

MOSE: Monocular Semantic Reconstruction Using NeRF-Lifted Noisy Priors

Zhenhua Du¹, Binbin Xu², Haoyu Zhang¹, Kai Huo¹, Shuaifeng Zhi^{1,†}

arXiv:2409.14019v1 [cs.CV] 21 Sep 2024

Abstract—Accurately reconstructing dense and semantically annotated 3D meshes from monocular images remains a challenging task due to the lack of geometry guidance and imperfect view-dependent 2D priors. Though we have witnessed recent advancements in implicit neural scene representations enabling precise 2D rendering simply from multi-view images, there have been few works addressing 3D scene understanding with monocular priors alone. In this paper, we propose MOSE, a neural field semantic reconstruction approach to lift inferred image-level noisy priors to 3D, producing accurate semantics and geometry in both 3D and 2D space. The key motivation for our method is to leverage generic class-agnostic segment masks as guidance to promote local consistency of rendered semantics during training. With the help of semantics, we further apply a smoothness regularization to texture-less regions for better geometric quality, thus achieving mutual benefits of geometry and semantics. Experiments on the ScanNet dataset show that our MOSE outperforms relevant baselines across all metrics on tasks of 3D semantic segmentation, 2D semantic segmentation and 3D surface reconstruction.

Index Terms—Semantic Scene Understanding, Representation Learning, Deep Learning for Visual Perception.

I. INTRODUCTION

COMPREHENSIVELY understanding the high-level semantics of 3D scenes is crucial for various downstream applications, including augmented reality, robot navigation, and autonomous driving. Given the prohibitive costs of acquiring dense 3D scans and corresponding annotations, it would always be desirable if a dense semantic 3D map could be achieved simply from its multi-view 2D observations, benefiting from the widely available training corpora. There have been great related advancements using SfM and RGB-D vSLAM systems, projecting 2D semantic predictions into 3D space and fusing them via prescribed rules like Bayesian fusion [1], [2]. Nevertheless, addressing this task with monocular cues remains challenging, especially when confronted with inconsistent 2D semantic labels and inaccurately reconstructed geometry. As an attempt in addressing this task, in this paper,

Manuscript received: April, 7, 2024; Revised July, 10, 2024; Accepted August, 23, 2024.

This paper was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments.

Research presented in this paper has been partly supported by the NSFC (No. 62201603), the CPSF (No. 2023TQ0088, No. GZC20233539), and the Research Program of NUDT (No. ZK22-04).

[†]Zhenhua Du, Haoyu Zhang, Kai Huo, Shuaifeng Zhi[†] (*corresponding author*) are with National University of Defense Technology.

Emails: dzh11@nudt.edu.cn, zhanghaoyu19@nudt.edu.cn, huokai2001@163.com, zhishuaifeng@outlook.com.

²Binbin Xu: xu.binbin@outlook.com.

Source code: <https://github.com/ZhenhuaDu11/Mose>.

Digital Object Identifier (DOI): see top of this page.

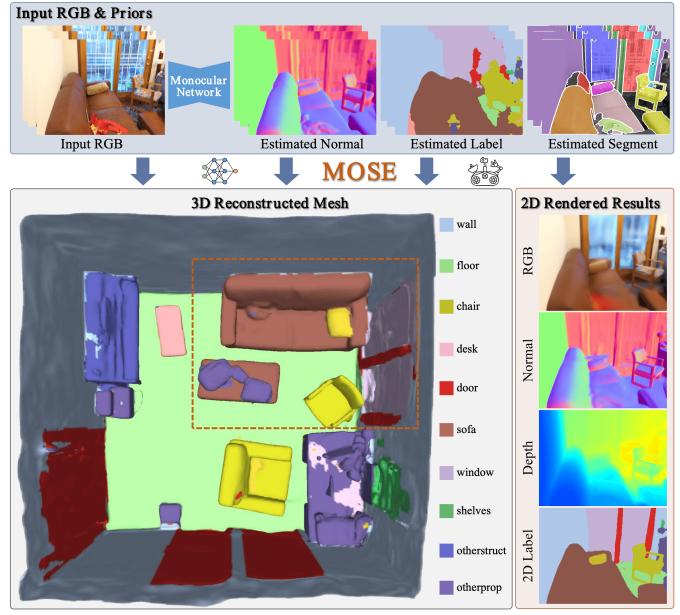


Fig. 1. 3D indoor semantic reconstruction. Taking RGB images and noisy 2D scene priors from monocular networks (upper portion), our method MOSE, is able to reconstruct the 3D smooth semantic map of the scene and render 2D associated results (bottom portion).

our objective is to reconstruct a consistent 3D semantic map of indoor scenes only with imperfect multi-view 2D priors.

Recently proposed neural implicit representations (NeRF) [3] have shown impressive performance in capturing intricate appearance and geometric details from only RGB images, with a clear 3D awareness of multi-view consistency. Semantic-NeRF [4] employs an extra MLP to represent the semantic field of the scene, demonstrating how 2D semantic predictions benefit from the self-similarity inherent in the compact scene encodings. However, due to the lack of sufficient surface constraints, Semantic-NeRF and its variants [5], [6] often produce floaters when extracting 3D semantic maps. VolSDF [7] and NeuS [8], on the other hand, improve geometry quality of NeRF by parameterizing the density field as an SDF field, and alternatively use depth and normal priors for even higher quality [9], [10]. Built upon the work of [4] and [7], Manhattan-SDF [11] learns a joint representation of scene geometry and semantics motivated by Manhattan-world assumption within indoor scenes. Though improved monocular 3D semantic mapping is achieved, [11] only considers three coarse semantic classes (wall, floor, and others), which limits its applicability in cases demanding detailed categorization.

In this paper, we propose MOSE, to concurrently achieve high-quality 3D geometry scans as well as finer-grained se-

mantic labeling purely from a series of 2D images along with imperfect priors including normals, semantics and segments. Our motivation comes from the observation that there is a lack of effective region-wise coherence within learned object semantics and texture-less geometry, leading to obvious defects in terms of incongruous semantic patches and fractured surfaces. Our MOSE exploits local segments as basis to encourage further smoothness of semantics, and furthermore relies on semantics to guide the learning process in texture-less regions. With negligible overhead during training, our approach produces accurate semantic 3D scans and surpasses existing baselines with a clear margin.

To summarize, we propose MOSE, an implicit monocular semantic reconstruction system, whose key contributions are listed as follows:

- We introduce a neural semantic reconstruction system capable of reconstructing smooth 3D semantic maps only from images and noisy 2D scene priors, which obtains state-of-the-art 3D semantic understanding performance on the challenging ScanNet scenes.
- To better utilize inconsistent semantic supervision, a locally-consistent fusion strategy using class-agnostic image segments (e.g., SAM [12], super-pixels [13]) is proposed, obtaining smooth and accurate semantics.
- To mitigate geometric degradation in texture-less regions, we introduce a semantically-weighted geometric regularization term to encourage stronger smoothness on dominant semantic structures.

