Extending a Large View Synthesis Model for Multi-view Panoptic Segmentation

Kwonyoung Ryu[1], In-Jae Lee[4], Jonghyun Jin[2], Hyunjee Lee[2], Jongmin Lee[3], Jaesik Park[4]

[1]POSTECH [2]POSCO DX [3]Chung-Ang University [4]Seoul National University

Abstract. Large view synthesis models synthesize novel views through cross-view attention without explicit 3D representations, and recent studies have shown that they learn accurate spatial correspondence from RGB supervision alone. We observe that this correspondence generalizes beyond appearance. When non-photorealistic signals such as binary encoded panoptic labels are passed through the model, they are propagated to novel views with consistent spatial structure. These results indicate that the correspondence learned for RGB view synthesis can also propagate view-independent per-pixel labels. From this observation, we study how large view synthesis models can be extended beyond appearance rendering to 3D scene understanding. We propose a panoptic segmentation pipeline that reuses a frozen view synthesis model to propagate panoptic labels from input views to novel views, without 3D reconstruction or any segmentation-specific training of the view synthesis model. Given panoptic labels on the input views, we encode them into binary channel representations and pass them through the same model to render target-view segmentation. On ScanNet, our method achieves segmentation quality on par with Gaussian based approaches requiring explicit 3D reconstruction, while outperforming them in novel view synthesis by more than 7 dB. The label propagation also transfers across datasets, surpassing these approaches on Replica without any fine-tuning.

Keywords: Novel View Synthesis, Large View Synthesis Models, 3D Panoptic Segmentation, Label Propagation, Multi-view 3D Understanding

1 Introduction

Predicting panoptic segmentation from novel viewpoints requires assigning semantic labels and instance identities to every pixel in an unobserved view, given only a few images of the scene. Embodied agents need this capability to reason about objects beyond the currently observed view. Example applications include anticipating scene content before navigation and augmenting training data from a small number of annotated views.

Most existing approaches first reconstruct a 3D representation such as NeRF [30] or 3D Gaussian Splatting [22], and then lift 2D panoptic labels into the recovered

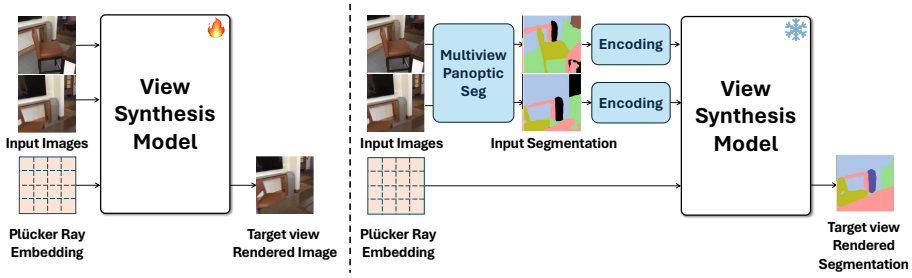

Fig. 1: Illustration of a large-view synthesis model and its extension to multi-view panoptic segmentation. In the **left** figure, a feed-forward large view synthesis model renders a novel target image from the input source views. In the **right** figure, the same model is bootstrapped for panoptic segmentation. We propose a simple signal-level extension that repurposes the learned cross-view correspondence of a view synthesis model for 3D scene understanding beyond RGB rendering.

representation through per-scene optimization [4, 40]. This design couples reconstruction quality with segmentation quality. Errors in the recovered geometry propagate directly to the lifted labels. Recent feed-forward variants predict 3D Gaussians with semantic features in a single forward pass [13, 51], eliminating per-scene optimization but still coupling reconstruction quality with segmentation in a shared 3D representation.

These limitations motivate an alternative that avoids explicit 3D reconstruction entirely. Meanwhile, recent NVS methods replace explicit 3D representations with latent token-based scene modeling [19, 21, 36, 38, 44]. These large view synthesis models attend across source and target tokens to synthesize novel views directly, without any 3D inductive bias such as NeRF [30] or 3DGS [22] in the network architecture [21]. Recent studies [19, 44] show that these models learn cross-view spatial correspondence even without explicit 3D supervision, and that their attention layers learn accurate cross-view spatial correspondence between source and target views. However, all existing analyses focus exclusively on target RGB view appearance. Whether the learned implicit correspondence extends to signals beyond appearance and can thus serve tasks beyond view synthesis remains unexplored.

We investigate this question by analyzing the implicit correspondence of a large view synthesis model [44] fine-tuned only for target RGB reconstruction via gradient-based saliency analysis [41]. We find that the model attends to geometrically corresponding regions in source views regardless of whether the input is RGB format or non-photorealistic per-pixel label encodings. The model thus learns geometric relationships between views rather than appearance-specific features, allowing view-consistent per-pixel signals such as panoptic labels to transfer to novel views through the same large view synthesis model.

We introduce a panoptic segmentation pipeline that bootstraps a large view synthesis model [44] trained solely for RGB image reconstruction. We first seg-

ment the input views using a shared query decoder to obtain cross-view-consistent panoptic labels, because the propagation stage preserves the labels provided as input, and instance identities must be consistent before encoding. The target RGB rendering path reconstructs the novel target-view image from source views, and the segmentation path encodes the panoptic labels as binary channel encodings and passes them through the same novel view synthesis model to produce target-view segmentation. Our method achieves 33.56 PSNR and 0.5949 mIoU on novel views in the ScanNet dataset [12], higher than the baseline SIU3R [51] (25.88 PSNR, 0.5894 mIoU), which trains explicit 3D Gaussians with dedicated segmentation objectives. Because the rendering transformer remains frozen, our pipeline preserves full rendering quality, exceeding Gaussian-based methods by more than 7 dB.

Our contributions are as follows.

- To our knowledge, this is the first work to extend large view synthesis models beyond appearance rendering to 3D scene understanding, demonstrating that explicit 3D reconstruction is not a prerequisite for multi-view panoptic segmentation.
- We demonstrate, through gradient-based saliency analysis, that a large view synthesis model trained only for RGB reconstruction propagates panoptic labels to novel views. The view synthesis model receives no segmentation-specific supervision.
- We present a modular pipeline that decouples novel view rendering and segmentation, allowing independent replacement of either component as the respective fields advance.

