PanoRecon: Real-Time Panoptic 3D Reconstruction from Monocular Video

Dong Wu[1] Zike Yan[2] Hongbin Zha[1]
[1]National Key Lab of GAI, School of IST
PKU-SenseTime Joint Lab of MV
Peking University
[2]AIR, Tsinghua University
riserwu@stu.pku.edu.cn, yanzike@air.tsinghua.edu.cn, zha@cis.pku.edu.cn

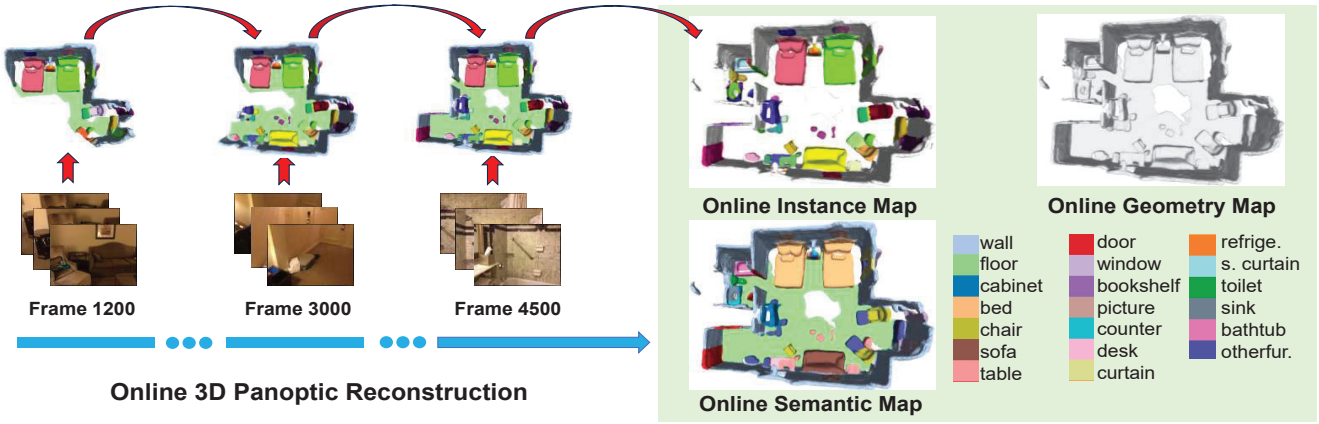

Figure 1. We present PanoRecon, which realizes an online reconstruction at the level of *stuff* and *things* with only monocular video as input. The system runs in real-time, and performs online 3D geometry reconstruction as well as dense semantic labeling for all map region and segmentation of individual *things*.

Abstract

We introduce the Panoptic 3D Reconstruction task, a unified and holistic scene understanding task for a monocular video. And we present PanoRecon - a novel framework to address this new task, which realizes an online geometry reconstruction alone with dense semantic and instance labeling. Specifically, PanoRecon incrementally performs panoptic 3D reconstruction for each video fragment consisting of multiple consecutive key frames, from a volumetric feature representation using feed-forward neural networks. We adopt a depth-guided back-projection strategy to sparse and purify the volumetric feature representation. We further introduce a voxel clustering module to get object instances in each local fragment, and then design a tracking and fusion algorithm for the integration of instances from different fragments to ensure temporal coherence. Such design enables our PanoRecon to yield a coherent and accurate panoptic 3D reconstruction. Experiments on ScanNetV2 demonstrate a very competitive geometry reconstruction result compared with state-of-the-art reconstruction methods, as well as promising 3D panoptic segmentation result with only RGB input, while being real-time. Code is available at: https://github.com/Riser6/PanoRecon.

1. Introduction

3D scene understanding from a posed monocular video is a fundamental task of 3D computer vision and robotics. The understanding ability to infer the underlying geometric structures and recognize object with semantics immediately, is critical towards many downstream applications, such as Augmented and Virtual Reality(AR and VR), interior modeling and human-robot interaction.

Recently, significant advances have been made in 3D geometric reconstruction. Many depth-fusion methods [1–5] and feature-fusion methods [6–12] are proposed, each with its own strengths(see Sec. 2.3). Although the watertight reconstruction of whole scene can be produced, different objects in the scene are unable to be decoupled due to lack of the ability to distinguish instance. On the other

hand, many algorithms are proposed to perform semantic segmentation [13–16] or instance segmentation [17–22] in 3D scene but require dedicated depth sensors, which are more expensive, less compact than cameras. Thus we aim to view the three separate tasks as a new unified task, called *panoptic 3D reconstruction*. For panoptic 3D reconstruction, we aim to recover the surface geometry of the scene from a monocular video and also assign semantic and instance label for geometry elements. Following the task format of 2D panoptic segmentation [23], for "stuff' elements that refer the structural regions of similar texture or material such as wall, floor, the instance ID is often ignored, but for "things" elements that denote countable objects such as chair, table, both semantic label and instance ID need to be distinguished(see Fig. 1).

To address this panoptic 3D reconstruction task, we propose a novel framework that jointly infers geometric structure, semantic label and instance id from a monocular video, called PanoRecon. As illustrated in Fig 2, PanoRecon incrementally performs 3D geometric reconstruction and 3D panoptic segmentation(consists of 3D semantic segmentation and 3D instance segmentation) in a view-independent 3D feature volume. Given a posed monocular video, we successively split it into multiple non-overlapping fragments. Then in order to form a sparse and valid 3D feature volume of each fragment, we adopt a depth-guided back-projection strategy to reduce erroneous feature allocation(detailed in Sec. 3.1). After getting the fused 3D feature volume, we introduce a Voxel-wise Prediction Network to decode semantic and geometric primitive for each voxel in the feature volume(detailed in Sec. 3.2). With the hybrid primitives, we introduce a voxel clustering module to get instance detection of each local fragment, and then a Tracking and Fusion module is designed for the integration of instances fragment by fragment to ensure global consistency(detailed in Sec. 3.3), yielding the final panoptic reconstruction.

