PCF-Lift: Panoptic Lifting by Probabilistic Contrastive Fusion

Runsong Zhu[1], Shi Qiu[1], Qianyi Wu[2], Ka-Hei Hui[1], Pheng-Ann Heng[1], and Chi-Wing Fu[1]

[1] The Chinese University of Hong Kong
[2] Monash University

Abstract. Panoptic lifting is an effective technique to address the 3D panoptic segmentation task by unprojecting 2D panoptic segmentations from multi-views to 3D scene. However, the quality of its results largely depends on the 2D segmentations, which could be noisy and error-prone, so its performance often drops significantly for complex scenes. In this work, we design a new pipeline coined **PCF-Lift** based on our **P**robabilistic **C**ontrastive **F**usion (PCF) to learn and embed probabilistic features throughout our pipeline to actively consider inaccurate segmentations and inconsistent instance IDs. Technical-wise, we first model the probabilistic feature embeddings through multivariate Gaussian distributions. To fuse the probabilistic features, we incorporate the probability product kernel into the contrastive loss formulation and design a cross-view constraint to enhance the feature consistency across different views. For the inference, we introduce a new probabilistic clustering method to effectively associate prototype features with the underlying 3D object instances for the generation of consistent panoptic segmentation results. Further, we provide a theoretical analysis to justify the superiority of the proposed probabilistic solution. By conducting extensive experiments, our PCF-lift not only significantly outperforms the state-of-the-art methods on widely used benchmarks including the ScanNet dataset and the challenging Messy Room dataset (4.4% improvement of scene-level PQ), but also demonstrates strong robustness when incorporating various 2D segmentation models or different levels of hand-crafted noise.

Keywords: Panoptic Lifting · Probabilistic Contrastive Fusion · Probabilistic Feature Embeddings

1 Introduction

3D panoptic segmentation [11,15,32,33,39,41,51] is a challenging 3D vision task, requiring the prediction of both semantic segmentation labels and instance segmentation labels. This task enables a comprehensive understanding of 3D scenes, thus facilitating many downstream applications, *e.g.*, VR/AR, robotics, *etc.*

Observing that the generalizability and applicability of existing 3D panoptic segmentation methods are limited by the scarcity of 3D training data, recent studies [3, 25, 40, 43, 49] suggest the idea of leveraging 2D panoptic segmentation

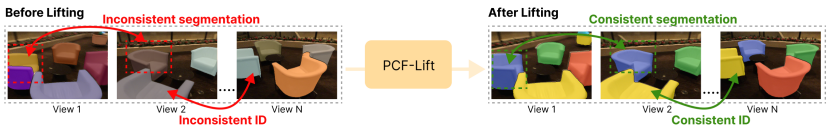

Fig. 1: Our PCF-Lift method unprojects 2D panoptic segmentation predictions to 3D domain, facilitating the generation of consistent panoptic segmentation masks. For simplicity and clarity, we highlight instance segmentation masks.

information predicted by foundation models [9, 24, 50]. Although 2D panoptic segmentation only provides image-based understanding, panoptic lifting [3, 40] is an effective technique to learn implicit 3D panoptic fields from 2D panoptic predictions, supporting the generation of coherent and view-consistent panoptic segmentation across different views.

Recent works [3, 40, 44] on panoptic lifting mainly focus on solving the challenging instance-related issues, given that semantic predictions can be effectively associated across different views [49]. In practice, as shown in Fig. 1, one direct issue is *inconsistent IDs* that the same 3D object is assigned to different instance IDs in two different views for 2D panoptic prediction, which complicates the straightforward fusion process. To bypass the *inconsistent IDs* issue, Panoptic Lifting [40] directly fits instance ID permutations during the training process to learn deterministic feature embeddings for instance representation. Furthermore, Contrastive Lift [3] extends the scalability of Panoptic Lifting by using the contrastive learning technique to optimize the deterministic feature embeddings. However, existing methods still struggle to achieve satisfactory performance on complex scenes, as they overlook another interrelated issue: *inconsistent segmentation*. Specifically, due to the presence of inaccurate segmentation, the same object can be segmented inconsistently across two views: *e.g.*, the chair in Fig. 1 is segmented into two parts in "view 1" but as a whole in "view 2". As existing methods use deterministic feature embeddings to learn 3D instance segmentation, the issue of *inconsistent segmentation* inherently introduces noise to training data and thus poses a robustness challenge for the models.