2 Related Work

Scene Understanding and Reconstruction. Panoptic segmentation in 3D assigns semantic labels and instance identities to every element of a scene. Some methods directly segment pre-scanned point clouds [26, 33, 39], but they assume the 3D geometry is already available. When only images are given, the dominant paradigm lifts 2D predictions into a 3D representation through per-scene optimization. NeRF-based approaches [3, 14, 27, 40, 57] fuse noisy 2D labels into a consistent volumetric model, while Gaussian Splatting variants [22] embed semantic or instance features directly into primitives [34, 48, 50, 53, 58–60]. A related direction distills open-vocabulary features from 2D foundation models into radiance fields [4, 20, 23]. All of these approaches require posed images, dense captures, and minutes to hours of optimization per scene. Recent feed-forward methods remove per-scene optimization by predicting labeled 3D representations in a single forward pass without pose information [13, 51, 61]. However, they still depend on explicit 3D intermediates such as pixel-aligned Gaussians [13, 51] or dense point maps [61]. As a result, reconstruction quality directly constrains segmentation quality, and the two components cannot be upgraded independently. Our approach passes encoded panoptic labels through a large view synthesis

model that has never been trained on segmentation. This requires no 3D reconstruction, no task-specific training, and no architectural modification.

Large View Synthesis Models. Novel view synthesis has traditionally relied on explicit 3D representations optimized per scene, such as volumetric fields [1, 2, 30, 31] or Gaussian primitives [17, 22, 55]. Feed-forward variants predict these representations from sparse inputs in a single pass, with some requiring posed images [6, 8, 16, 46, 54, 56] and others jointly recovering geometry and cameras from unposed images [28, 42, 45, 47, 52]. Despite their diversity, all assume that an explicit geometric intermediate must be recovered before rendering.

Another line of work synthesizes views directly through learned token-to-token mappings without explicit 3D structure [19, 21, 36–38, 44]. Early models in this line require ground-truth camera poses [21, 36], but subsequent work progressively removes this dependency through self-supervised pose prediction [19] and by eliminating pose representations entirely [44]. These models use cross-view attention to model correspondence between source and target views, and recent analyses [19, 44] show that this correspondence reflects 3D scene structure. Yet they have been explored exclusively for appearance synthesis. Whether the implicit correspondence extends to other per-pixel signals has not been investigated. We build upon Less3Depend [44] and show through gradient attribution analysis that its learned implicit correspondence is input-agnostic, propagating panoptic labels to novel views without any task-specific adaptation.

2D Panoptic Segmentation. Our pipeline uses off-the-shelf 2D panoptic segmentation as a modular source-view component. The task was introduced as a unification of semantic and instance segmentation [24], with early solutions building on detection [15] or encoder-decoder [7] architectures. The dominant modern formulation treats panoptic segmentation as mask classification via transformer decoders [10, 11, 18], and has been further extended toward open-vocabulary capabilities [25] and unified multi-task frameworks [18]. In parallel, self-supervised vision transformers [5, 32] serve as dense-prediction backbones, bridged to segmentation heads via adapter modules [9]. We combine these components to extract panoptic labels from each input view and propagate them through the large view synthesis model. Since no part of our pipeline is trained on 3D data, the 2D segmentation model can be replaced with any improved future method without modifying the rest of the system.

3 Method

In this section, we briefly review large view synthesis models (Sec. 3.1) and their inherent implicit correspondence capability (Sec. 3.2). We then present our approach to bootstrapping the view synthesis model for multi-view panoptic segmentation (Sec. 3.3), introduce a panoptic encoding and decoding scheme (Sec. 3.4), and discuss training (Sec. 3.5) and design properties (Sec. 3.6).

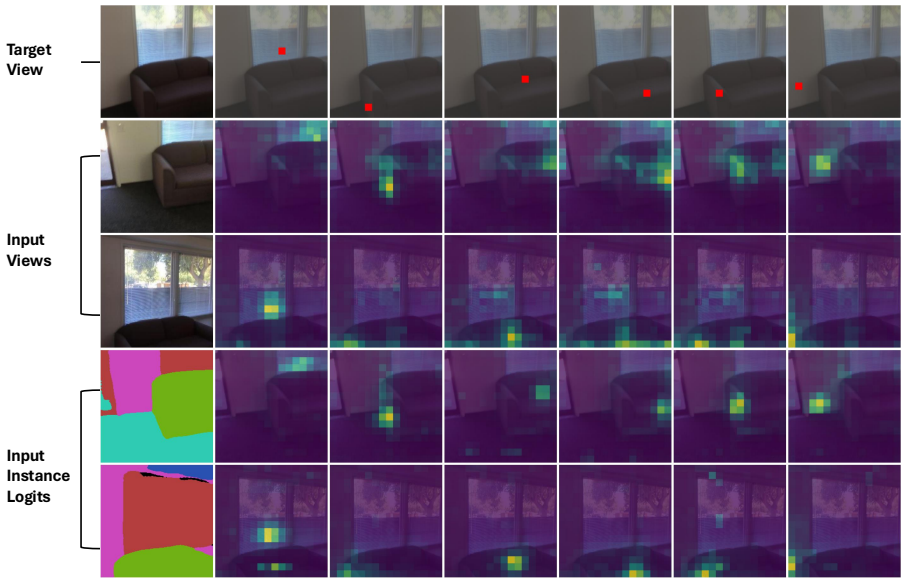

Fig. 2: Visualization of gradient saliency from target to source views. The saliency maps highlight similar source regions whether the model receives RGB images or binary instance encodings. This indicates that the learned cross-view correspondence of the model is not limited to appearance. Row 1 shows the rendered target view, with each column indicating the location of a different query patch in red. For each query patch, Rows 2-3 and Rows 4-5 show the gradient saliency over the same two source views, using RGB inputs and binary instance encodings, respectively.

3.1 Preliminaries : Large View Synthesis Models

Novel view synthesis aims to render images from unobserved viewpoints given a set of source images. Given N source views $\{I_i^{\mathrm{s}}\}_{i=1}^{N}$ with $I_i^{\mathrm{s}} \in \mathbb{R}^{H\times W\times 3}$, the goal is to produce a target image $\hat{I}^{\mathrm{t}} \in \mathbb{R}^{H\times W\times 3}$ at a novel viewpoint. Recent large view synthesis models such as Less3Depend [44] and RayZer [19] achieve strong performance even in the unposed sparse-view setting, where neither source nor target camera poses are provided. These models take N source images as input and directly synthesize the target view by inferring relative geometry internally, without constructing any explicit 3D representation. In this paper, we adopt Less3Depend [44], which builds upon LVSM [21], a transformer-based renderer that tokenizes source views via a frozen DINOv2 [32] encoder and synthesizes the target view through joint attention across all source and target tokens. Unlike LVSM, Less3Depend [44] does not require camera poses, enabling novel view synthesis from unposed sparse views. This formulation exposes a learned source-to-target correspondence through attention, which we later reuse to propagate view-consistent panoptic labels without modifying the rendering model.