The contributions of our work can be summarized as follows:

- We introduce the task of Panoptic 3D Reconstruction from posed monocular video, which aims for holistic 3D scene understanding by jointly reasoning scene geometry and instance-level semantics.
- We propose a novel system, PanoRecon, which realizes coherent while high-detailed geometry reconstruction as well as reasonable panoptic segmentation of the scene, and run in real-time.
- The experimental results on ScanNetV2 [24] show that PanoRecon achieves competitive geometry reconstruction quality compared with state-of-the-art methods and quite considerable 3D panoptic segmentation results in absence of depth input.

2. Related Work

2.1. Semantic SLAM and Neural Field

Recent developments in deep learning have also enabled the integration of rich semantic information within Simultaneous Localization and Mapping (SLAM). [25–29] Combine the semantic segmentation network with SLAM system can incrementally compute semantic 3D map of the environment, but no instance information is included. As the pioneer of object-level SLAM, SLAM++ [30] represents the scene with known CAD object models. NodeSLAM [31] and DSP-SLAM[32] take this direction of using a category-specific learnt shape prior to build the scene instead of CAD model. [33–38] drop the requirement for prior shape knowledge and instead take advantage of 2D instance segmentation masks to obtain object-level scene map. Our work will further get rid of the dependence of 2D semantic mask prior, and directly make prediction for 3D map at one stage. In addition, most of these systems rely on the RGBD sensor data input [26, 27, 30–37], some of them focus more on pose estimation but do not recover the continuous scene surface geometry [25, 27, 38]. Our method takes only RGB image sequence as input while recovering surface geometry and instance-level semantics of the environment.

Neural fields have recently been used as a flexible representation of the whole scene [39–41]. SemanticNeRF [42] and iLabel [43] reveal the coherence properties of neural fields via adding semantic output channel, but they do not explicitly model each semantic entity's geometry. PNF [44] and Panoptic NeRF [45] take a set of images, 3D bounding primitives and 2D predictions as input, and can render RGB color and panoptic segmentation map at arbitrary view. To represent multiple objects, [46–48] take pre-computed instance masks as additional input and conditioned object representation with learnable activation code, but all these methods are still trained offline. vMAP [49] represents each object by a small MLP and takes depth as additional input, which can detect and optimise object instances on-the-fly. However, these neural field methods rely on time-consuming per-scene optimization, while our method uses a feed-forward neural network to directly predict scene geometry and instance-level semantics in real-time.

2.2. Panoptic Segmentation

Kirillov et al. [23] proposes the task of 2D panoptic segmentation, establishing a unified, holistic scene understanding task for a single RGB image. This unified task aim to assign each pixel in 2D image with an semantic label and an instance ID. The semantic labels can be divided into two parts: 'stuff' and 'thing', where 'stuff' labels do not distinguish instance ID. VPSNet [50] proposes and explores a new video extension of this unified task, called video panop-

tic segmentation(VPS). [51–53] present depth-aware VPS network by introducing additional depth information. In contrast to these existing methods, we try to directly inference the instance-level semantics in the 3D space, and are more concerned with the global results rather than the results of individual frames.

Another line of works take RGBD sequences or point cloud as input and output the semantic label of 3D input data. Broadly speaking, there are two kinds of approaches to solve this problem: (1) Predict semantics of 2D image using 2D segmentation network [54–56] and back-project the semantic label to 3D data [27, 36, 57]. (2) Directly inference the semantic labels in the 3D space. [13–16] take voxelized point clouds as input and then apply 3D convolution on the voxel grid, which output the class semantics of each voxel regardless of the instance ID. In order to produce instance-level semantics of per-voxel, proposal-based methods [17, 18] and grouping-based methods [19–22] are presented to detect object instances. Different from these works, we propose a baseline to a relatively untouched problem of 3D panoptic segmentation without depth sensor.

2.3. 3D Reconstruction

There has been a long history of researches on multi-view stereo reconstruction [1, 58], which pose 3D reconstruction as per-pixel depth estimation. Recent works [2–5] extend the classical MVS with neural network to construct 3D cost volume with multi-view features, which is used to regress the dense depth maps. Although these methods have shown strong results, performing especially well on recovering highly detailed geometry, a known drawback is that the estimation of each depth map is independent, so continuity across different frames is not constrained, and this often leads to artifacts.

An alternative method has been proposed to address 3D reconstruction in Altas [6], which proposed direct prediction of TSDF and label volume from back-projected image features with a 3D CNN. [7, 9–11] adopt a sparse 3D CNN to perform feature volume-based reconstruction fragment by fragment for better efficiency and scalability, and utilize a GRU to fuse fragment feature volumes over time for better coherence. The main advantage of these method is that the 3D CNN can learn to produce smooth and consistent surfaces. However, the geometric reconstruction with these methods still remains coarse. In order to recover more accurate surfaces, [9, 12] propose to use multi-view depth estimates as guidance to enhance the scene representation and produce high-resolution predictions. Moreover, although these methods can produce watertight reconstruction of the whole scene, the scene entities can not be split to different objects due to lack of instance-level information.

We share a similar pipeline with PlanarRecon [59], where we recover panoptic watertight mesh instead of low-polygonal geometry of planar instances. [60, 61] address the panoptic 3D scene reconstruction task from a single image rather than monocular video. Our work will combine the advantages of multi-view stereo method and volumetric-based method, achieving highly detailed and coherent reconstruction, and for the first time show an instance-level reconstructed map from a posed monocular video.

3. Methods

Given a sequence of monocular video frames $\{I_t\}$ and their camera poses $\{\xi_t\} \in \mathbb{SE}(3)$ provided by a SLAM system, our goal is to incrementally reconstruct the underlying geometric structures and recognize objects as well as semantics.

To achieve online reconstruction, we sequentially select suitable key frames from the incoming image stream following [7]. A new incoming frame is selected as key frame if the camera motion is greater than a predefined threshold [7]. We then split the key frame stream into multiple non-overlapping fragments $\mathcal{F}_i = \{I_{i,j}\}_{j=1}^{N}$, each of which consists of N consecutive key frames. With these fragments, we then perform online panoptic 3D reconstruction.

