Supplementary Materials for: 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

1. Depth-guided Feature Volume Construction

The depth-guided feature volume construction is detailed in Fig. 1. We use an efficient MVS network [1] to obtain the depth estimations of key frames. And the 2D feature representations of key frames are extracted using a 2D convolution neural network [2], and then are back projected to voxels within a predefined distance from the corresponding estimated depth surface along the ray. The view-independent feature volume is obtained by directly averaging the features from different views. We gradually build the feature volume in a coarse-to-fine pyramid paradigm with the hierarchy 2D image feature. At each pyramid level, we perform occupancy classification for each voxel, and only occupied voxels are further upsampled and passed to the next pyramid level. At the final fine level, we obtain a sparse feature volume consisting of a set of occupied voxels, which are used to perform 3D panoptic reconstruction.

2. Loss Function

Occupancy Classification. The occupancy loss L_O is defined as the binary cross-entropy (BCE) between the predicted occupancy score and the ground-truth occupancy value. The supervision is applied to all the coarse-to-fine levels.

TSDF Regression. The TSDF loss L_T between the TSDF prediction $\hat{T}_k$ and the groundtruth TSDF T_k is formulated 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. The supervision is applied to all the coarse-to-fine levels.

Semantic Classification. The semantic loss L_S is defined as the cross-entropy (CE) between the predicted semantic score and the ground-truth semantic category. The supervision is applied to all the coarse-to-fine levels.

Offset Regression. The offset loss L_D is defined as L1 loss between the predicted 3D displacement and the groundtruth 3D displacement. The supervision is applied to the final fine level.

Differentiable Matching. Denote the ground-truth matching labels $\mathcal{M} = \{(i,j)\} \subset \mathcal{A} \times \mathcal{B}$ between the local instance detections $\mathcal{O}_i$ from the current fragment $\mathcal{F}_i$ and the global instance reconstruction $\mathcal{O}^g_{i-1}$ from all previous fragments. But there are some instances in $\mathcal{O}_i$ may not find their correspondences in $\mathcal{O}^g_{i-1}$ and vice versa. We denote the unmatched labels for $\mathcal{O}_i$ and $\mathcal{O}^g_{i-1}$ as $\mathcal{I} \subseteq \mathcal{A}$ and $\mathcal{J} \subseteq \mathcal{B}$, respectively. Given these labels, we minimize the negative log-likelihood of the optimal matching matrix $\mathcal{M}^*$ as follows inspired by [3, 4].

$$\begin{aligned} L_M = &- \sum_{(i,j)\epsilon\mathcal{M}} \log \mathcal{M}^*_{i,j} - \sum_{i\epsilon\mathcal{I}} \log \mathcal{M}^*_{i,N+1} \\ &- \sum_{j\epsilon\mathcal{J}} \log \mathcal{M}^*_{M+1,j} \end{aligned} \tag{1}$$

where M and N are the number of instances in $\mathcal{O}_i$ and $\mathcal{O}^g_{i-1}$, respectively.

Joint Loss. PanoRecon is optimized by a joint loss consisting of several loss terms:

$$L = \alpha_1 L_O + \alpha_2 L_T + \alpha_3 L_S + \alpha_4 L_D + \alpha_5 L_M \tag{2}$$

3. Evaluation Metrics

3.1. Metrics of 3D Geometry Reconstruction

The definitions of metrics used to evaluate 3D Geometry Reconstruction are detailed in Tab. 1