In this paper, we introduce an effective probabilistic contrastive fusion solution to collectively address the two issues. For the issue of *inconsistent segmentation*, we propose to learn probabilistic feature embeddings rather than deterministic feature embeddings used in [3, 40]. Our key insight is the development of probabilistic features, which correspond to distributions, allows for the incorporation of uncertainty modeling and enhances robustness to noise. This leads to a stable model optimization and, ultimately, results in a more reliable representation of instance information. Given these benefits, we develop probabilistic feature embeddings based on multivariate Gaussian distributions. Particularly, to enable contrastive learning among different Gaussian distributions, we devise the Probability Product (PP) Kernel [18] to measure the probabilistic feature similarities. Accordingly, a probabilistic clustering algorithm is introduced

in the inference phase to generate consistent panoptic segmentation, using the measured probabilistic feature similarities. For the issue of *inconsistent IDs*, we devise the contrastive loss to fuse the probabilistic feature embeddings. In addition to formulating the loss term with each single view observation as inspired by the prior work [3], we introduce a novel cross-view constraint to facilitate an effective model training from segmentation results with inconsistent instance IDs. By dynamically exploiting feature pairs from different views, our proposed constraint offers an effective formulation to further enhance the feature consistency of the same 3D object instance across multiple views, and thus improves the panoptic lifting performance.

To evaluate the effectiveness of our method, we conduct experiments on the ScanNet dataset [12] and the Messy Rooms dataset [3]. Furthermore, we conduct experiments to demonstrate the robustness of our probabilistic method to variations in segmentation models and different levels of noise. In addition to the experimental outcomes, our theoretical analysis from an optimization perspective demonstrates that the proposed probabilistic representation can be seen as a more flexible and generalized form compared to the previous deterministic representation [3].

Our contributions are summarized as follows:

- We introduce a probabilistic contrastive fusion solution (PCF-Lift) to effectively unproject the 2D panoptic segmentations to the 3D domain by collectively considering the issues of *inconsistent segmentation* and *inconsistent IDs*.
- To fuse the probabilistic feature embeddings modeled by multivariate Gaussian distributions, we reformulate the contrastive loss with the Probability Product kernel and propose a novel cross-view constraint to further encourage the multi-view consistency.
- Coupled with a new probabilistic clustering algorithm, our proposed method outperforms the state-of-the-art methods consistently on the ScanNet dataset and the Messy Room dataset.

2 Related Works

2.1 Traditional 2D and 3D Panoptic Segmentation

The 2D Panoptic segmentation task was initially introduced in [23]. Despite the notable advancements made by subsequent works [8–10, 37, 47], it remains challenging to panoptically understand individual images, while avoiding inconsistent instance recognitions across different image views of the scene.

To enhance the panoptic understanding of the real world, 3D panoptic segmentation focuses on segmenting pre-computed 3D structures [15,32,41,51] (*e.g.*, point clouds, voxels) or performing simultaneous 3D reconstruction or segmentation from 2D images [11, 33, 39]. However, its generalizability is rather limited, largely due to the significant difference in scale between 2D and 3D training data. Given the capabilities of recent 3D reconstruction techniques [6, 20, 31] and 2D

panoptic segmentation models [9, 50], in this work, we explore 3D panoptic segmentation by leveraging multi-view images without explicit 3D input data.

2.2 Multi-view Fusion

Recently, researches in 3D reconstruction have made tremendous progress [2, 6, 20, 29, 31]. Beyond novel view synthesis, we can utilize 3D reconstruction techniques as a tool to fuse 2D information like semantics and features in the 3D space [5, 14, 17, 21, 25, 35, 43, 46, 48]. For example, Semantic-NeRF [49] learns a semantic field from the 2D semantic segmentation, demonstrating the robustness and effectiveness of the semantic fusion. Later works [21, 25, 43] propose to fuse the unsupervised 2D dense feature for various segmentation and editing applications, showcasing the potential to adapt the 2D zero-shot models to 3D domain. In this work, we study the task of panoptic fusion, integrating both instance and semantic information to obtain a comprehensive understanding of the 3D scene. This task is significantly more challenging than the previous semantic fusion task [49], since we need to incorporate additional instance fusion and achieve instance label consistency. Unlike recent methods (*e.g.*, Panoptic-Lifting [40] and Contrastive Lift [3]), which utilize deterministic feature embeddings to represent instance information, we develop a novel probabilistic solution, actively considering the error-prone and noisy nature of the 2D segmentations, such that we can duly enhance the effectiveness and robustness of instance fusion.

2.3 Probabilistic Representation

Probabilistic feature embedding is a popular tool employed in various tasks, *e.g.*, image generation [22, 38], normal estimation [1], video understanding [36], point cloud understanding [4], and prototype embeddings for few-shot detection [42]. Motivated by its high capabilities in estimating aleatoric uncertainty [13] and addressing data noise, we innovate a new panoptic-lifting pipeline for 3D panoptic segmentation by formulating new modules based on probabilistic feature embeddings, including a reformulated contrastive loss with the probability product kernel [18] and an effective cross-view constraint to enhance the multi-view consistency. Further, we also design a novel probabilistic clustering algorithm to facilitate the generation of consistent panoptic segmentation.