An overview of our proposed method is shown in Fig 2. We divide our framework into three major components, namely Depth-guided Feature Allocation and Fusion, Voxel-wise Prediction Network, and Instance Detection, Tracking and Fusion. Below, we will introduce these components in detail.

3.1. Depth-guided Feature Allocation and Fusion

Overview. The task of panoptic 3D reconstruction requires coherent geometry recovered from the whole scene, also detailed reconstruction of foreground objects. So we choose to utilize the high-detailed depth prior from an MVS network to guide the volxel feature allocation and fusion. As MVS is a well-studied problem, we leverage an off-the-shelf method [5] due to its outstanding efficiency and accuracy.

Feature Allocation. Unlike most existing volume-based methods that allocate dense feature volume along each ray, our method allocates sparse feature volumes from locations only where the surface is likely located according to the depth prior, similar to [9, 10]. Given a predicted depth map and 2D feature which extracted by a 2D CNN from key frame image, our method back-projects the 2D feature only to the voxels within a predefined distance $\triangle d$ from the corresponding estimated depth surface along the ray. This simple and direct approach can not only sparse the feature volume for efficiency, but also avoid the problem of irrelevant 2D feature filling due to the occlusion of foreground objects. An illustration of this module is included in supplementary material.

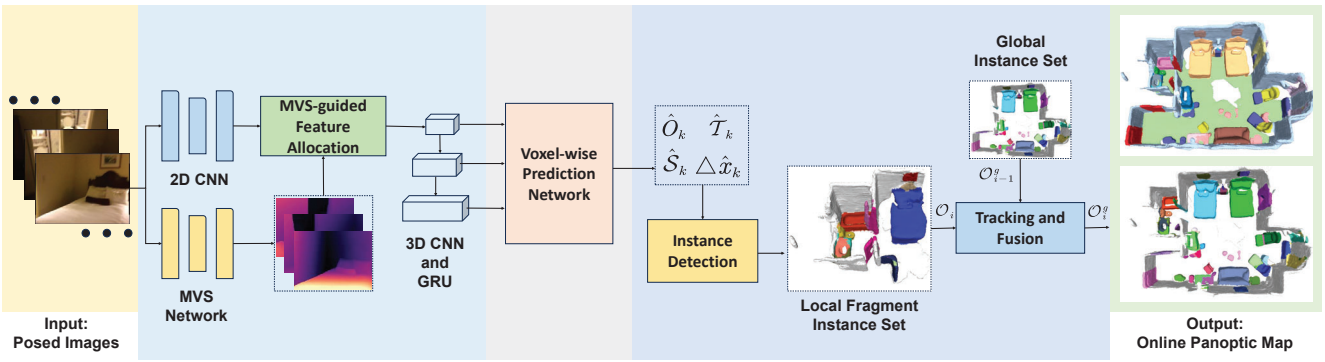

Figure 2. **PanoRecon Framework.** PanoRecon back-projects the 2D image features into a fragment bound volume $\mathcal{F}_i$ with the guidance of MVS depth estimation, and gradually processes the feature volume in a coarse-to-fine paradigm with the 3D CNN and GRU, then forms a temporal coherent feature volume. A voxel-wise prediction network will be used to decode geometric and semantic primitives $\{\hat{O}_{i,k}, \hat{T}_{i,k}, \hat{S}_{i,k}, \triangle\hat{x}_{i,k}\}$ for each voxel in this fragment. Then the hybrid primitives $\{\hat{S}_{i,k}, \triangle\hat{x}_{i,k} + x_{i,k}\}$ are sent to the instance detection module to obtain the instance objects $\mathcal{O}_i = \{O_{i,m}\}$ in this fragment. The tracking and fusion module matches the current instances $\mathcal{O}_i$ with the global instances $\mathcal{O}^g_{i-1}$ from all previous fragments, and then performs fusion for matched instances as well as initialization for new instances, yielding the current global panoptic reconstruction.

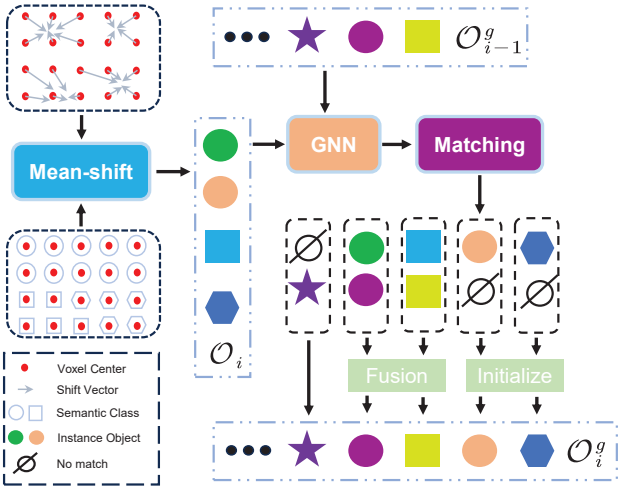

Figure 3. **2D illustration of Instance Detection, Tracking and Fusion.** We use different shapes to represent different semantic classes, and use different colors to distinguish instances. The red dot means the coordinate center of each voxel.

Feature Fusion. After back-projecting view-dependent 2D features into 3D volume, the view-independent feature volume is obtained by directly averaging the features from different views. Thanks to the guidance of depth prior, we do not encouter the noise problem caused by the fusion of irrelevant feature, mentioned in [8]. Following [7], we use 3D sparse convolution to efficiently process the feature volume and adopt the Gated Recurrent Unit(GRU) to perform feature fusion between current-fragment reconstruction and previously reconstructed global volume for temporal coherence, in a coarse-to-fine paradigm.

3.2. Voxel-wise Prediction Network

Overview. After obtaining the general 3D feature volume, there are four branches applied to decode the voxel-wise properties of geometric and semantic for the panoptic 3D reconstruction task.