3.2 Implicit Correspondence in a Large View Synthesis Model

In this section, we analyze whether the geometric correspondence learned by feed-forward large view synthesis models generalizes beyond RGB inputs, and show that this property enables panoptic label propagation to novel views. To examine whether this correspondence is specific to RGB inputs, we visualize gradient saliency [41] of the large view synthesis model [44]. For a single target patch, we backpropagate through the decoder and encoder and record the absolute gradient magnitude at each source token. This directly measures how much each source region causally influences the target output. For both input types, we provide the target-view latent Plücker estimated from RGB inputs.

As shown in Fig. 2, the gradient saliency reveals spatial correspondence between the target and source views. For each query patch on the target view, the saliency concentrates on geometrically corresponding regions in the source views (Rows 2,3). To test whether this correspondence depends on the input modality, we replace the source RGB inputs with binary instance encodings while keeping the same pose conditioning from the RGB rendering path. The correspondence persists under this change (Rows 4,5). The consistent saliency patterns across both modalities indicate that the model resolves spatial correspondence from geometric pose rather than input content, which enables panoptic label propagation through the rendering path.

3.3 Bootstrapping NVS for Panoptic Understanding

Overall Pipeline. As illustrated in Fig. 3(a), our method operates in two parallel paths through the same large view synthesis model. Given N unposed input views $\{I_i^{\mathrm{s}}\}_{i=1}^{N}$ with $I_i^{\mathrm{s}} \in \mathbb{R}^{H\times W\times 3}$ and a target image I^{t}, the model first encodes the input views into a scene latent $\mathbf{z}$ and estimates the target camera embedding:

$$\mathbf{z} = \mathcal{E}(I_1^{\mathrm{s}}, \ldots, I_N^{\mathrm{s}}, \mathbf{z}_0), \quad \hat{\mathbf{e}}^{\mathrm{t}} = \mathcal{P}(I^{\mathrm{t}},\ \mathbf{z}), \tag{1}$$

where $\mathbf{z}_0$ is a learnable latent refined by attending to source-view tokens, $\mathcal{E}$ is the scene encoder, $\mathcal{R}$ is the render decoder, and $\hat{\mathbf{e}}^{\mathrm{t}}$ is a latent Plücker embedding estimated by the pose estimator $\mathcal{P}$ from the unposed target image I^{t} and the scene latent $\mathbf{z}$ following [44]. The rendering path synthesizes the target-view image through the decoding stage as follows,

$$\hat{I}^{\mathrm{t}} = \mathcal{R}(\mathbf{z},\ \hat{\mathbf{e}}^{\mathrm{t}}). \tag{2}$$

For the segmentation path, a shared query decoder $\mathcal{D}$ first extracts cross-view consistent panoptic labels from the input views. As shown in Fig. 3(b), the labels are then encoded into binary channel representations $\mathbf{b}_i \in \{0,1\}^{H\times W\times 3}$. These encodings are then passed through the same encoder and decoder as follows.

$$\mathbf{z}^{\mathrm{seg}} = \mathcal{E}(\mathbf{b}_1^{\mathrm{s}}, \ldots, \mathbf{b}_N^{\mathrm{s}}, \mathbf{z}_0), \quad \hat{\mathbf{b}}^{\mathrm{t}} = \mathcal{R}(\mathbf{z}^{\mathrm{seg}},\ \hat{\mathbf{e}}^{\mathrm{t}}), \tag{3}$$

where $\hat{\mathbf{e}}^{\mathrm{t}}$ and $\mathbf{z}_0$ are reused from the rendering path and all weights of $\mathcal{E}$ and $\mathcal{R}$ are trained with RGB reconstruction only. The continuous output $\hat{\mathbf{b}}^{\mathrm{t}} \in [0,1]^{H\times W\times 3}$

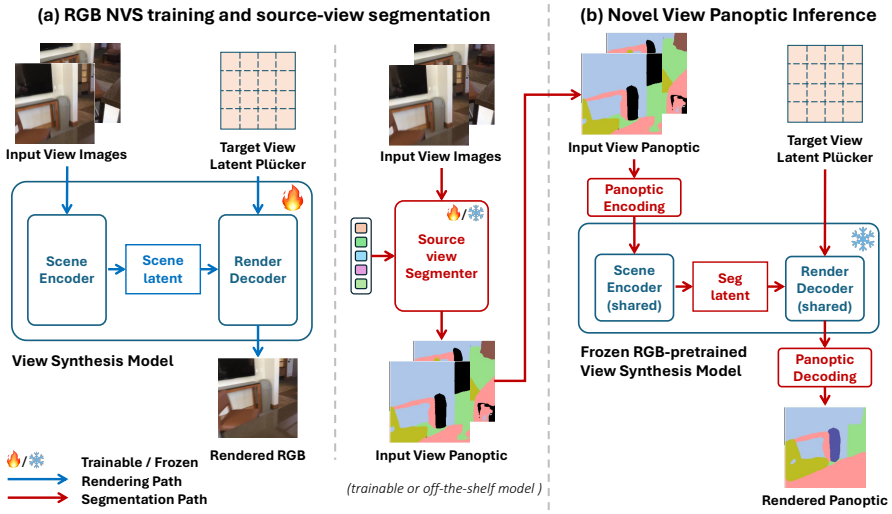

Fig. 3: Overall pipeline. (a) RGB NVS training and source-view segmentation: the view synthesis model (blue) is trained to encode input-view images into a scene latent and render the target-view RGB image from the target-view latent Plücker embedding. A source-view segmenter (red), implemented either as our trainable shared query decoder or an off-the-shelf model, predicts panoptic labels for the input views. (b) Novel-view panoptic inference: the input-view panoptic labels produced in (a) are encoded and passed through the frozen RGB-pretrained view synthesis model to propagate them to the target view. The resulting target-view binary output is converted into the final panoptic map by panoptic decoding. The view synthesis model is trained only for novel-view RGB synthesis, not for semantic or panoptic segmentation.

is then decoded into the final panoptic map P^{t}. As shown in Sec. 3.2, the correspondence learned by the view synthesis model remains stable when RGB inputs are replaced with binary instance encodings. Since the model is trained solely for RGB image reconstruction, panoptic propagation requires only lightweight processing around the existing rendering path.