II. RELATED WORK

Neural Implicit Representation. There is a trend to encode a scene into an implicit function by training a coordinate-based neural network [14], [15]. In particular, Neural Radiance Field (NeRF) [3] has opened up a line of research of representing scenes as volumetric density fields to learn geometry and appearance simultaneously. However, the density-based methods face challenges when extracting high-fidelity surfaces. Following-up works Neus [8] and VolSDF [7] correlate the signed distance function (SDF) with density, enabling accurate surface reconstruction. To reconstruct larger scenes, estimated geometric priors such as depth and normal maps, are often introduced to aid representation learning [9], [10], [16]. Among these methods, NeURIS [9] effectively reconstructs room-scale scenes by adaptively using noisy normal priors. However, due to the inherent shape ambiguity of implicit representation when fitting multi-view training images [17], degradation still occurs in texture-less regions even when the rendering results align well with 2D supervision. In this work, we propose a compact implicit representation that captures scene appearance, geometry and semantics from only RGB and 2D noisy neural priors.

Neural Semantic Reconstruction. Most traditional semantic reconstruction methods [1], [2] typically project 2D predicted labels into a fused semantic map. Due to the lack of perceptual awareness of neighboring pixels, they often struggle to handle complex scenarios. Utilizing neural networks [18], SceneCode [19] encodes the scene into compact latent codes

and jointly optimize the geometry and semantics. Inspired by the success of NeRF [3], Semantic-NeRF [4] and its following-up works [5], [6] attempt to learn a 3D consistent semantic field via lifting 2D semantic labels. However, these methods solely learn the density field and face challenges in extracting a high-quality 3D semantic map directly. Adopting the surface representation of VolSDF [7], Manhattan-SDF [11] simultaneously learns scene geometry and semantics, enabling monocular 3D semantic mapping of indoor scenes. However, [11] is limited to three coarse semantic classes and exhibits fluctuating surface geometry. In this work, we propose MOSE, aiming to simultaneously achieve accurate 3D geometry scans as well as smooth semantic labeling.

III. METHOD

Scanning the scene with a monocular camera, our objective is to reconstruct an accurate and smooth 3D semantic map of indoor scenes. As illustrated in Fig. 1, MOSE takes as input posed RGB images $\{\mathcal{I}_k\}$, as well as noisy 2D scene priors from off-the-shelf predictors, including semantics class labels $\{\mathcal{S}_k\}$, estimated normals $\{\mathcal{N}_k\}$ and class-agnostic segment masks $\{\mathcal{M}_k\}$. To faithfully reconstruct semantic 3D scans, we encode scene geometry, appearance and semantics into a neural field representation and conduct joint optimization to achieve mutual benefits (Sec. III-A). To overcome the inherent discontinuity of 2D semantic predictions [20], [21], we propose a locally-consistent fusion strategy (LCF) (Sec. III-B) to improve semantics’ local coherence and smoothness by leveraging generic 2D segment masks as a prior, which has been found to be particularly useful in producing accurate object-level semantics. In addition, we show further benefits of acquiring accurate semantics by improving geometry quality using the proposed semantically-weighted geometric regularization (SGR) (Sec. III-C).

A. Scene Representation and Rendering

MOSE adopts a NeRF-based representation to compactly encode scene appearance, geometry and semantics (see Fig. 2). Specifically, three MLPs are used: for spatial points $\mathbf{x} = (x, y, z)$ and their viewing direction $\mathbf{v} = (\theta, \phi)$, a color MLP encodes the appearance as a color field $\mathbf{c}(\mathbf{x}, \mathbf{v})$, an SDF MLP represents the geometry as a signed distance function (SDF) field $d(\mathbf{x})$, and a semantic MLP encodes its semantic class label as a semantic field $\mathbf{s}(\mathbf{x})$.

To learn SDF field, we adopt the neural surface representation from NeuS [8], which transforms the volume density $\sigma(\mathbf{x})$ to an SDF value $d(\mathbf{x})$. We sample 3D points \mathbf{x}_i along the camera ray \mathbf{r} , and utilize volume rendering to accumulate its color \mathbf{c}_i , normal \mathbf{n}_i and semantic logits \mathbf{s}_i :

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^N w_i \mathbf{c}_i, \hat{\mathbf{N}}(\mathbf{r}) = \sum_{i=1}^N w_i \mathbf{n}_i, \hat{\mathbf{S}}(\mathbf{r}) = \sum_{i=1}^N w_i \mathbf{s}_i, \quad (1)$$

where $\hat{\mathbf{C}}$, $\hat{\mathbf{N}}$, $\hat{\mathbf{S}}$ are the predicted 2D color, normal and semantic logits of ray \mathbf{r} , respectively; $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$ is the discrete opacity of i -th sampled point, $w_i = \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j)$ represent its transmittance, $\delta_i = \|x_{i+1} - x_i\|_2$ is the distance between adjacent sampled points.