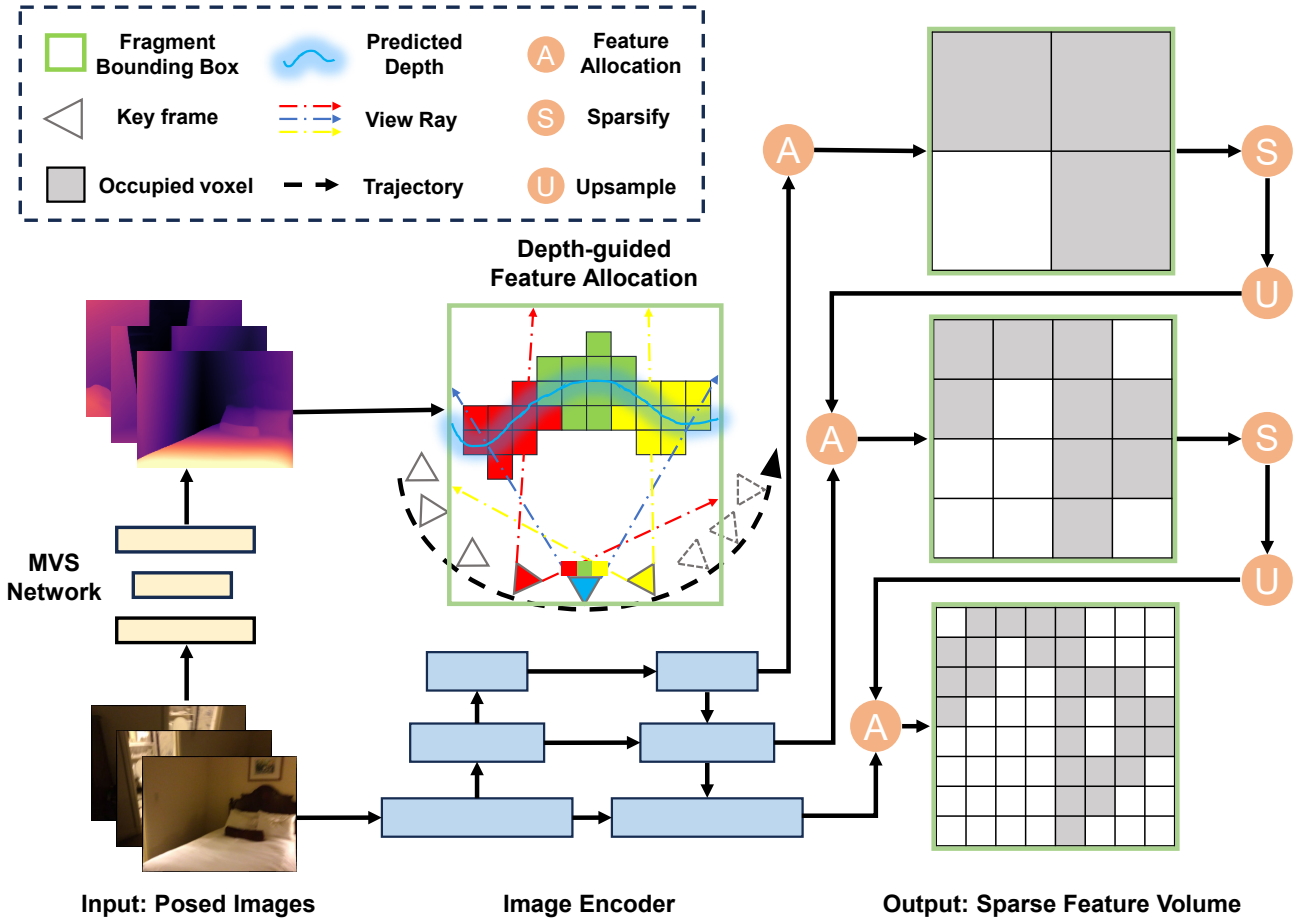

Figure 1. **2D illustration of depth-guided feature volume construction.**

Metrics	Definition
Comp	$mean_{p^* \in P^*}(min_{p \in P}\|\|p - p^*\|\|)$
Acc	$mean_{p \in P}(min_{p^* \in P^*}\|\|p - p^*\|\|)$
Recall	$mean_{p^* \in P^*}(min_{p \in P}\|\|p - p^*\|\| < .05)$
Prec	$mean_{p \in P}(min_{p^* \in P^*}\|\|p - p^*\|\| < .05)$
F-score	$\frac{2 \times Recall \times Prec}{Recall + Prec}$

Table 1. **Metrics of 3D Geometry Reconstruction.** P and P^* are the predicted and ground truth point clouds.