Occupancy prediction branch. Given a voxel-wise feature, the occupancy prediction branch can predict the per-voxel occupancy score $\hat{O}_k$. The voxel whose occupancy score is lower than a predefined threshold θ will be sparsified. The supervision of this branch is defined as cross-entropy(BCE) between the predicted occupancy score and the ground-truth occupancy values.

TSDF prediction branch. After the occupancy analysis, our model will obtain a set of features of occupied voxels. We apply a simple MLP for regressing TSDF $\hat{T}_k$. The TSDF loss between the TSDF prediction and the groundtruth TSDF is defined as: $L_T = |\ell(\hat{T}_k) - \ell(T_k)|$, where $\ell(x) = sgn(x)\log(|x| + 1)$ is the log scale function and $sgn(\cdot)$ is the sign function.

Semantic prediction branch. Paralleled with the TSDF prediction branch. We apply an additional MLP to produce semantic scores of classes we considered for each occupied voxel. The class with highest score will be be regraded as the predicted semantic label $\hat{S}_k$. The cross-entropy loss is used to train this branch.

Offset prediction branch. With the semantic predictions, our model can distinguish voxels of different semantic classes, but still may fail to separate two instances who have

the same semantic label, especially when they're close together. Inspired by [62], we also estimate the voxel-wise center shift vector $\triangle\hat{x}_k$, which represents the offset from each voxel to its instance center. The instance center is defined as the coordinate mean of all voxels belong to this instance. The predicted 3D shift vector $\triangle\hat{x}_k$ is supervised by L1 loss.

3.3. Instance Detection, Tracking and Fusion

Overview. The panoptic 3D reconstruction task requires instance-level reconstruction rather than voxel-level reconstruction. Hence, we perform a clustering-based instance segmentation scheme to detect all potential instance objects $\mathcal{O}_i = \{O_{i,m}\}$ in each fragment $\mathcal{F}_i$. Meanwhile, we maintain a set of global instance object reconstruction $\mathcal{O}^g_{i-1} = \{O^g_{i-1,n}\}$ from previous fragments. In order to associate instance detections between $\mathcal{O}_i$ and $\mathcal{O}^g_{i-1}$ and integrate instance detections from different fragments to obtain a globally consistent 3D instance object reconstruction, we introduce an instance-level tracking and fusion module. An illustration of this procedure is shown in Fig. 3.

Voxel clustering. Once we have geometric and semantic primitives for each occupied voxel, we group the voxels to form instance in each local fragment $\mathcal{F}_i$. With the estimated shift vector $\triangle\hat{x}_k \in \mathbb{R}^3$, we shift every voxel x_k toward its instance center, making the voxels of the same instance spatially closer to each other, defined as:

$$x^{'}_k = x_k + \triangle\hat{x}_k \tag{1}$$

Then, we ignore the voxels belong to the background (*stuff* region) and use the predicted semantic label $\hat{S}$ to split voxels of different foreground classes. With the shifted coordinate $x^{'}_k$, we adopt a simple yet efficient mean-shift clustering algorithm [63] to perform intra-class instance segmentation. Then, we get the instance detection sets $\mathcal{O}_i$ in current fragment $\mathcal{F}_i$.

Instance tracking. In order to get the correspondences between two sets of object instances, $\mathcal{O}_i$ and $\mathcal{O}^g_{i-1}$, we first compute the similarity $S(m,n)$ between instance $O_{i,m} \in \mathcal{O}_i$ and every instance $O^g_{i-1,n} \in \mathcal{O}^g_{i-1}$. We initialize instance embedding $e_{i,m}$ of each instance $O_{i,m}$ by average pooling all voxel features of this instance. Inspired by [59, 64], we resort to a GNN with attention mechanism [65] to allow message passing between different instances, and then get augmented embedding vectors $\bar{e}_{i,m}$ and $\bar{e}^g_{i-1,n}$, of $O_{i,m}$ and $O^g_{i-1,n}$ respectively. The similarity between them is formulated as:

$$S(m,n) = \left\langle \bar{e}_{i,m}, \bar{e}^g_{i-1,n} \right\rangle \tag{2}$$

Then we use a differentiable Sinkhorn algorithm [66, 67] to solve the optimal matching $\mathcal{M}^*$:

$$\mathcal{M}^* = \arg\min_{\mathcal{M}} \sum_{m,n} \mathcal{M}(m,n) S(m,n) \tag{3}$$

Where $\mathcal{M}(m,n) \in \{0,1\}$. We will include the loss function of this module in the supplementary material.

Integration to global map. After instance tracking, there are three possible cases, as shown in Fig. 3. If some global instances in $\mathcal{O}^g_{i-1}$ do not find their correspondences since they may not be visible in current fragment $\mathcal{F}_i$, we will keep these instances unchanged. In addition, for a local instance in $\mathcal{O}_i$ may not find its correspondence in global set, it could be a new instance observed for the first time, so we initialize this new instance to $\mathcal{O}^g_i$. Lastly, if two instances, $O_{i,m}$ and $O^g_{i-1,n}$ have been successfully matched, we will integrate them into a global instance $O^g_{i,n}$ and adopt a GRU module to fuse their instance embeddings:

$$e^g_{i,n} = \mathrm{GRU}(\bar{\mathrm{e}}^{\mathrm{g}}_{\mathrm{i-1,n}}, \bar{\mathrm{e}}_{\mathrm{i,m}}) \tag{4}$$

where $\bar{e}^g_{i-1,n}$ and $\bar{e}_{i,m}$ are used as the hidden state and input to the GRU module.

3.4. Implementation Details

Our 2D CNN architecture is a Feature Pyramid Network [68] using EfficientNetV2-S [69] as backbone, and we use torchsparse [70] for 3D sparse convolution. Our network is trained using Adam optimizer on four Nvidia RTX 3090 GRU for 60 epochs. Following [7], we use three coarse-to-fine layers and set the voxel size of each layer as 16cm, 8cm, 4cm respectively. The predefined distance $\triangle d$ is set to be 32cm. The TSDF truncation distance is set as three times of the voxel size for each layer. The occupancy pruning threshold is set to 0.