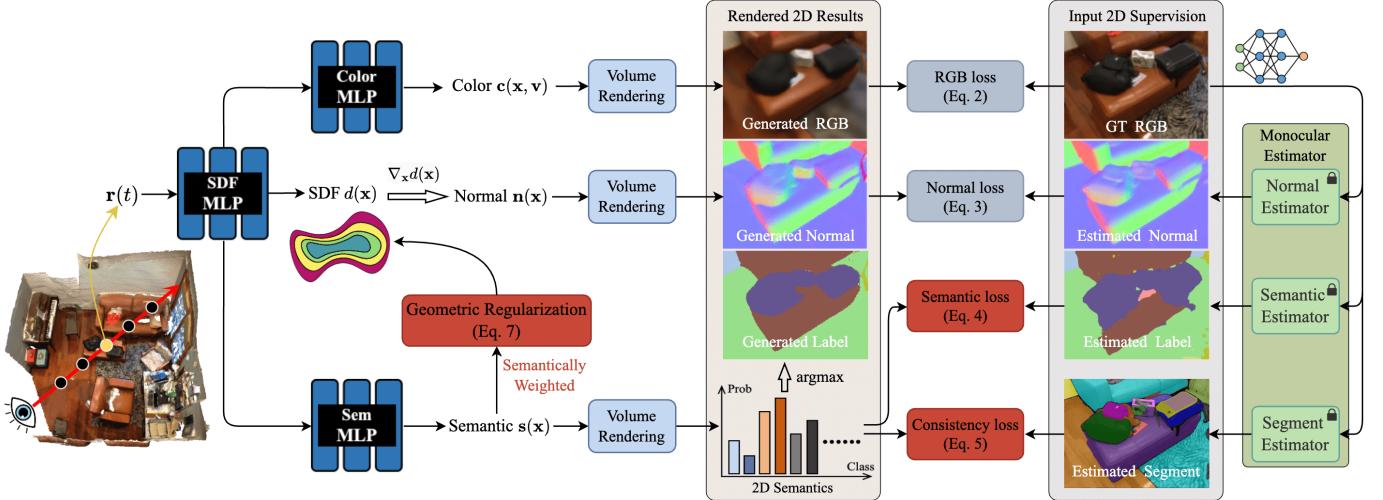


Fig. 2. **Overview of MOSE.** Utilizing RGB images and estimated normals, semantic labels, as well as segment masks, MOSE learns the color field, signed distance function (SDF) field and semantic field of the scene through an implicit neural representation. To address the discontinuity of 2D semantic predictions, we propose a locally-consistent fusion strategy (Sec. III-B) leveraging 2D segmentation techniques. Semantically-weighted geometric regularization (Sec. III-C) is further introduced to bring benefits to both the SDF field and semantic field.

We use photometric loss \mathcal{L}_c , normal loss \mathcal{L}_n and semantic loss \mathcal{L}_s to supervise the network:

$$\mathcal{L}_c = \sum_{\mathbf{r} \in \mathcal{R}} \|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_1, \quad (2)$$

$$\mathcal{L}_n = \sum_{\mathbf{r} \in \mathcal{R}} \|\hat{\mathbf{N}}(\mathbf{r}) - \mathbf{N}(\mathbf{r})\|_1 \cdot \Omega_n(\mathbf{r}), \quad (3)$$

$$\mathcal{L}_s = - \sum_{\mathbf{r} \in \mathcal{R}} \sum_{l=1}^L p^l(\mathbf{r}) \log \hat{p}^l(\mathbf{r}), \quad (4)$$

where \mathcal{R} are the sampled rays, \mathbf{C} and \mathbf{N} are the corresponding RGB color and normal prior, respectively. Following [9], we apply normal supervision only to pixels satisfying a pre-defined patch-wise similarity metric, indicated by $\Omega_n(\cdot)$. With a set of predefined L semantic classes, $p^l(\mathbf{r})$ is the rendered semantic probability at class l after softmax normalization, and $p^l(\mathbf{r})$ is the input 2D semantic supervision.

It is worth to note that we prevent gradients from semantics (Eq. 4) to SDF MLP. We empirically observe that the volume density (i.e., SDF value) often distorts to accommodate input noisy labels, leading to significant geometric degradation. Similar finding has also been found in [5]. As shown in Tab. III, adopting this strategy benefits both surface reconstruction and semantic understanding.

B. Locally-Consistent Fusion Strategy

Supervised by imperfect semantic predictions [20], [21], reasonable 2D labels can be directly rendered from a joint compact scene representation [4], [9]. However, we observe frequent discontinuity and inconsistency among semantic rendering, especially on objects shown as noisy patches and bleeding edges in Fig. 3. We suspect that the element-wise fusion lacks an explicit awareness of the local coherence of semantics. To address this issue, we propose to rely on generic images segments obeying image structures like boundaries. We introduce a locally-consistent fusion strategy (LCF), leveraging on 2D segmentation techniques (e.g. SAM [12], superpixel segmentation [13]), to perceive neighboring pixels and output locally consistent semantics.

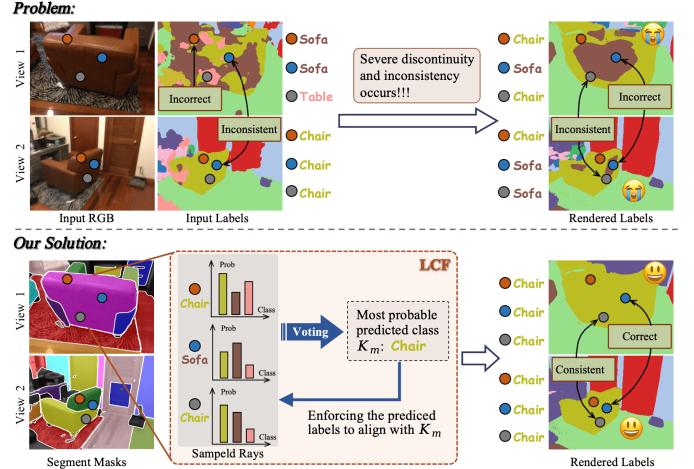


Fig. 3. **Overview of locally-consistent fusion strategy.** Severe discontinuity and inconsistency of semantics can be observed when directly inputting noisy multi-view labels into a NeRF-based fusion system (upper part). Our LCF strategy utilizes 2D segment priors to enforce consistent and accurate semantic distributions within each segment mask (bottom part).

Without loss of generality, we use SAM [12] to process each multi-view image and generate class-agnostic masks $\{\mathcal{M}_k\}$. We also ablate the choice of using super-pixels [13] in Sec. IV-D. During training, we encourage learned semantic distribution to be consistent within each local segment mask (Fig. 3). Specifically, sampled rays located within the same mask m are clustered into a group \mathcal{R}_m . For each group, we vote the most probably predicted class K_m by counting the number of occurrences of each class. We also tried other voting strategies like soft voting but find negligible difference. We hence stick to current simple yet efficient choice. During training, after obtaining the voted labels of \mathcal{R}_m , the semantic consistency loss is defined as follows:

$$\mathcal{L}_{con} = - \sum_{m=1}^M \sum_{\mathbf{r} \in \mathcal{R}_m} \log \hat{p}^{K_m}(\mathbf{r}), \quad (5)$$

where $\hat{p}^{K_m}(\mathbf{r})$ is the rendered semantic probability of the

voted class K_m . Although 2D masks $\{\mathcal{M}_k\}$ vary across frames, we incorporate LCF in a frame-wise manner and learn a more consistent and accurate semantic field.

C. Semantically-Weighted Geometric Regularization

Accurately reconstructing scene geometry from images alone is difficult even with normal prior, since degradation usually occurs in low-texture regions due to their inherent shape ambiguity [17]. As shown in Fig. 4, though the rendered normal maps have closely matched input normal priors, the reconstructed walls still exhibit obvious fragmentation.