Multi-view Panoptic Segmentation. Since our pipeline propagates input-view panoptic labels through the large view synthesis model, instance identities must be consistent across all input views. Conflicting labels for the same object would lead to ambiguous propagation. To this end, we adopt a shared query decoder $\mathcal{D}$ following Mask2Former [10], built on a frozen DINOv2 [32] encoder with ViT-Adapter [9]. A shared set of learnable object queries $\mathbf{Q}$ is fed into the decoder along with features from all input views simultaneously as follows.

$$\{(\mathbf{m}_j,\, c_j)\}_{j=1}^{Q} = \mathcal{D}(\mathbf{Q},\, I_1, \ldots, I_N), \tag{4}$$

where $\mathbf{Q} = \{\mathbf{q}_j\}_{j=1}^{Q}$ is a set of Q learnable object queries, $\mathbf{m}_j \in [0,1]^{N\times H\times W}$ is the predicted mask, and $c_j \in \mathcal{C}$ is the predicted semantic class for the j-th query. Since each query attends to all views jointly, it explicitly represents

a potential object instance across views, ensuring consistent instance identities without requiring explicit cross-view matching.

3.4 Panoptic Encoding and Decoding

The large view synthesis model transfers implicit correspondence through three output channels. Given a panoptic map with semantic labels $S(p)$ and instance IDs $I(p) \in \{0, \ldots, K-1\}$ for each pixel p, we encode each pixel as $\mathbf{b}(p) = \mathcal{B}(I(p)) \in \{0,1\}^3$, where $\mathcal{B}$ denotes the 3-bit binary encoding ($K \leq 8$). We also store a look-up table $\mathcal{L}$ mapping each instance index to its semantic class. At inference, the model outputs $\hat{\mathbf{b}}^{\mathrm{t}} \in [0,1]^{H \times W \times 3}$ for the target view, and we recover the panoptic map as,

$$\hat{I}^{\mathrm{t}}(p) = \mathcal{B}^{-1}\big(\mathrm{round}(\hat{\mathbf{b}}^{\mathrm{t}}(p))\big), \quad S^{\mathrm{t}}(p) = \mathcal{L}\big(\hat{I}^{\mathrm{t}}(p)\big). \tag{5}$$

Pixels where any channel k satisfies $|\hat{b}_{:,:,k}(p) - 0.5| \leq \tau$ are marked as uncertain and left unlabeled. We set $\tau = 0.2$ for the best performance.

We choose binary over continuous encoding for robustness to rendering artifacts. When the model blends two binary codewords at instance boundaries, the interpolated values converge toward 0.5, which is maximally distant from both valid states and reliably rejected by thresholding. A finer encoding would narrow the inter-codeword gaps, causing boundary interpolation to alias into valid but incorrect codewords and introduce phantom instances that cannot be filtered out. A single 3-bit propagation pass represents at most eight active instance IDs. We handle scenes with more instances through a multi-pass decoding scheme, described in the supplementary (Appendix B).

3.5 Training Scheme

The default variant fine-tunes the NVS model ($\mathcal{E}$, $\mathcal{P}$, $\mathcal{R}$) with the photometric loss of Less3Depend [44], using RGB reconstruction only. We train the shared query decoder $\mathcal{D}$ with the standard Mask2Former [10] objective on the input-view labels. The NVS model receives no segmentation-specific loss, so panoptic propagation relies entirely on the correspondence learned from photometric supervision. The segmentation path runs only a forward pass through $\mathcal{E}$ and $\mathcal{R}$, and no gradient flows into the rendering weights from the binary encodings. This keeps the rendering transformer identical to the frozen view synthesis model. The modular variants take source-view labels from an off-the-shelf segmenter such as SAM2 [35] or PanSt3R [61], and require no training within our pipeline. The propagation also holds when we replace Less3Depend [44] with RayZer [19] as the NVS backbone, showing that it applies across different large view synthesis models. We report the segmentation loss terms and weights, together with the RayZer results, in the supplementary (Appendices A and D).

3.6 Discussion

A key reason the label propagation works is that panoptic labels are view-independent. The semantic class and instance identity of a surface point do not change with the viewpoint of the observer. The model only needs to establish accurate spatial correspondence between views, and the labels transfer directly. This design separates source-view segmentation from target-view propagation. The large view synthesis model and the source-view segmenter can be replaced without changing the propagation stage, without retraining the rest of the pipeline. Replacing the source-view segmenter with PanSt3R [61] transfers to Replica [43] zero-shot and outperforms a reconstruction-based baseline on every metric, without any fine-tuning of the rest of the pipeline (Sec. 4.5).

4 Experiments

In this section, we evaluate rendering quality, novel-view segmentation accuracy, low-overlap behavior, and cross-dataset transfer. We evaluate whether panoptic propagation through a frozen NVS model preserves rendering quality while reaching competitive segmentation accuracy, under standard overlap (Sec. 4.2) and sparse overlap (Sec. 4.3). We compare design alternatives that couple segmentation with the NVS model (Sec. 4.4). We then assess the decoupled design through modular composition with an off-the-shelf segmenter and zero-shot transfer to a new dataset (Sec. 4.5).

4.1 Experimental Setup

Implementation Details. We use ScanNet [12] for training and evaluation. Each training sample consists of 2 input views and 2 target novel views, all resized to 224×224. Following SIU3R [51], view pairs are sampled with depth-based IoU in $[0.3, 0.8]$ for both training and evaluation. For evaluation, we use 2 input views and 4 target views on 1,860 selected view pairs in the same IoU range from the validation set. We train for 100 epochs on 8 NVIDIA RTX A6000 GPUs with a total batch size of 64, taking approximately 3 hours.