4. Experiments

4.1. Dataset, Metrics and Baseline.

Dataset. We perform our experiments on ScanNet(V2) [24]. ScanNet consists of 1613 scans across 807 distinct scenes with RGB images, groundtruth depths, camera poses, surface reconstructions and instance-level semantic segmentations. Following previous works [6, 7], we use the official train/val/test split to train and evaluate our method in order to make a fair comparison.

Metrics. We follow the 3D geometry metrics used in [6, 7] to evaluate the 3D geometry reconstruction performance of our approach. Among these 3D metrics, we regard the *F-score* as the most important metric to measure 3D geometry reconstruction quality since both the accuracy and

completeness of the reconstruction are taken into account. Inspired by [6], we transfer the semantic and instance label from the predicted mesh to the groundtruth mesh using nearest neighbor lookup on each vertices. The standard mIoU, mAP@50 and mAP@25 are used to evaluate the prediction of semantic label and instance label respectively. The detailed definitions of all these metrics above are included in the supplementary material.

Baseline. For the evaluation of 3D geometry reconstruction, we compare our method with multi-view depth estimation methods [1–5] and end-to-end reconstruction methods [6–9]. Among them, NeuralRecon [68] and Zuo et al. [9] are two end-to-end incremental volumetric reconstruction methods that are the most relevant ones to our approach. In order to evaluate the prediction of semantic label , we compare our method with some prior methods that include depth as input [14–16, 24, 36, 71–73], as well as Atlas [6] using only RGB images as input, as we do. As for the evaluation of instance label prediction, in addition to some previous works with 3D input [17, 74–78], we also compare with an original but classic algorithm [56], which also include only 2D images as input.

4.2. Results

Evaluation of 3D Geometry Reconstruction. The experimental results of 3D reconstruction on ScanNet dataset are shown in Tab. 1. Our method outperforms existing online feature fusion methods [7, 9], and slightly better than the state of the art depth fusion method [5] in terms of *F-score* metric. The advantage on *F-score* is due to the fact that our method have achieved a good balance between the accuracy and completeness. With the assistance of MVS depth, our method can recover more complete and detailed geometry than the pure feature fusion method [7]. Thanks to the de-noise ability of feature fusion, our method can recover more coherent and accurate geometry than [5]. The qualitative comparison and analysis are shown in Fig. 4. Our method presents more complete and coherent reconstruction of the whole scene while preserving many details of foreground objects. Moreover, we additionally conduct evaluation of geometry reconstruction at instance level, and provide average metrics for each semantic category in the supplementary material.

Evaluation of 3D Panoptic Segmentation. The quantitative results of 3D semantic segmentation on ScanNetV2 are reported in Tab. 2. Despite the unfair setup since we use only RGB data, our method is still surprisingly competitive with(even beats) some prior works with 3D input data. In the same case with only RGB images as input, our method outperforms Altas by a large margin(+18.4) in terms of mIoU. As the qualitative comparison shown in Fig. 5, the advance of performance is due to the improvement of both reconstruction quality and segmentation accuracy. It is worth mentioning that due to the setting of online panoptic reconstruction, our method do not have the access to full batch data with global context, which is extremely challenging for geometry reconstruction and semantic segmentation. In addition, we also report the 3D instance segmentation results in Tab. 3. In absence of depth sensor and post process procedure, such as Non-max suppression(NMS), our method still achieves decent performance of instance segmentation. The quantitative result of 3D instance segmentation is illustrated in the Fig. 2 of the supplementary material, our method is able to accurately reconstruct while successfully splitting the 3D scene into multiple instance objects.

	Method	Online	Comp↓	Acc↓	Recall↑	Prec↑	**F-score↑**
Depth Fusion	COLMAP [1]	-	6.9	13.5	0.634	0.505	0.558
	MVDNet [2]	-	**4.0**	24.0	**0.831**	0.208	0.329
	DPSNet [3]	-	4.5	28.4	0.793	0.223	0.344
	GPMVS [4]	-	10.5	19.1	0.423	0.339	0.373
	SimRec [5]	-	6.2	**10.1**	0.636	**0.536**	**0.577**
Feature Fusion	Atlas [6]	x	8.3	10.1	0.566	0.600	0.579
	Vortx [8]	x	**8.1**	**6.2**	**0.605**	**0.689**	**0.643**
	NeuRec [7]	✓	13.7	**5.6**	0.470	**0.678**	0.553
	Zuo et al. [9]	✓	11.0	5.8	0.505	0.665	0.572
	Ours	✓	**8.9**	6.4	**0.530**	0.656	**0.584**

Table 1. **Quantitative Result of 3D Geometry Reconstruction on ScanNetV2 test set.** We highlight the best results for Depth Fusion, Offline Feature Fusion and Online Feature Fusion methods in green, magenta, and cyan, respectively. The experimental results of Simplerecon [5] are reproduced with the official codebase and published weights. Other results come from [9].

Method	with Depth	mIoU↑
ScanNet [24]	✓	30.6
PointNet++ [14]	✓	33.9
SPLATNet [71]	✓	39.3
3DMV [72]	✓	48.4
SegFusion [73]	✓	51.5
PanopticFusion [36]	✓	52.9
SparseConvNet [15]	✓	72.5
MinkowskiNet [16]	✓	**73.4**
Altas [6]	x	34.0
Ours	x	52.4

Table 2. **Quantitative Result of 3D Semantic Segmentation on ScanNetV2 test set.** The experiment results of other methods are borrowed from Altas [6] and the ScanNetV2 benchmark [24].