Recall that previous methods [7], [8] usually use Eikonal loss [22] to regularize the SDF field:

$$\mathcal{L}_{eik} = \sum_{\mathbf{x} \in \mathcal{X}} (\|\nabla_{\mathbf{x}} d(\mathbf{x})\|_2 - 1)^2, \quad (6)$$

where \mathcal{X} denotes a set of sampled points \mathbf{x} and $d(\mathbf{x})$ is the associated SDF field. One straightforward solution to enhance the continuity of such texture-less planar regions is to increase the weighting of Eikonal loss. However, aggressively applying excessive strength to the Eikonal term worsens geometry by over-smoothing objects (Fig. 4). Thanks to the accurate semantics encouraged by LCF, we could adaptively adjust the strength of Eikonal loss based on learned semantic classes, and propose a semantically-weighted geometric regularization (SGR), defined as follows:

$$\mathcal{L}_{sgr} = \sum_{\mathbf{x} \in \mathcal{X}} (1 + \Phi(\hat{p}_{\mathbf{x}}))(\|\nabla_{\mathbf{x}} d(\mathbf{x})\|_2 - 1)^2, \quad (7)$$

where $\hat{p}_{\mathbf{x}}$ denotes the rendered multi-class semantic probability of sampled points \mathbf{x} , $\Phi(\hat{p}_{\mathbf{x}})$ is a semantically-adjusted weighting function, defined as:

$$\Phi(\hat{p}_{\mathbf{x}}) = \begin{cases} \sum_{l \in \mathcal{P}} \hat{p}^l(\mathbf{x}) & \text{if } \hat{l}_{\mathbf{x}} \in \mathcal{P} \\ 0 & \text{if } \hat{l}_{\mathbf{x}} \in \mathcal{O}, \end{cases} \quad (8)$$

where $\hat{l}_{\mathbf{x}}$ is the predicted semantic label of sampled points \mathbf{x} , we consider \mathcal{P} to be dominant indoor structures like walls, floors and ceilings, which are also planar regions, and \mathcal{O} to be other object classes. We dynamically adjust the strength of our geometric regularization across different semantic classes, ensuring a stronger SDF smoothness on planar regions while preserving objects' details. Though SGR does not explicitly optimize for semantics, we observe improvements in semantics owned to better geometry quality, as shown in Tab. III.

D. Training and Implementation Details

We train our model with the following loss function:

$$\mathcal{L} = w_c \mathcal{L}_c + w_n \mathcal{L}_n + w_s \mathcal{L}_s + w_{con} \mathcal{L}_{con} + w_{sgr} \mathcal{L}_{sgr}, \quad (9)$$

where we set the weighting factors as $w_c = w_n = 1$, $w_s = w_{con} = 0.5$, $w_{sgr} = 0.1$, respectively. We sample 512 rays for each batch and optimize using the Adam optimizer [23] with a learning rate of 2×10^{-4} . Our MLPs are trained on 1 NVIDIA V100 GPU for 160,000 iterations (~ 10 hours). Hyperparameters of MLPs are similar to [9] and [11] for a fair comparison. To obtain the 3D semantic map, we use the Marching Cubes algorithm [24] for extracting surface mesh and compute the semantic labels of each vertex on the mesh.

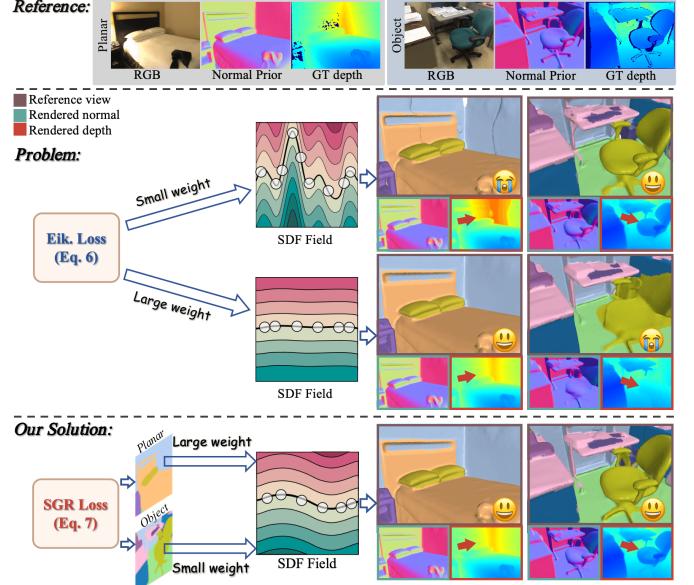


Fig. 4. Overview of semantically-weighted geometric regularization. Achieving a balance between planar and object regions is challenging with the widely-used Eikonal loss [22]: large loss weights lead to discontinuity in texture regions, while small loss weights result in loss of object details. Our proposed semantically-weighted geometric regularization (SGR) dynamically adjusts the regularization strength across different semantic classes, resulting in more accurate surface reconstruction. Semantics also benefit from the more accurate radiance field.

IV. EXPERIMENTS

A. Datasets, metrics and baselines

Datasets. We validate our approach on the popular ScanNet [25] dataset. ScanNet is a large-scale real-world indoor RGB-D video dataset consisting of 1613 indoor scenes with ground-truth camera poses, surface reconstructions and semantic annotations. Following previous works [9], [11], we select 8 representative scenes and use provided camera poses to perform experiments. As for training and evaluation of semantics, we adopt the widely used NYU-40 convention. Input images are resized to 640×480 pixels as well as other monocular priors, and we equally sampled 10% images of each scene to train our model [9].

Neural Priors. For normal priors, we use the SNU network [26] provided by NeuRIS [9]. For input semantic labels, we use DeepLabV3+ [20] with a ResNet-101 backbone [18] and re-train it on the ScanNet-frames-25k dataset (a subset of ScanNet) for 50 epoches. Both normal network and semantic networks are trained on the training split of ScanNet [25] and evaluated on 8 scenes from the validation set. For generic 2D segment priors, we use the standard SAM [12] and only keep segments larger than 4000 pixels to encourage consistency in a wider locality. In Sec. IV-D, we also validate our method using Mask2Former [21] and classical graph-based super-pixels [13] as semantics and segments predictions, respectively.