3 Method

Given a set of posed images $\{\mathcal{I}\}$ associated with 2D panoptic segmentation predictions (*i.e.*, semantic masks $\{\mathcal{H}\}$ and instance masks $\{\mathcal{K}\}$) generated by 2D segmentation models, our goal is to learn accurate 3D panoptic fields that can be rendered into consistent panoptic segmentation results over different views.

The overview of PCF-Lift is illustrated in Fig. 2. Specifically, the 3D panoptic fields include a semantic field, an instance field, a density field, and a color field. Particularly, considering the significance of instance-related issues, we propose

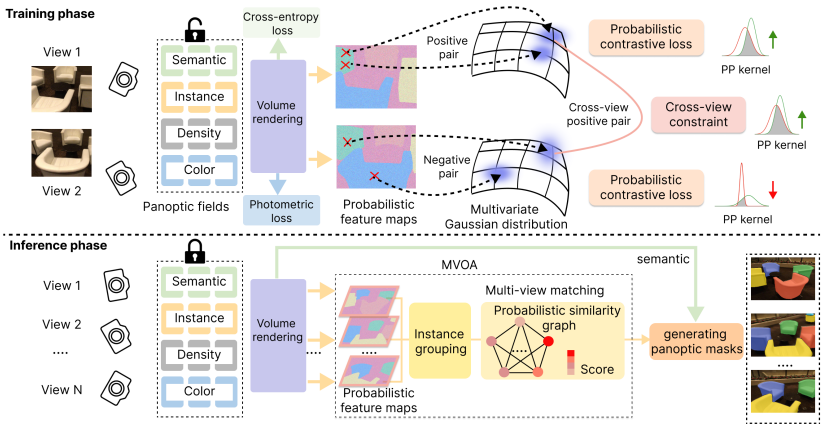

Fig. 2: Overview of PCF-Lift. The 3D panoptic fields include a semantic field, an instance field, a density field, and a color field. To solve the instance-related issues, we propose to learn probabilistic feature embeddings in the instance field (see Sec. 3.1). During the training phase, given two camera views, we can render the probabilistic feature maps from the instance field via volume rendering. To optimize the probabilistic instance field, we devise the probabilistic contrastive loss with Probability Product (PP) kernel [18], and propose a cross-view constraint to further enhance the feature consistency from different views (see Sec. 3.2). Similarly, we can render the semantic and color predictions, and adopt photometric loss and cross-entropy loss to optimize the semantic field, the density field, and the color field. During the inference phase, we design a novel multi-view object association (MVOA) algorithm for the generation of consistent panoptic segmentations (see Sec. 3.3).

to learn probabilistic feature embeddings in the instances field (see Sec. 3.1). During the training phase, we jointly train the whole panoptic fields. Given two camera views, we can render the probabilistic feature maps from the instance field via volume rendering [19]. To optimize the probabilistic feature embeddings, we develop the contrastive loss with the Probability Product kernel [18], and propose a cross-view constraint to further enhance the feature consistency from different views (see Sec. 3.2). Similarly, the semantic predictions and color predictions are also rendered via volume rendering, where the semantic field, the density field, and the color field are optimized by the photometric loss and cross-entropy loss [3, 6, 40, 49]. In the inference phase, we design a probabilistic clustering (*i.e.*, multi-view object association) algorithm to effectively identify the prototype features of underlying 3D object instances for the generation of consistent panoptic segmentation results across any given views (see Sec. 3.3).

3.1 Probabilistic Feature Embeddings

To provide a robust instance representation, we propose to learn probabilistic feature embeddings in the instance field, which maps each 3D point to a random

variable $\mathcal{F}$, modeled as an N-dimensional multivariate Gaussian distribution $\mathcal{F} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Here, μ is the mean vector, indicating the central feature values and $\mathbf{\Sigma}$ is a diagonal covariance matrix $\text{diag}(\sigma^2)$ with $\sigma^2 = ({\sigma^{(1)}}^2, {\sigma^{(2)}}^2, \cdots, {\sigma^{(N)}}^2)$. Concretely, for any given query point $\mathbf{x} \in \mathbb{R}^3$, the instance field predicts $(\mu, \sigma^2) \in \mathbb{R}^{2N}$. Similar to the rendering of the color field, for each pixel in a given camera view, we can render its corresponding Gaussian distribution feature via volume rendering. A significant advantage of using probabilistic features rather than deterministic features is the ability of Gaussian distributions that assign different covariance values for uncertainty modeling. This property is crucial as it aids in reducing the impact of noise, thereby enhancing the robustness and accuracy of feature embeddings in representing complex 3D scenes.