Baselines. We organize baselines into two groups to evaluate different aspects of our method. Per-view segmentation methods (Mask2Former [10], LSeg [29]) establish upper bounds on input-view segmentation, since they are applied directly to ground-truth images without any rendering. Joint NVS and scene understanding methods (LSM [13], SIU3R [51]) are the most direct comparisons, as they also target novel view panoptic segmentation from sparse unposed views. Both construct explicit 3D Gaussian representations for label propagation, whereas our method propagates labels through the implicit correspondence in a frozen NVS model. All learning-based baselines are re-trained on the same ScanNet split.

Evaluation protocol. We evaluate segmentation on both input views and rendered novel views separately. Input view metrics measure source segmentation

Table 1: Comparisons of RGB reconstruction and panoptic segmentation on ScanNet [12]. Input-view segmentation reflects the quality of the shared query decoder, and novel-view segmentation measures the combined effect of rendering and panoptic propagation.

	Novel View Synthesis			Input Views (2D-only)		Novel Views (3D-aware)	
	PSNR↑	SSIM↑	LPIPS↓	mIoU↑	PQ↑	mIoU↑	PQ↑
2D segmentation							
Mask2Former [10]	-	-	-	**0.6186**	0.5925	-	-
LSeg [29]	-	-	-	0.3976	-	-	-
Joint NVS and scene understanding							
LSM [13]	20.96	0.7245	0.3176	0.2810	-	0.2707	-
SIU3R [51]	25.88	<u>0.8220</u>	<u>0.1831</u>	<u>0.5899</u>	**0.6565**	<u>0.5894</u>	**0.6565**
Ours	**33.56**	**0.9109**	**0.1149**	**0.6186**	<u>0.5949</u>	**0.5949**	<u>0.6092</u>

quality independent of the rendering pipeline. Novel view metrics evaluate the full system including panoptic propagation.

Metrics. For novel view synthesis, we report PSNR, SSIM, and LPIPS. For scene understanding, we report semantic mIoU and panoptic quality (PQ) [24] defined as

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\mathrm{TP}|}}_{\mathrm{SQ}} \times \underbrace{\frac{|\mathrm{TP}|}{|\mathrm{TP}| + \frac{1}{2}|\mathrm{FP}| + \frac{1}{2}|\mathrm{FN}|}}_{\mathrm{RQ}}, \tag{6}$$

where TP, FP, and FN denote matched, unmatched predicted, and unmatched ground-truth segments, respectively. SQ measures boundary precision of matched segments, and RQ measures instance-level recognition. We report metrics on input views and on novel target views.

4.2 Main Results: Novel View Reconstruction and Segmentation

Table 1 shows the evaluation results of our method against baselines on ScanNet [12]. The results show that panoptic propagation preserves rendering quality while achieving competitive segmentation with a frozen novel view synthesis model.

Panoptic propagation preserves rendering quality. Our method achieves 33.56 dB PSNR, the highest among all baselines. The segmentation path does not update the rendering weights, so RGB synthesis uses the same model as the original NVS path. The baselines SIU3R [51] and LSM [13] couple reconstruction with understanding in a shared 3D Gaussian representation, reaching at most 25.88 dB. This separation preserves the rendering performance of the NVS backbone.

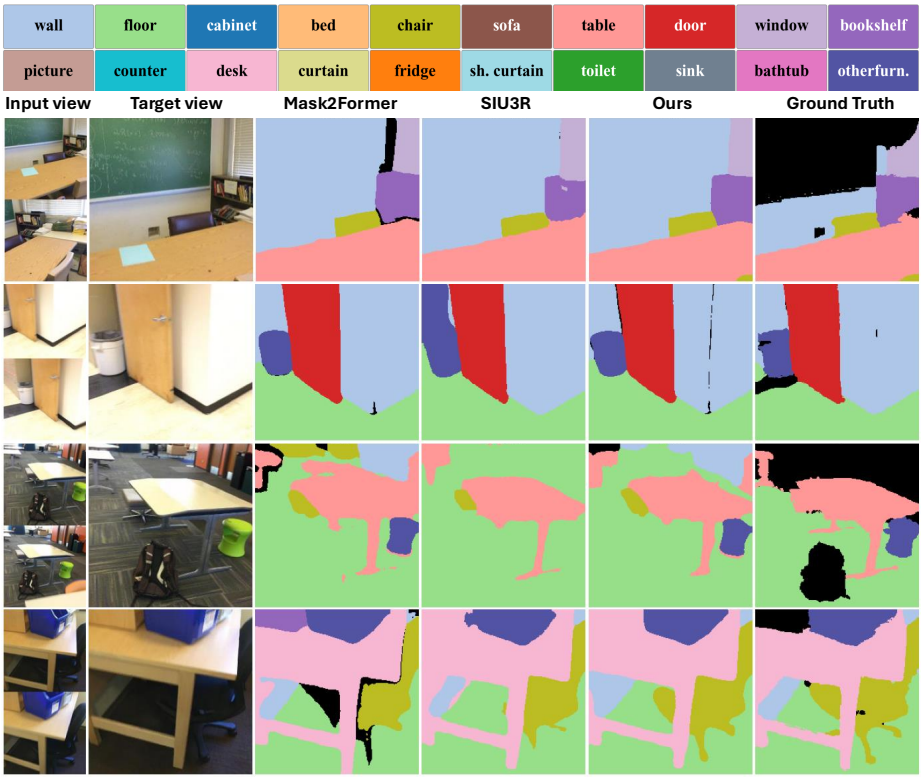

Fig. 4: Qualitative results of novel-view panoptic segmentation on ScanNet [12]. From left to right: Input view, Target view, Panoptic segmentation results from Mask2Former [10], SIU3R [51], ours, and ground truth.

Frozen propagation achieves competitive segmentation. On input views, our method matches the per-view Mask2Former [10] baseline in mIoU (0.6186) at the first row of Table 1, as both share the same query-based decoder architecture. On novel views, our method reaches 0.5949 mIoU, achieving better results than SIU3R [51] (0.5894) despite requiring no joint training. At the last row, the mIoU drop from input views (0.6186) to novel views (0.5949) reflects propagation error. Binary channel encoding loses fine-grained boundary detail, and regions with low source-view overlap accumulate uncertainty.

SIU3R [51] shows better results than our method in terms of PQ (0.6565 vs. 0.6092). We attribute this gap to their multi-view mask aggregation mechanism, which rasterizes semantic attributes through 3D Gaussians and enforces cross-view instance consistency in 3D space. Our propagation operates in 2D token space without explicit geometric reasoning, making instance boundaries more susceptible to spatial ambiguity in the frozen attention.