Metrics. Our method focuses on the task of 3D semantic reconstruction, 2D semantic segmentation and 3D geometry reconstruction from multi-view images. For 3D semantic segmentation, we transfer semantic labels from the reconstructed 3D mesh to the ground-truth mesh for the evaluation purpose. For 2D semantic segmentation, we render 2D semantic images

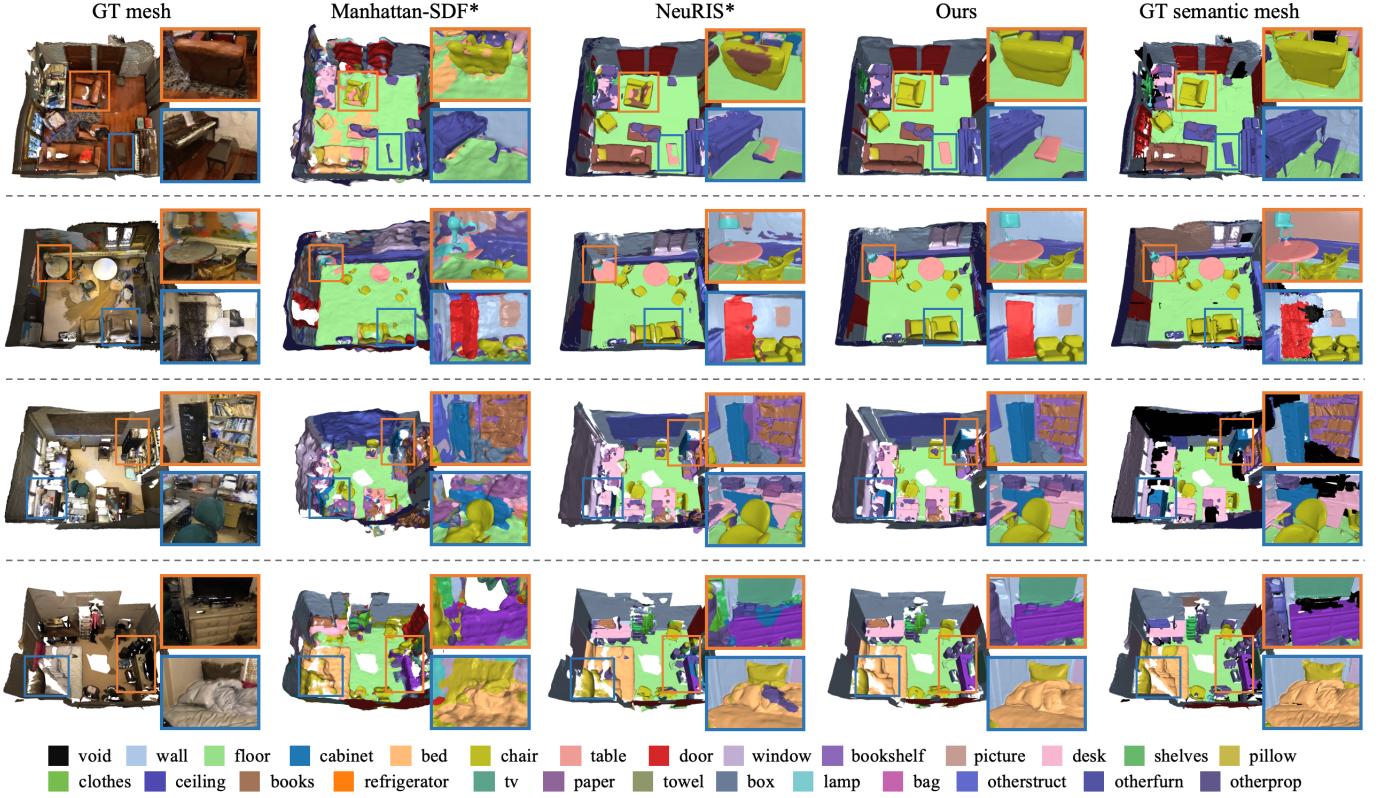


Fig. 5. Qualitative comparisons of 3D semantic reconstruction results. Our method is able to produce smoother 3D semantic map and align well with GT results, while Manhattan-SDF* and NeuRIS* exhibits severe inconsistencies of semantic labels.

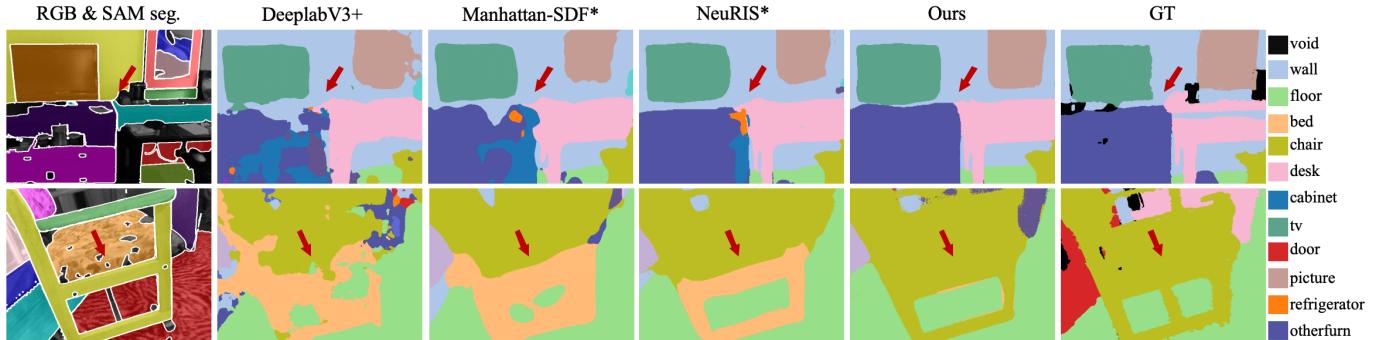


Fig. 6. Qualitative comparisons of 2D rendering semantic results. Inputting noisy labels from DeeplabV3+ [20], both Manhattan-SDF* and NeuRIS* struggle to render reasonable 2D semantic images. Utilizing SAM [12] segments, our method outperforms baselines with more consistent results.

at training viewpoints and compare them to ground-truth labels. We adopt total pixel accuracy (Acc), average pixel accuracy (mAcc), and mean class-wise intersection over union (mIoU) as the semantic evaluation metrics. For geometry reconstruction, we follow the evaluation procedure of [11], utilizing six standard metrics: accuracy (Acc), completeness (Comp), precision (Prec), recall, F-score and chamfer distance (CD).

Baseline. We compare our method to the following baselines: (1) **NeuRIS***: As NeuRIS [9] is only built for geometry reconstruction, we enhance NeuRIS with extra semantic attribute by integrating a semantic MLP. As discussed in Sec. III-A, we also prevent gradients from the semantic loss (Eq. 4) back to SDF branch, and denote this version as NeuRIS*. (2) **Manhattan-SDF** and **Manhattan-SDF***: Since Manhattan-SDF [11] mainly considers three coarse semantic classes:

floor, wall and others, we extend it to learn NYU-40 classes and denote it as Manhattan-SDF* while keeping all other components intact. We also attempted to stop the gradient of semantic loss (Eq. 4) like NeuRIS* and ours, but found no significant gains for Manhattan-SDF*. (3) **S-NeRF**: We re-trained density-based Semantic-NeRF [4] using the same images and predicted labels.