3.2. Metrics of 3D Semantic Segmentation

Following [6], we transfer the semantic labels from the predicted mesh into the ground truth mesh using nearest neighbor lookup on the vertices. And use the standard mIoU(Mean Intersection over Union) to evaluate the quality of 3D Semantic Segmentation. The definition of mIoU is formulated as:

$$\begin{aligned} mIoU &= \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}} \\ &= \frac{1}{k}\sum_{i=1}^{k}\frac{TP}{FN + FP + TP} \end{aligned} \tag{3}$$

where k represents the total number of categories, i represents groundtruth value, j represents predicted value, and p_{ij} means predicting i as j.

3.3. Metrics of 3D Instance Segmentation

We transfer the instance labels from the predicted mesh into the ground truth mesh using nearest neighbor lookup on the vertices, and use standard mAP@50 and mAP@25 to evaluate the prediction of instance label. The definitions of mAP and related metrics as follows.

Method	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	otherfu.	picture	refrige.	s. curtain	sink	sofa	table	toilet	window	avg.
NeuRec [5]	60.5	68.0	61.8	60.7	58.7	69.9	43.5	65.0	50.9	64.7	58.4	60.5	63.9	71.1	63.4	64.3	67.3	36.9	60.5
Ours	**71.8**	**70.7**	**62.6**	**67.6**	**65.9**	**74.3**	**51.0**	**72.0**	**60.2**	**70.3**	**61.0**	**65.7**	**71.4**	**74.7**	**67.8**	**69.2**	**77.7**	**45.7**	**66.6**

Table 2. **Quantitative Result of Instance-level 3D Reconstruction on ScanNetV2 val set.** We compare our method with a representative online feature fusion method [5], and report average F-score [%] metric for each semantic foreground category(*things* category).

IoU (Intersection over Union). To decide whether a prediction is correct w.r.t to an object or not, IoU is used. It is defines as the intersection between the predicted instance mask and actual instance mask divided by their union.

$$IoU = \frac{Area\,of\,Overlap}{Area\,of\,Union} \tag{4}$$

A prediction is considered to be True Positive if $IoU \geq threshold$, and False Positive if $IoU < threshold$.

Precision and Recall. Before introducing mAP, we present the definitions of precision and recall first. Recall is the True Positive Rate i.e. Of all the actual positives, how many are True positives predictions. Precision is the Positive prediction value i.e. Of all the positive predictions, how many are True positives predictions.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

mAP (mean Average Precision). In order to calculate mAP, first, we need to calculate AP per class. For each categoty, there are K prediction-groundtruth mask pairs $\{M_i^{pred}, M_i^{gt}\}_{i=1}^{K}$, if $IoU(M_i^{pred}, M_i^{gt}) > threshold$, as well as K confidence scores $\{conf\}_{i=1}^{K}$, correspondingly. In order of confidence score, we will obtain a curve where precision changes as recall increases, called PR-curve. The area under the PR-curve refers to the AP(Average Precision) of this categoty. The mAP is calculated by finding Average Precision(AP) for each class and then average over all N classes.

$$mAP = \frac{1}{N}\sum_{j=1}^{N} AP_j \tag{7}$$

mAP@50 indicates that $threshold$ is set to be 0.5, and mAP@25 indicates that $threshold$ is set to be 0.25.

4. Evaluation on Instance-level 3D Reconstruction

As shown in Fig. 2, our method is able to accurately reconstruct while successfully splitting the 3D scene into multiple instance objects. For the evaluation of instance-level 3D Reconstruction, we obtain instance-level mesh by using the groundtruth instance-level bounding box to crop scene-level mesh. As shown in Tab. 2, we compare our method with a representative online feature fusion method [5] in terms of F-score metric. It is obviously that our method greatly outperform NeuralRecon [5] in the instance-level 3D reconstruction. With the assistance of MVS depth, our method can recover more complete and detailed geometry of foreground objects than the pure feature fusion method [5].

5. Ablation of tracking and fusion

We evaluate an offline instance segmentation paradigm that initially reconstructs the voxel map with semantic and geometric primitives of the entire scene and then performs voxel clustering to obtain the offline instance segmentation. The comparison in Tab. 3 reveals that the offline paradigm performs worse than online paradigm. As depicted in Fig. 3, the shifted coordinates of large object tend to be distributed across multiple clusters due to the limited receptive field of each fragment, leading to their segmentation into multiple instances. In contrast, the proposed tracking and fusion module in online paradigm can effectively match and fuse the same instance across different fragments.