Qualitative comparison. Fig. 4 compares novel-view panoptic segmentation across methods. Our method propagates panoptic labels through the same frozen

Table 2: Evaluation under different overlap ranges on ScanNet [12]. Under low overlap, the PQ gap between SIU3R and ours narrows from 0.0473 to 0.0011, indicating that attention-based propagation degrades more gracefully than geometry-dependent rasterization. All the results use the same model reported in Table 1.

Overlap	Method	PSNR↑	SSIM↑	LPIPS↓	mIoU↑	PQ↑
High (0.3–0.8)	SIU3R [51]	25.88	0.8220	0.1831	0.5894	**0.6565**
	Ours	**33.56**	**0.9109**	**0.1149**	**0.5949**	0.6092
					PQ gap	0.0473
Low (0.01–0.3)	SIU3R [51]	21.33	0.7147	0.2934	0.5458	**0.5737**
	Ours	**26.47**	**0.7711**	**0.2544**	**0.5477**	0.5726
					PQ gap	0.0011

Table 3: Results of design alternatives. We evaluate several strategies in Fig. 5. Rows 1 and 2 correspond to the architecture in Fig. 5(a), where Row 1 does not freeze the parameters of the unposed NVS model but trainable. Row 3 corresponds to the design in Fig. 5(b), while Row 4 corresponds to our proposed model in Fig. 5(c). The column 2 (NVS) indicates whether the model parameters of the unposed NVS model were released for training (learnable) or kept frozen.

Strategy	Fig.	NVS	PSNR↑	SSIM↑	LPIPS↓	mIoU↑	PQ↑
Joint decoder features	Fig. 5(a)	Release	24.76	0.7502	0.2774	0.4146	0.4716
Joint decoder features	Fig. 5(a)	Freeze	**33.56**	**0.9109**	**0.1149**	0.2008	0.1816
Segment rendered image	Fig. 5(b)	Freeze	**33.56**	**0.9109**	**0.1149**	**0.6239**	0.5740
Ours (Propagation)	Fig. 5(c)	Freeze	**33.56**	**0.9109**	**0.1149**	0.5949	**0.6092**

rendering model used for RGB synthesis, so the segmentation boundaries align with the rendered image appearance. SIU3R [51] produces segmentation through 3D Gaussian rasterization separately from RGB rendering, which can misalign the predicted labels with the rendered image. Mask2Former segments the rendered image directly, so any rendering imperfection propagates into the segmentation.

4.3 Emerging Properties at Low-Overlap Scenarios

Table 2 evaluates robustness under sparse geometric overlap, using view pairs with depth-based IoU in $[0.01, 0.3]$ following SIU3R [51]. This setting stresses label propagation because large target regions are weakly observed or dis-occluded from the source views, making explicit geometry estimation less reliable. Our method maintains a 5.14 dB PSNR advantage and slightly higher mIoU under this setting. More importantly, the PQ gap to SIU3R shrinks from 0.0473 under standard overlap to 0.0011 under low overlap. This suggests that the frozen NVS transformer retains a learned extrapolation capability. Even when explicit

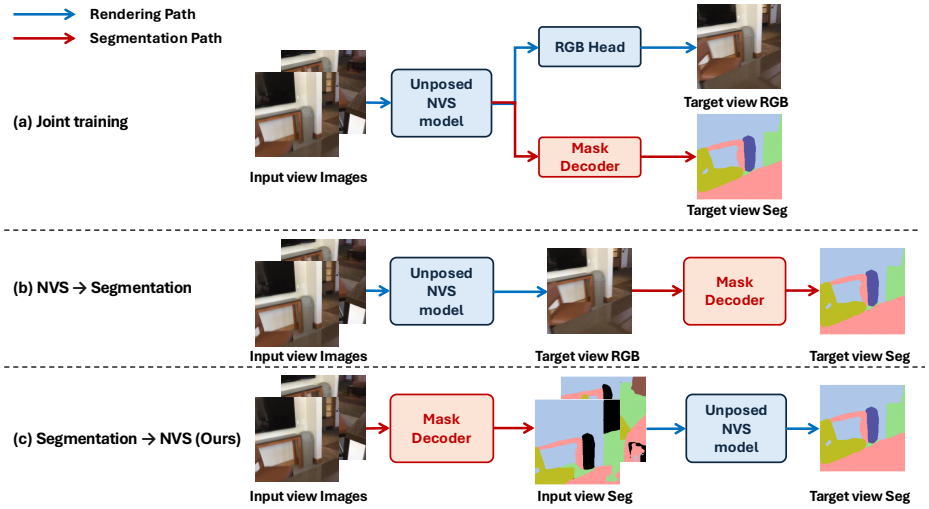

Fig. 5: Design alternatives. (a) Joint training attaches a segmentation head to intermediate features of the NVS rendering decoder. (b) NVS→Segmentation applies a per-view segmenter, Mask2Former [10], to the rendered novel-view image. (c) Our method first segments the input views, encodes the panoptic maps into binary channel representations, and propagates them through the frozen NVS model to produce target-view segmentation. Quantitative comparisons are reported in Table 3.

reconstruction becomes unreliable under sparse overlap, the token-level correspondence still transfers useful panoptic structure.

4.4 Design Alternatives for Novel-View Panoptic Segmentation

Table 3 compares three strategies for integrating panoptic segmentation with the NVS model [44], validating the design of our decoupled propagation approach. The three strategies, illustrated in Fig. 5, differ in how segmentation is integrated with the rendering pipeline.

The first strategy (Fig. 5(a)) attaches a segmentation head to intermediate features of the frozen rendering decoder. As shown in the second row of Table 3, these features are optimized for appearance reconstruction rather than semantic discrimination, causing segmentation performance to collapse to 0.2008 mIoU and 0.1816 PQ. In the first row of Table 3, allowing the NVS model to be jointly optimized with the segmentation head partially recovers segmentation accuracy, but introduces interference between photometric and segmentation objectives. This joint training degrades both rendering quality (24.76 dB) and segmentation performance (0.4146 mIoU, 0.4716 PQ).