B. Evaluation and Comparison

3D Semantic Segmentation. Tab. I presents the overall quantitative results of 3D semantic segmentation. Compared with baselines, our approach demonstrates significant improvement across all metrics. It is noteworthy that Manhattan-SDF* with extended 40 semantic classes yields relatively poor performance (46.6% mIoU). This deficiency can be attributed to its joint optimization strategy, which only covers

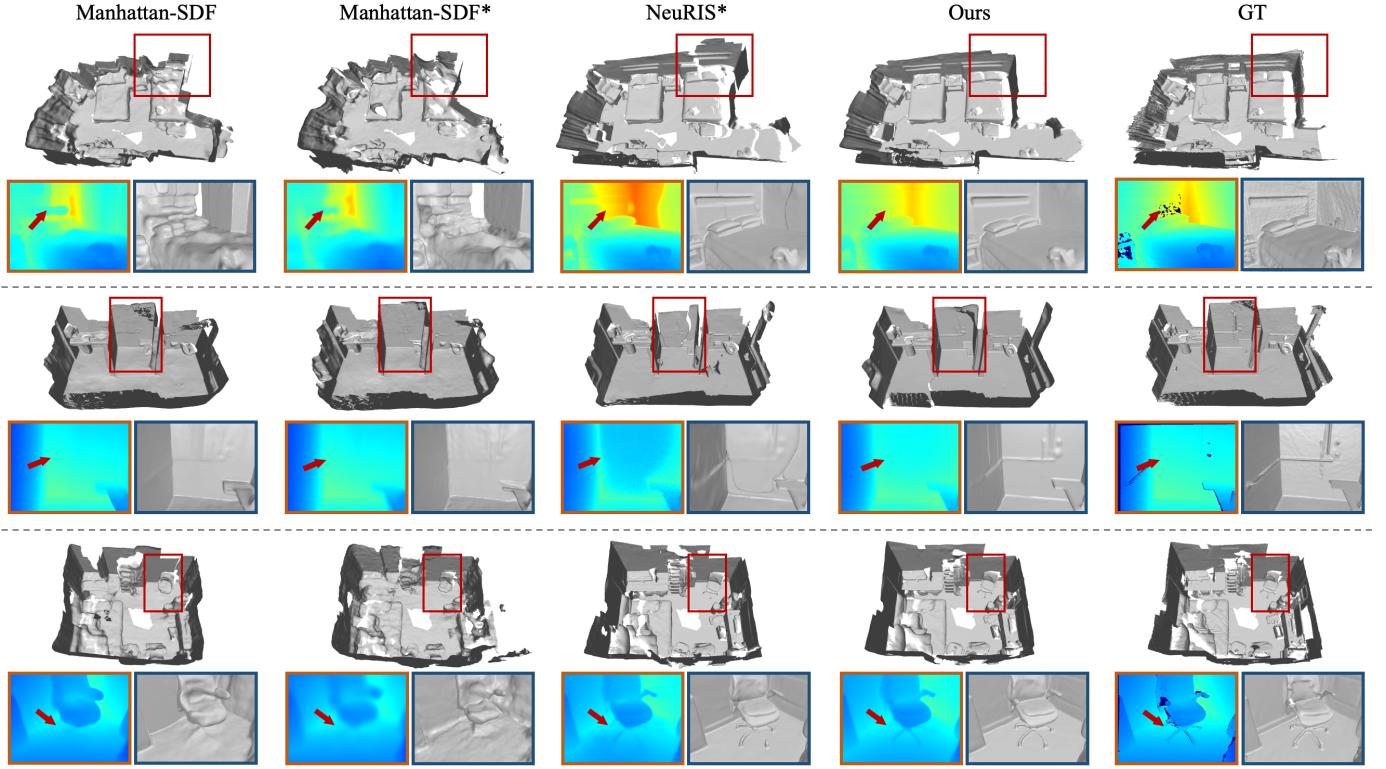


Fig. 7. Qualitative comparisons of surface reconstruction results. We compare our surface reconstruction results to vanilla Manhattan-SDF [11], Manhattan-SDF*, and NeuRIS* (same geometry to NeuRIS [9]). Both Manhattan-SDF and Manhattan-SDF* exhibit poor geometry and lose many object details (left two columns). Although estimated normals effectively aid surface reconstruction, NeuRIS* struggles on texture-less regions (third column). With the proposed SGR, our method produces superior results not only on texture-less regions (top two rows), but also on object regions (last row).

TABLE I
QUANTITATIVE RESULTS OF FUSED 3D/2D SEMANTIC SEGMENTATION

Method	3D Semantics			2D Semantics		
	Acc↑	mAcc↑	mIoU↑	Acc↑	mAcc↑	mIoU↑
DeeplabV3+	\	\	\	0.613	0.763	0.503
S-NeRF	0.471	0.644	0.361	0.635	0.783	0.532
Manhattan*	0.628	0.672	0.466	0.643	0.793	0.545
NeuRIS*	0.631	0.766	0.529	0.662	0.819	0.573
Ours	0.647	0.789	0.562	0.693	0.844	0.619

TABLE II
QUANTITATIVE RESULTS OF SURFACE RECONSTRUCTION

Method	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑	CD↓
S-NeRF	0.162	0.109	0.193	0.297	0.233	0.065
Manhattan	0.065	0.066	0.642	0.615	0.628	0.025
Manhattan*	0.122	0.077	0.457	0.471	0.464	0.099
NeuRIS*	0.051	0.041	0.746	0.756	0.751	0.014
Ours	0.045	0.037	0.773	0.779	0.776	0.011

two planar semantic regions (i.e., wall, floor) and thus lacks constraints for finer-grained objects. Compared to Manhattan-SDF*, NeuRIS* exhibits higher semantic fusion performance (+6.3% mIoU), benefiting from its more precise geometry through the utilization of normal priors. MOSE, built upon NeuRIS*, outputs a smoother 3D semantic map (further +3.3% mIoU) thanks to proposed LCF and SGR modules. Qualitative results of 3D semantic reconstruction are presented in Fig. 5. Inputting erroneous 2D labels with coarse neural priors, our MOSE reconstructs more accurate and consistent semantic maps, while Manhattan-SDF* and NeuRIS-S* struggle to

ensure labels continuity.

2D Semantic Segmentation. We also render learned semantics into image space and report 2D semantic segmentation metrics in Table I. Compared with the input DeeplabV3+ labels, both baseline methods and MOSE are capable of producing higher quality 2D semantic images, benefiting from the compact implicit scene representation. Among these methods, our MOSE achieves the best performance across all metrics, surpassing the second-highest NeuRIS* score by 4.6% mIoU and outperforming the input DeeplabV3+ labels by 11.6% mIoU. Fig. 6 presents qualitative results and ours are also visually much better than other methods by predicting coherent labels.