Ablation Study of different designs of choices. In Tab. 4, we conduct ablation experiments to verify the effect of different designs of choices. We experiment with different number of keyframes in each fragment(row (a), (b)) rather than the default 9 views(row (f)), and replace Mnas-

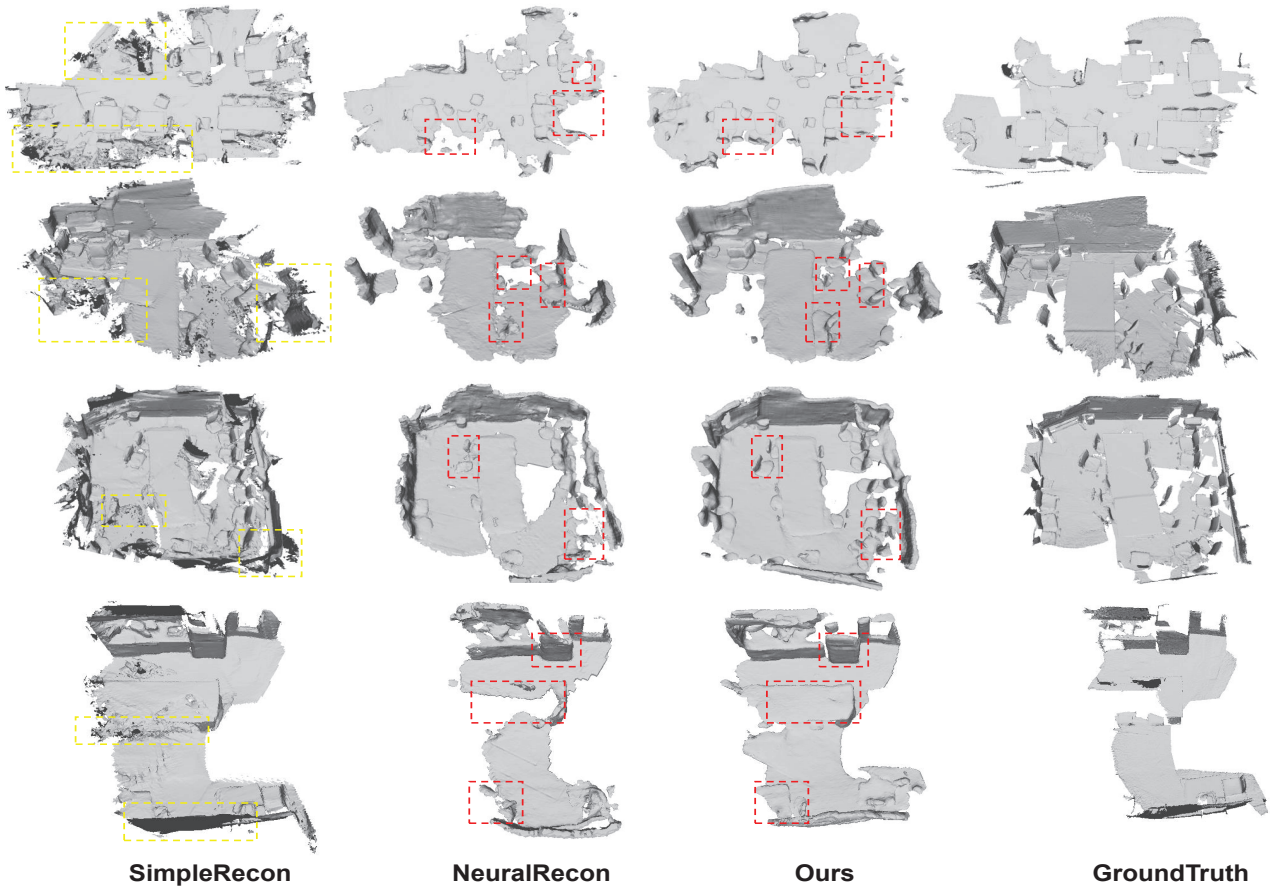

Figure 4. **Qualitative Result of 3D Geometry Reconstruction on ScanNet val set.** Since SimpleRecon relies on the non-learnable depth fusion method, it generates artifacts and duplicate surfaces (highlighted with the yellow boxes). With the incorporation of MVS depth, our geometry reconstruction results are more complete while preserving more details of foreground objects(highlighted with the red boxes) compared with NeuralRecon.

Method	with Depth	AP50↑	AP25↑
SGPN [74]	✓	0.143	0.390
ASIS [75]	✓	0.199	0.422
Gspn [76]	✓	0.306	0.544
3D-SIS [17]	✓	0.382	0.558
SegGroup [77]	✓	0.445	0.637
PBNet [78]	✓	**0.747**	**0.825**
MaskRCNN [56]	x	0.058	0.261
Ours	x	0.227	0.484

Table 3. **Quantitative Result of 3D Instance Segmentation on ScanNetV2 test set.** The experiment results of other methods are borrowed from the ScanNetV2 benchmark [24].

Net [79] as backbone(row (c)) instead of EfficientNetV2-S(row (d). We also conduct an ablation study to verify the effect of depth guidance strategy(row (d), (e), (g)). By comparing the row (d) and row (f), it is obvious that introducing the strategy of depth-guided feature allocation brings considerable improvement to all metrics. We also attempt to adopt the traditional TSDF-Fusion to form TSDF volume, and then enhance the feature volume by concatenating with the TSDF Volume(row(e)). But this strategy does not bring the expected improvement, so we abandon it in our final scheme. There is a significant improvement of geometric quality by directly using gt depth as input without any fine-tuning or re-training(row (g)), which indicates our method can benefit from more advanced depth estimation models. In addition, the ablation of tracking and fusion module is detailed in supplementary material.