Method	AP50↑	AP25↑
Ours with offline paradigm	0.223	0.471
Ours with online paradigm	**0.264**	**0.497**

Table 3. **Ablation of different paradigms on ScanNetV2 val set.**

6. Supplementary Video

In the supplementary video, we demonstrate the incremental panoptic 3D reconstruction process of PanoRecon in real-time applications.

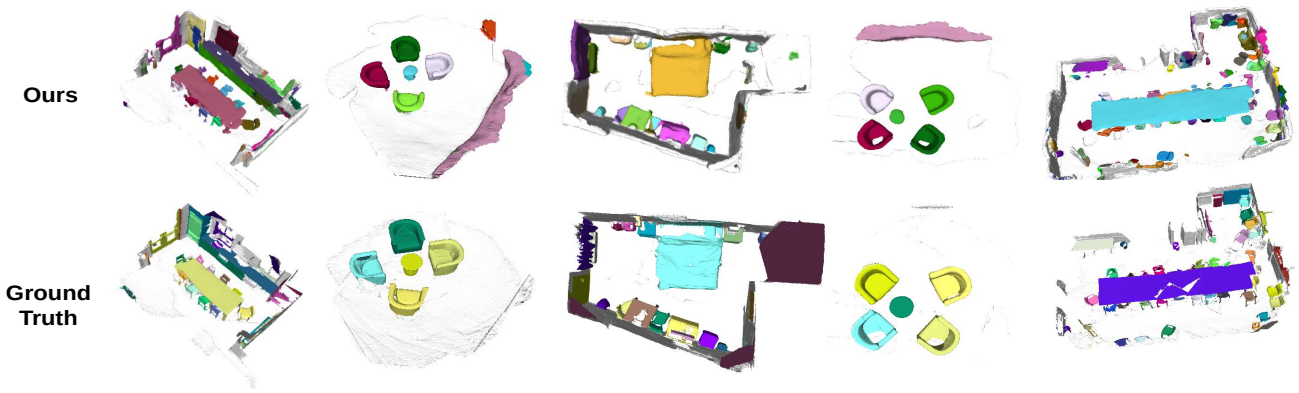

Figure 2. **Qualitative Result of 3D Instance Segmentation on ScanNetV2 val set.** We are able to accurately reconstruct while segmenting the 3D scene in instance-level despite without a depth sensor.

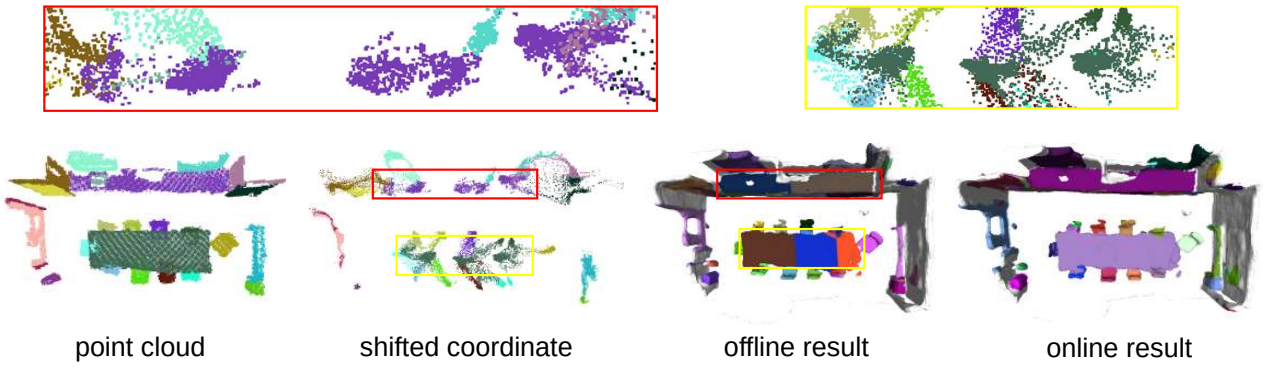

Figure 3. **Qualitative comparison of two different paradigms.**

References

[1] 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

[2] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 1

[3] 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. 1

[4] 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. 1

[5] 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

[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. 2