The second strategy (Fig. 5(b)) applies a per-view segmenter [10] to each rendered novel-view image independently. While this avoids modifying the NVS model, the third row of Table 3 shows that PQ drops to 0.5740 due to inconsistent

Table 4: Modular composition and cross-dataset transfer. The Segmenter column lists the source-view segmenter and its encoder. SIU3R and our *M2F* variant share a Mask2Former [10] head but differ in encoder, CroCo [49] for SIU3R and DINOv2 [32] for ours. *PanSt3R* [61] is an off-the-shelf predictor that adds no segmentation-loss training. "Pose" indicates whether the target view uses the ground-truth pose (GT) or the estimated latent Plücker embedding (latent). We evaluate Replica [43] zero-shot, without any Replica fine-tuning.

Method	Segmenter	Pose	ScanNet [12]				Replica [43] (zero-shot)			
			PSNR↑	SSIM↑	PQ↑	mIoU↑	PSNR↑	SSIM↑	PQ↑	mIoU↑
SIU3R [51]	M2F (CroCo)	GT	25.88	0.822	**0.657**	0.589	14.05	0.527	0.186	0.171
Ours	M2F (DINOv2)	latent	**33.56**	**0.911**	0.609	**0.595**	–	–	–	–
	PanSt3R [61]	GT	25.02	0.782	0.510	0.366	21.89	0.751	0.246	0.256
		latent	28.28	0.850	0.580	0.593	**23.52**	**0.786**	**0.394**	**0.454**

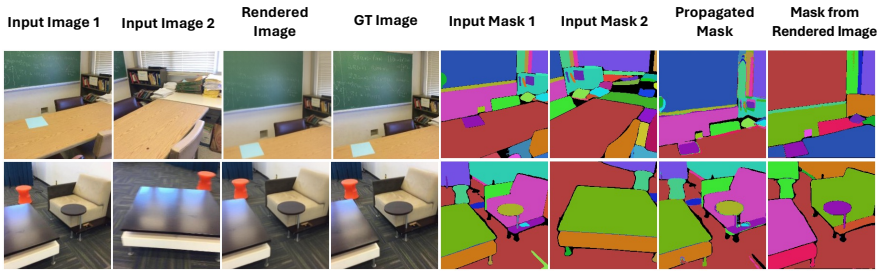

Fig. 6: Replacing the input view segmentation model. The novel view synthesis model propagates SAM2 [35] predictions (Input Mask 1 and 2) to the novel viewpoint (Propagated Mask) while preserving instance identity across views.

instance IDs across views, indicating that high rendering quality alone does not ensure cross-view panoptic consistency.

Finally, our method (Fig. 5(c)) propagates panoptic logits decoded from the source views through the frozen NVS model. As shown in the fourth row of Table 3, this decoupled design preserves the full rendering quality (33.56 dB) while achieving the best panoptic quality (0.6092 PQ), outperforming joint-training alternatives without compromising rendering fidelity.

4.5 Modular Composition and Cross-Dataset Transfer

Our decoupled design lets us replace the source-view segmenter with an off-the-shelf predictor without retraining the rest of the system. We choose PanSt3R [61] because it produces source-view panoptic labels that transfer across datasets, which lets us test propagation under cross-dataset evaluation. We replace our shared query decoder with it to supply source-view labels, testing whether the frozen propagation path generalizes beyond our trained segmentation module, and report the results in Table 4. On ScanNet, *Ours(PanSt3R)* stays competitive without any ScanNet fine-tuning of the segmenter, and *Ours(M2F)* keeps

the rendering advantage over SIU3R [51]. On Replica, *Ours(PanSt3R)* outperforms SIU3R [51] on every RGB and segmentation metric (23.52 vs. 14.05 dB PSNR, 0.394 vs. 0.186 PQ, 0.454 vs. 0.171 mIoU). The higher ScanNet PQ of SIU3R [51] thus partly reflects in-domain specialization, while reconstruction-free propagation transfers better across datasets. We therefore frame our contribution as reconstruction-free label propagation with high rendering quality and transferable segmentation, not as overall PQ superiority. The latent-query variant matches or exceeds the GT-pose variant on both datasets. The estimated latent Plücker embedding therefore carries enough view information for instance-level propagation.

The same modularity holds for a class-agnostic segmenter. As the source-view segmenter, we replace Mask2Former [10] with SAM2 [35] and propagate its masks through the same frozen NVS model. Fig. 6 shows that the propagated masks preserve instance identity across views with comparable visual quality. These results show that the propagation stage can take labels from different source-view segmenters without retraining.

5 Conclusion

This work demonstrates that a view synthesis model, trained for novel view RGB image reconstruction, propagates panoptic labels to novel viewpoints with quality comparable to methods that build explicit 3D representations with dedicated segmentation objectives. The only learned segmentation component is a shared query decoder that produces source-view labels, and the frozen rendering transformer performs the cross-view propagation. The implicit correspondence inside the transformer thus generalizes beyond appearance and encodes geometric structure that can transfer view-consistent dense prediction signals such as panoptic labels. This result suggests that large view synthesis models can serve as an alternative propagation mechanism to explicit 3D reconstruction for novel-view panoptic labels. As long as the transferred signal stays valid across viewpoints, large view synthesis models may also be applicable to other view-consistent dense prediction signals.

Limitation. Our pipeline requires a full forward pass of the view synthesis model for every target view, which becomes a bottleneck when only segmentation at a known viewpoint is needed. The propagation also assumes that the transferred signal stays valid across viewpoints. This holds for semantic classes and instance identities. View-dependent quantities such as depth do not satisfy this assumption, and extending the propagation to them would require a camera-aware transformation within the propagation path.

Acknowledgement. This work was supported by IITP grant (RS-2021-II211343: AI Graduate School Program at Seoul National Univ. (5%), RS-2025-25442338: AI star Fellowship Support Program at Seoul National Univ. (20%), and RS-2026-25517417: Development of Compression, Reconstruction, and Rendering Technologies for Free-viewpoint Media. (75%)). POSCO DX Company, Ltd. provided generous support for this research.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021) 4
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19697–19705 (2023) 4
3. Bhalgat, Y., Laina, I., Henriques, J.F., Zisserman, A., Vedaldi, A.: Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. In: Advances in Neural Information Processing Systems (2023) 3
4. Bhalgat, Y., Laina, I., Henriques, J.F., Zisserman, A., Vedaldi, A.: N2f2: Hierarchical scene understanding with nested neural feature fields. In: European Conference on Computer Vision. pp. 197–214. Springer (2024) 2, 3
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021) 4
6. Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 4
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: International Conference on Learning Representations (2015) 4
8. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: European Conference on Computer Vision (2024) 4
9. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. In: International Conference on Learning Representations (2023) 4, 7
10. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 4, 7, 8, 9, 10, 11, 13, 14, 15
11. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Advances in Neural Information Processing Systems. vol. 34, pp. 17864–17875 (2021) 4
12. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017) 3, 9, 10, 11, 12, 14
13. Fan, Z., Zhang, J., Cong, W., Wang, P., Li, R., Wen, K., Zhou, S., Kadambi, A., Wang, Z., Xu, D., Ivanovic, B., Pavone, M., Wang, Y.: Large spatial model: End-to-end unposed images to semantic 3d. In: Advances in Neural Information Processing Systems (2024) 2, 3, 9, 10
14. Fu, X., Zhang, S., Chen, T., Lu, Y., Zhu, L., Zhou, X., Geiger, A., Liao, Y.: Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In: International Conference on 3D Vision. pp. 1–11. IEEE (2022) 3