3D Surface Reconstruction. As demonstrated in Tab II, our method achieves the best reconstruction performance as well. Although Manhattan-SDF [11] applies Manhattan constraints to regularize planar regions, and utilizes sparse depth from point cloud, both Manhattan-SDF and Manhattan-SDF* exhibit lower geometric quality compared to NeuRIS* and ours with normal priors. These results demonstrate the unique effectiveness of normal priors in multi-view 3D reconstruction, which has been validated by [9], [16]. [11] lacks supervision for objects where Manhattan assumption is not satisfied (first row of Fig. 7). Please note that NeuRIS* shares the same geometric quality to NeuRIS, as the SDF MLP receives no gradients from semantics. MOSE further enhances the geometric quality (+2.5% F-score compared to NeuRIS*). As shown in Fig. 7, MOSE shows better visualization results not only on texture-less regions (top two rows) but also on more

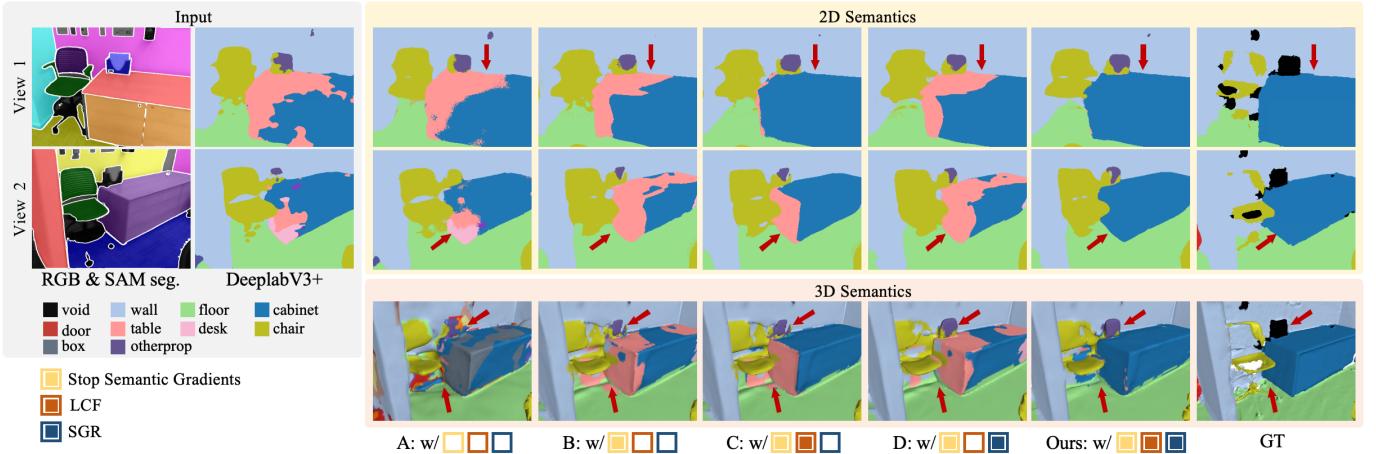


Fig. 8. **Fused 3D/2D semantic results of ablation study.** Comparing Model-A with Model-B, the volume density (i.e., SDF value) often distorts to accommodate input noisy labels, and preventing SDF MLP from the influence of semantic loss brings significant benefits to the semantic field. Comparing Model-B with Model-C, after adopting LCF module, our method is able to produce more accurate and consistent semantic labels. Comparing Model-C with ours MOSE, with the help of SGR module, smoother semantic field can be learned benefiting from more precise radiance field.

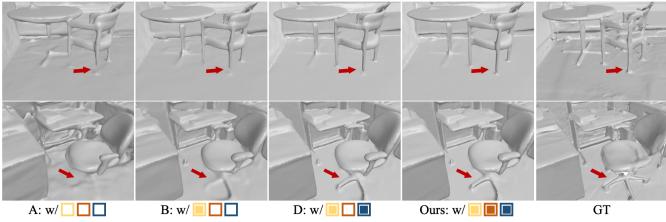


Fig. 9. **Surface reconstruction results of ablation study.** Comparing Model-A with Model-B, preventing semantic gradients to other branches leads to significant improvements in surface reconstruction. Compared with Model-B, ours preserve more object details with the help of SGR module. Model-C is not presented here as LCF module does not affect SDF field.

TABLE III
ABALATION STUDIES OF OUR DESIGN CHOICES

Model	Stop	LCF	SGR	mIoU(3D)↑	mIoU(2D)↑	F-Score↑	CD↓
A	✗	✗	✗	0.328	0.532	0.617	0.023
B	✓	✗	✗	0.529	0.573	0.751	0.014
C	✓	✓	✗	0.553	0.597	0.751	0.014
D	✓	✗	✓	0.529	0.580	0.775	0.011
Ours	✓	✓	✓	0.562	0.619	0.776	0.011

object details (last row).

C. Ablation Studies

In order to evaluate the effectiveness of each component in our method, we conduct ablation studies with different settings and variants: **Model-A**: directly integrate a semantic MLP into NeuRIS; **Model-B**: prevent SDF MLP of Model-A from the influence of semantics (i.e., NeuRIS*). Thus Model-B has the same geometry quality to NeuRIS [9]; **Model-C**: built upon Model-B, with the LCF module. Note that we use Eikonal loss (Eq. 6) to regularize the SDF field in Model-A, Model-B and Model-C; **Model-D**: built upon Model-B, with the SGR module; **Ours**: integrate both LCF and SGR modules. We report quantitative results in Tab III and present qualitative results in Fig. 8 and Fig. 9.

Stopping Semantic Gradient. Comparing Model-A against Model-B, we observe that preventing semantic gradients to other branches leads to improvements in both geometry and semantics evaluation. We attribute this to the negative impact

of altering geometry to over-fit erroneous 2D labels, hence adhering to the radiance field of NeuRIS is more accurate.

Locally-Consistency Fusion Strategy. As LCF module concentrates on semantics, here we focus on the tasks of 3D and 2D semantic segmentation. Comparing Model-C with Model-B, the LCF module brings significant improvement in both 3D and 2D semantics. Fig. 8 also shows that inconsistent input 2D semantics under different views lead to a blending of two labels for the same object, while our proposed LCF is able to mitigate this effect and produce accurate labels.

Semantically-Weighted Geometric Regularization. Comparing Model-C with ours in Tab. III, our designed SGR module brings benefits to both geometry and semantics. We observe that Model-D achieves almost the same geometry evaluation metrics as ours. We attribute this to the fact that SGR focuses on planar and object regions, and the fused semantics are already sufficiently accurate even without LCF module. Fig. 7 (Note that Model-B is equivalent to NeuRIS*) and Fig. 9 shows that our SGR module not only effectively improves the reconstruction quality of planar regions, but also preserves more object details. In addition, benefiting from more accurate radiance field, MOSE is able to produce smoother semantics, especially when combined with LCF module (second-to-last columns of Fig. 8).