Runtime analysis. We conduct runtime evaluation on a desktop computer equipped with Intel i9-12900KS CPU and NVIDIA RTX3090 GPU. Our method takes an average of 700 ms to process one fragment including 9 key frames, which composed of 430ms for MVS Network, 30 ms for 2D CNN, 137 ms for 3D Network(consists of 3D CNN, GRU and voxel-wise prediction network), and 103 ms for

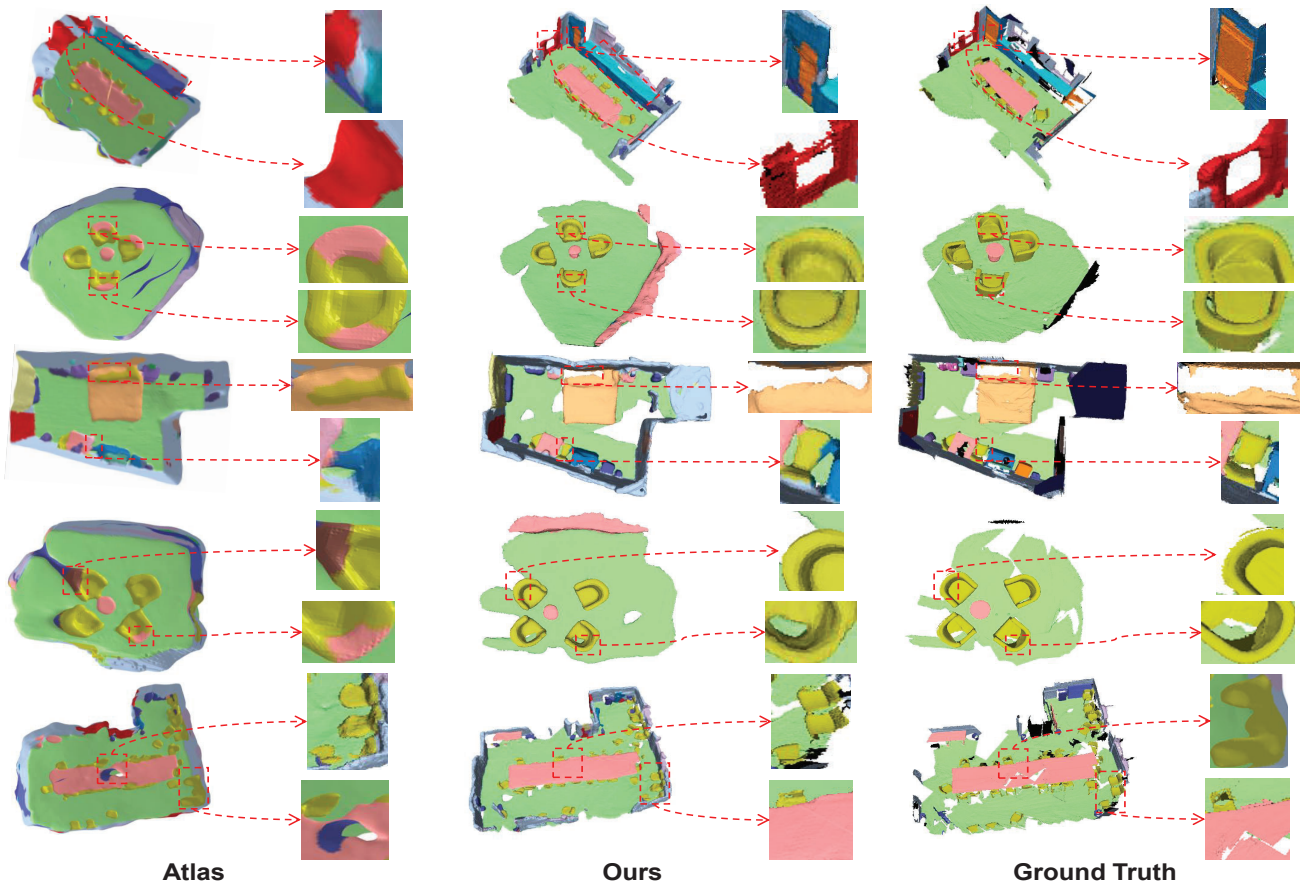

Figure 5. **Qualitative Result of 3D Semantic Segmentation on ScanNetV2 val set.** Our method produces more accurate geometry reconstruction and semantic labeling than Atlas, though Atlas have the access to full batch data with global context.

Method	Recall↑	Prec↑	**F-score↑**	mIoU↑
(a) 7 views.	0.580	0.680	0.623	**54.3**
(b) 11 views.	0.552	0.666	0.600	52.5
(c) MnasNet.	0.551	0.655	0.596	52.6
(d) Ours w/o guide proj.	0.548	0.661	0.597	53.6
(e) Ours with tsdf agu.	0.582	0.679	0.624	53.7
(f) Ours	**0.593**	**0.681**	**0.631**	54.2
(g) with depth sensor.	0.761	0.874	0.810	55.5

Table 4. **Ablation of different designs of choices on ScanNetV2 val set.**

instance detection, tracking and fusion. Since keyframes are created at a far lower frequency than the framerate, our model still achieves a real-time panoptic reconstruction of 12.85 key frames per second (KFPS).

5. Conclusion

In this work, we introduce the task of panoptic 3D reconstruction from a posed monocular video. This unified task aims to obtain the holistic understanding of global scene consisting of geometric reconstruction, 3D semantic segmentation and 3D instance segmentation. We also present a novel system, PanoRecon, for real-time panotic 3D reconstruction. The main idea of PanoRecon is to use the volumetric feature representation to perform geometry reconstruction and panoptic segmentation fragment by fragment. Experiments show that PanoRecon achieves competitive geometry reconstruction compared with the state-of-the-art methods and promising 3D panoptic segmentation, while running in real-time. We hope this work will help to push forward research towards more comprehensive holistic scene understanding and introduce new algorithm challenges and additional insights to this area.

Acknowledgement

We thank anonymous reviewers and AC for their fruitful comments and suggestions. This work is supported by NSFC (U22A2061, 62176010), and 230601GP0004.