15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2961–2969 (2017) 4
16. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. In: International Conference on Learning Representations (2024) 4
17. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024) 4
18. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2989–2998 (2023) 4
19. Jiang, H., Tan, H., Wang, P., Jin, H., Zhao, Y., Bi, S., Zhang, K., Luan, F., Sunkavalli, K., Huang, Q., Pavlakos, G.: Rayzer: A self-supervised large view synthesis model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025) 2, 4, 5, 8
20. Jiao, S., Dong, H., Yin, Y., Jie, Z., Qian, Y., Zhao, Y., Shi, H., Wei, Y.: Clip-gs: Unifying vision-language representation with 3d gaussian splatting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4670–4680 (2025) 3
21. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snavely, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. In: International Conference on Learning Representations (2025) 2, 4, 5
22. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023) 1, 2, 3, 4
23. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19729–19739 (October 2023) 3
24. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9404–9413 (2019) 4, 10
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 4
26. Kolodiazhnyi, M., Vorontsova, A., Konushin, A., Rukhovich, D.: Oneformer3d: One transformer for unified point cloud segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20943–20953 (2024) 3
27. Kundu, A., Genova, K., Yin, X., Fathi, A., Pantofaru, C., Guibas, L.J., Tagliasacchi, A., Dellaert, F., Funkhouser, T.: Panoptic neural fields: A semantic object-aware neural scene representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12871–12881 (2022) 3
28. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision. pp. 71–91. Springer (2024) 4
29. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022) 9, 10

30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (2020) 1, 2, 4
31. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022) 4
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research **3**(1) (2024) 4, 5, 7, 14
33. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 3
34. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 3
35. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations (2025) 8, 14, 15
36. Sajjadi, M.S.M., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lučić, M., Duckworth, D., Dosovitskiy, A., Uszkoreit, J., Funkhouser, T., Tagliasacchi, A.: Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6229–6238 (June 2022) 2, 4
37. Sajjadi, M.S., Duckworth, D., Mahendran, A., Van Steenkiste, S., Pavetic, F., Lucic, M., Guibas, L.J., Greff, K., Kipf, T.: Object scene representation transformer. In: Advances in Neural Information Processing Systems. vol. 35, pp. 9512–9524 (2022) 4
38. Sajjadi, M.S., Mahendran, A., Kipf, T., Pot, E., Duckworth, D., Lučić, M., Greff, K.: Rust: Latent neural scene representations from unposed imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17297–17306 (2023) 2, 4
39. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In: IEEE International Conference on Robotics and Automation (2023) 3
40. Siddiqui, Y., Porzi, L., Bulò, S.R., Müller, N., Nießner, M., Dai, A., Kontschieder, P.: Panoptic Lifting for 3D Scene Understanding With Neural Fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 2, 3
41. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: International Conference on Learning Representations Workshop (2014) 2, 6
42. Smart, B., Zheng, C., Laina, I., Prisacariu, V.A.: Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. arXiv preprint arXiv:2408.13912 (2024) 4

43. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica Dataset: A Digital Replica of Indoor Spaces. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019) 9, 14
44. Wang, H., Ye, K., Li, Y., Chen, W., Chen, B.: The less you depend, the more you learn: Synthesizing novel views from sparse, unposed images without any 3d knowledge. In: International Conference on Learning Representations (2026) 2, 4, 5, 6, 8, 13
45. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5294–5306 (2025) 4
46. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021) 4
47. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024) 4
48. Wang, Y., Wei, X., Lu, M., Kang, G.: Plgs: Robust panoptic lifting with 3d gaussian splatting. IEEE Transactions on Image Processing (2025) 3
49. Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., Jérôme, R.: CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In: Advances in Neural Information Processing Systems (2022) 14
50. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., Zhang, J.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In: Advances in Neural Information Processing Systems. pp. 19114–19138 (2024) 3
51. Xu, Q., Wei, D., Zhao, L., Li, W., Huang, Z., Ji, S., Liu, P.: SIU3R: Simultaneous Scene Understanding and 3D Reconstruction Beyond Feature Alignment. In: Advances in Neural Information Processing Systems (2025) 2, 3, 9, 10, 11, 12, 14, 15
52. Ye, B., Liu, S., Xu, H., Xueting, L., Pollefeys, M., Yang, M.H., Songyou, P.: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In: International Conference on Learning Representations (2025) 4
53. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: European Conference on Computer Vision (2024) 3
54. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021) 4
55. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19447–19456 (2024) 4
56. Zhang, K., Bi, S., Tan, H., Xiangli, Y., Zhao, N., Sunkavalli, K., Xu, Z.: Gs-lrm: Large reconstruction model for 3d gaussian splatting. In: European Conference on Computer Vision. pp. 1–19. Springer (2024) 4

57. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 3
58. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21676–21685 (2024) 3
59. Zhu, R., Qiu, S., Wu, Q., Hui, K.H., Heng, P.A., Fu, C.W.: Pcf-lift: Panoptic lifting by probabilistic contrastive fusion. In: European Conference on Computer Vision. pp. 92–108. Springer (2024) 3
60. Zuo, X., Samangouei, P., Zhou, Y., Di, Y., Li, M.: Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. International Journal of Computer Vision **133**(2), 611–627 (2025) 3
61. Zust, L., Cabon, Y., Marrie, J., Antsfeld, L., Chidlovskii, B., Revaud, J., Csurka, G.: Panst3r: Multi-view consistent panoptic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025) 3, 8, 9, 14