2022. 2

[88] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 15

[89] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *ECCV*, 2022. 2

[90] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *CGF*, 41(2):641–676, 2022. 2

[91] Linning Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes. In *CVPR*, pages 8296–8306, 2023. 2

[92] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *NeurIPS*, 2019. 2

[93] Guoqing Yang, Fuyou Xue, Qi Zhang, Ke Xie, Chi-Wing Fu, and Hui Huang. UrbanBIS: a large-scale benchmark for fine-grained urban building instance segmentation. In *SIGGRAPH*, 2023. 2, 4, 6, 7, 15, 16

[94] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman.

Volume rendering of neural implicit surfaces. *NeurIPS*, 34:4805–4815, 2021. 2

[95] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, volume 33, pages 2492–2502, 2020. 2

[96] yatengLG, Alias-z, and horffmanwang. ISAT with segment anything: Image segmentation annotation tool with segment anything. `https://github.com/yatengLG/ISAT_with_segment_anything`, 2023. 16

[97] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. GSPN: Generative shape proposal network for 3d instance segmentation in point cloud. In *CVPR*, 2019. 2

[98] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 2022. 2, 6

[99] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *CVPR*, pages 5449–5458, 2022. 2

[100] Yuqi Zhang, Guanying Chen, and Shuguang Cui. Efficient large-scale scene representation with a hybrid of high-resolution grid and plane features. *arXiv preprint arXiv:2303.03003*, 2023. 2, 13

[101] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 1, 2, 4, 6, 7

[102] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 13

[103] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global sfm by distributed motion averaging. In *CVPR*, pages 4568–4577, 2018. 2