References

[1] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 1, 3, 6

[2] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International conference on 3d vision (3DV)*, pages 248–257. IEEE, 2018. 3, 6

[3] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 6

[4] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019. 6

[5] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 1, 3, 6

[6] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020. 1, 3, 5, 6

[7] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 3, 4, 5, 6

[8] Noah Stier, Alexander Rich, Pradeep Sen, and Tobias Höllerer. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In *2021 International Conference on 3D Vision (3DV)*, pages 320–330. IEEE, 2021. 4, 6

[9] Xingxing Zuo, Nan Yang, Nathaniel Merrill, Binbin Xu, and Stefan Leutenegger. Incremental dense reconstruction from monocular video with guided sparse feature volume fusion. *IEEE Robotics and Automation Letters*, 2023. 3, 6

[10] Jihong Ju, Ching Wei Tseng, Oleksandr Bailo, Georgi Dikov, and Mohsen Ghafoorian. Dg-recon: Depth-guided neural 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18184–18194, 2023. 3

[11] Huiyu Gao, Wei Mao, and Miaomiao Liu. Visfusion: Visibility-aware online 3d scene reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17317–17326, 2023. 3

[12] Noah Stier, Anurag Ranjan, Alex Colburn, Yajie Yan, Liang Yang, Fangchang Ma, and Baptiste Angles. Finerecon: Depth-aware feed-forward network for detailed 3d reconstruction. *arXiv preprint arXiv:2304.01480*, 2023. 1, 3

[13] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 3

[14] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 6

[15] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 6

[16] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 2, 3, 6

[17] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019. 2, 3, 6, 7

[18] 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. *Advances in neural information processing systems*, 32, 2019. 3

[19] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 3

[20] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020.

[21] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021.

[22] Leyao Liu, Tian Zheng, Yun-Jou Lin, Kai Ni, and Lu Fang. Ins-conv: Incremental sparse convolution for online 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18975–18984, 2022. 2, 3

[23] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 2

[24] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 5, 6, 7

[25] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M Rehg. Joint semantic segmentation and 3d re-

construction from monocular video. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 703–718. Springer, 2014. 2

[26] Christian Hane, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3d scene reconstruction and class segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 97–104, 2013. 2

[27] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017. 2, 3

[28] Shuaifeng Zhi, Michael Bloesch, Stefan Leutenegger, and Andrew J Davison. Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11776–11785, 2019.

[29] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1689–1696. IEEE, 2020. 2

[30] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1352–1359, 2013. 2

[31] Edgar Sucar, Kentaro Wada, and Andrew Davison. Nodeslam: Neural object descriptors for multi-view shape reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 949–958. IEEE, 2020. 2

[32] Jingwen Wang, Martin Rünz, and Lourdes Agapito. Dsp-slam: Object oriented slam with deep shape priors. In *2021 International Conference on 3D Vision (3DV)*, pages 1362–1371. IEEE, 2021. 2

[33] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478. IEEE, 2017. 2

[34] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level slam. In *2018 international conference on 3D vision (3DV)*, pages 32–41. IEEE, 2018.

[35] Binbin Xu, Wenbin Li, Dimos Tzoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5231–5237. IEEE, 2019.

[36] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212. IEEE, 2019. 3, 6

[37] Leevi Raivio and Esa Rahtu. Online panoptic 3d reconstruction as a linear assignment problem. In *International Conference on Image Analysis and Processing*, pages 39–50. Springer, 2022. 2

[38] Weicai Ye, Xinyue Lan, Shuo Chen, Yuhang Ming, Xingyuan Yu, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Pvo: Panoptic visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2023. 2

[39] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[41] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[42] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 2

[43] Shuaifeng Zhi, Edgar Sucar, Andre Mouton, Iain Haughton, Tristan Laidlow, and Andrew J Davison. ilabel: Interactive neural scene labelling. *arXiv preprint arXiv:2111.14637*, 2021. 2

[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 2

[45] 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 *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 2

[46] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021. 2

[47] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022.

[48] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21764–21774, 2023. 2

[49] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vmap: Vectorised object mapping for neural field slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 952–961, 2023. 2

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

[51] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, 2021. 3

[52] Haobo Yuan, Xiangtai Li, Yibo Yang, Guangliang Cheng, Jing Zhang, Yunhai Tong, Lefei Zhang, and Dacheng Tao. Polyphonicformer: unified query learning for depth-aware video panoptic segmentation. In *European Conference on Computer Vision*, pages 582–599. Springer, 2022.

[53] Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Bin Luo, Jun-Yan He, Jin-Peng Lan, Yifeng Geng, and Xuansong Xie. Towards deeply unified depth-aware panoptic segmentation with bi-directional guidance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4111–4121, 2023. 3

[54] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 3

[55] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[56] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 6, 7

[57] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. Volumetric instance-aware semantic mapping and 3d object discovery. *IEEE Robotics and Automation Letters*, 4(3):3037–3044, 2019. 3

[58] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 3

[59] Yiming Xie, Matheus Gadelha, Fengting Yang, Xiaowei Zhou, and Huaizu Jiang. Planarrecon: Real-time 3d plane detection and reconstruction from posed monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6219–6228, 2022. 3, 5

[60] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34:8282–8293, 2021. 3

[61] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2023. 3

[62] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 5

[63] Dorin Comaniciu and Peter Meer. Mean shift analysis and applications. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1197–1203. IEEE, 1999. 5

[64] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 5

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[66] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 5

[67] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 5

[68] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5, 6

[69] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 5

[70] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, pages 685–702. Springer, 2020. 5

[71] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2530–2539, 2018. 6

[72] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 6

[73] Davide Menini, Suryansh Kumar, Martin R Oswald, Erik Sandström, Cristian Sminchisescu, and Luc Van Gool. A real-time online learning framework for joint 3d reconstruction and semantic segmentation of indoor scenes. *IEEE Robotics and Automation Letters*, 7(2):1332–1339, 2021. 6

[74] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018. 6, 7

[75] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2019. 7

[76] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 7

[77] An Tao, Yueqi Duan, Yi Wei, Jiwen Lu, and Jie Zhou. Seggroup: Seg-level supervision for 3d instance and semantic segmentation. *IEEE Transactions on Image Processing*, 31:4952–4965, 2022. 7

[78] Weiguang Zhao, Yuyao Yan, Chaolong Yang, Jianan Ye, Xi Yang, and Kaizhu Huang. Divide and conquer: 3d point cloud instance segmentation with point-wise binarization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 562–571, 2023. 6, 7

[79] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828, 2019. 7