D. Further Discussion

We also discuss the sensitivity of our system to different input labels and segmentation priors using labels and segments from Mask2Former [21] and unsupervised super-pixel techniques [13]. We use the COCO pre-trained Mask2Former [5], [27] and remap its output semantic labels to NYU-40 classes. Therefore, its predictions under NYU-40 convention are relatively lower than fine-tuned DeepLabV3+. For 2D segment masks, we take graph-based super-pixels and remove those smaller than 4000 pixels. We focus on evaluating their impact on semantics as the geometry quality is almost comparable. Qualitative results in Fig. 10 show that MOSE still exhibits commendable semantic fusion capability, indicating its robustness to input label qualities. Using super-pixels

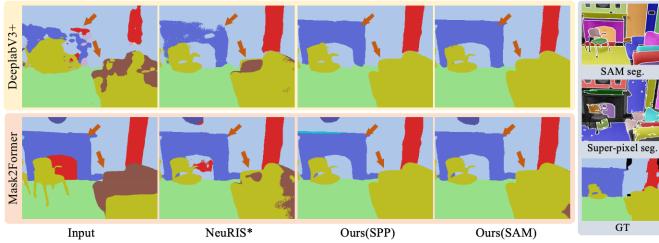


Fig. 10. **Qualitative results of different labels and segments.** We validate the effectiveness of MOSE by inputting Mask2Former [21] labels and superpixel segments [13]. Qualitative results demonstrate that our method is not sensitive to prior predictors.

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT LABELS AND SEGMENTS

Method	DeepLabV3+			Mask2Former		
	mIoU(3D)↑	mIoU(2D)↑	F-score↑	mIoU(3D)↑	mIoU(2D)↑	F-score↑
Input	\	0.503	\	\	0.440	\
NeuRIS*	0.529	0.573	0.751	0.427	0.471	0.751
Ours(SPP)	0.539	0.596	0.777	0.434	0.478	0.777
Ours(SAM)	0.562	0.619	0.776	0.450	0.489	0.772

results in slightly inferior scores compared to those obtained using SAM, mainly due to less accurate segment masks. Tab. IV confirms our observations and shows that our approach is not sensitive to the specific choices of prior predictors and maintains consistent performance across various types of prior predictors.

V. LIMITATIONS AND CONCLUSION

In this work, we proposed MOSE, a NeRF-based 3D scene understanding approach only using multi-view images and 2D priors. To address the particularly limited performance of existing works on object-level semantics as well as geometry of texture-less regions, we introduce segment-guided consistency and semantic-guided smoothness to improve these capability. Both quantitative and qualitative results on the real-world ScanNet dataset have validated the effectiveness of MOSE, achieving promising results in both semantics and geometry.

Limitations. Although MOSE demonstrates robustness to noisy priors, its performance may degrade when neural priors are incorrectly predicted across most viewpoints. Similar to other NeRF-based methods, MOSE still requires provided camera poses and relatively lengthy per-scene optimization, while in practice it would be more desirable if multi-view information could be efficiently accumulated to form a 3D model. Therefore, choosing the recently popular 3D Gaussian Splatting [28] as our novel backbone or conducting joint optimization of MLPs and camera states could serve as our valuable future work.

REFERENCES

- J. McCormac, A. Handa, A. J. Davison, and S. Leutenegger, “SemanticFusion: Dense 3D semantic mapping with convolutional neural networks,” in *IEEE International Conference on Robotics and Automation*, 2017.
- G. Narita, T. Seno, T. Ishikawa, and Y. Kaji, “Panopticfusion: Online volumetric semantic mapping at the level of stuff and things,” in *IEEE/RSJ Conference on Intelligent Robots and Systems*, 2019.
- B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis,” in *European Conference on Computer Vision*, 2020.
- S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, “In-Place Scene Labelling and Understanding with Implicit Scene Representation,” in *International Conference on Computer Vision*, 2021.
- Y. Siddiqui, L. Porzi, S. R. Bulò, N. Müller, M. Nießner, A. Dai, and P. Kontschieder, “Panoptic lifting for 3d scene understanding with neural fields,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- Y. S. Bhagat, I. Laina, J. F. Henriques, A. Vedaldi, and A. Zisserman, “Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion,” in *Neural Information Processing Systems*, 2023.
- L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, “Volume rendering of neural implicit surfaces,” in *Neural Information Processing Systems*, 2021.
- P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, “Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction,” in *Neural Information Processing Systems*, 2021.
- J. Wang, P. Wang, X. Long, C. Theobalt, T. Komura, L. Liu, and W. Wang, “Neuris: Neural reconstruction of indoor scenes using normal priors,” in *European Conference on Computer Vision*, 2022.
- Z. Li, X. Lyu, Y. Ding, M. Wang, Y. Liao, and Y. Liu, “Rico: Regularizing the unobservable for indoor compositional reconstruction,” in *International Conference on Computer Vision*, 2023.
- H. Guo, S. Peng, H. Lin, Q. Wang, G. Zhang, H. Bao, and X. Zhou, “Neural 3d scene reconstruction with the manhattan-world assumption,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.
- A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *International Conference on Computer Vision*, 2023.
- P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 59, pp. 167–181, 2004.
- L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, “Occupancy networks: Learning 3d reconstruction in function space,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, “Deepsdf: Learning continuous signed distance functions for shape representation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- M. Park, M. Do, Y. Shin, J. Yoo, J. Hong, J. Kim, and C. Lee, “H2O-SDF: Two-phase learning for 3d indoor reconstruction using object surface fields,” in *International Conference on Learning Representations*, 2024.
- K. Zhang, G. Riegler, N. Snavely, and V. Koltun, “Nerf++: Analyzing and improving neural radiance fields,” *arXiv preprint arXiv:2010.07492*, 2020.
- K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- S. Zhi, M. Bloesch, S. Leutenegger, and A. J. Davison, “SceneCode: Monocular dense semantic reconstruction using learned encoded scene representations,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *European Conference on Computer Vision*, 2018.
- B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.
- A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, “Implicit geometric regularization for learning shapes,” in *Proceedings of the International Conference on Machine Learning*, 2020.
- D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, 2015.
- W. E. Lorensen and H. E. Cline, “Marching Cubes: A high resolution 3D surface construction algorithm,” in *SIGGRAPH*, 1987.
- A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- G. Bae, I. Budvytis, and R. Cipolla, “Estimating and exploiting the aleatoric uncertainty in surface normal estimation,” in *International Conference on Computer Vision*, 2021.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision*, 2014.
- B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” in *SIGGRAPH*, 2023.