Intuitively, the similarity between two rendered Gaussian distributions indicates whether the corresponding pixels belong to the same instance. To quantify the similarities between different Gaussian distributions, we employ the Probability Product (PP) kernel [18] K_ρ. Specifically, given two Gaussian distributions $\mathcal{F}_i \sim \mathcal{N}(\mu_i, \mathbf{\Sigma}_i)$ and $\mathcal{F}_j \sim \mathcal{N}(\mu_j, \mathbf{\Sigma}_j)$, the corresponding kernel can be explicitly formulated as

$$K_\rho\left(\mathcal{F}_i, \mathcal{F}_j\right) = \left(\prod_{d=1}^{N}\left(\frac{\sigma_i^{(d)^2}}{\sigma_j^{(d)^2}} + \frac{\sigma_j^{(d)^2}}{\sigma_i^{(d)^2}}\right)/2\right)^{-\frac{1}{2}} \exp\left(-\sum_{d=1}^{N}\left(\mu_i^{(d)} - \mu_j^{(d)}\right)^2 / 4\left(\sigma_i^{(d)^2} + \sigma_j^{(d)^2}\right)\right). \tag{1}$$

Note that the results produced by the PP kernel fall within the range of 0 (minimal similarity) to 1 (maximal similarity), the same as the Radial Basis Function (RBF) kernel used in the deterministic method [3].

3.2 Training Probabilistic Feature Embeddings

Probabilistic Contrastive Loss. During the training phase, we leverage the contrastive learning technique to optimize the probabilistic feature embeddings. Specifically, the contrastive loss used in [7, 27, 34, 45] tends to increase the predicted feature similarities of predefined positive pairs, while decreasing the similarities of negative pairs for training. In our work, a positive pair is defined as two pixels belonging to the same instance, while a negative pair denotes two pixels of different instances within a single image, following the previous method [3]. Since we basically learn probabilistic features, the PP kernel is exploited to measure the features similarity. In general, the loss can be formulated as

$$\mathcal{L}_{\text{pixel-contra}} = -\frac{1}{|\Omega|}\sum_{u\in\Omega}\log\frac{\sum_{u'\in\Omega}\mathbf{1}_{(u,u')}\exp\left(K_\rho\left(\mathcal{F}_u, \mathcal{F}_{u'}\right)\right)}{\sum_{u'\in\Omega}\exp\left(K_\rho\left(\mathcal{F}_u, \mathcal{F}_{u'}\right)\right)}, \tag{2}$$

$$\mathcal{L}_{\text{concen}} = -\frac{1}{|\Omega|}\sum_{u\in\Omega}\log\left(K_\rho\left(\mathcal{F}_u, \frac{\sum_{u'\in\Omega}\mathbf{1}_{(u,u')}\mathcal{F}_{u'}}{\sum_{u'\in\Omega}\mathbf{1}_{(u,u')}}\right)\right), \tag{3}$$

$$\text{and } \mathcal{L}_{\text{contra}} = \mathcal{L}_{\text{pixel-contra}} + \mathcal{L}_{\text{concen}}, \tag{4}$$

where $\mathcal{L}_{\text{pixel-contra}}$ is the pixel-wise contrastive loss, $\mathcal{L}_{\text{concen}}$ is the concentrate loss term, $\mathbf{1}$ is the indicator function of positive pairs, Ω is the set of pixel sample, $\mathcal{F}_u$

and $\mathcal{F}_{u'}$ are the rendered features for pixels u and $u' \in \Omega$ via volume rendering, respectively. Different from previous method [3], we devise probabilistic similarity kernels to calculate a more effective contrastive loss for model training. By optimizing the loss Eq. (4), we can learn a more expressive 3D instance field for robust feature representations, as further analyzed in Sec. 3.4. To avoid the neural network from generating large covariances everywhere, we use an additional regularization term, $\mathcal{L}_{\text{reg}} := \log(\prod_{d=1}^{N} \sigma^{(d)^2})$, to penalize the large covariances.

Cross-view Constraint. To better train the probabilistic feature embeddings, we propose a cross-view constraint to enhance feature consistency for the same object across different views. Given the rendered learned probabilistic feature embedding sets $\{\mathcal{F}^m\}$ and $\{\mathcal{F}^n\}$ for two different views, we devise the PP kernel K_ρ and a predefined threshold τ to collect positive pairs $\mathcal{P}$ that belong to the same object: $\mathcal{P} = \{(\mathcal{F}_r, \mathcal{F}_s) \mid K_\rho(\mathcal{F}_r, \mathcal{F}_s) > \tau, \mathcal{F}_r \in \{\mathcal{F}^m\}, \mathcal{F}_s \in \{\mathcal{F}^n\}\}$. Moreover, as we expect to maximize the similarities between the features that belong to the same 3D object in different views, the cross-view constraint is defined as:

$$\mathcal{L}_{\text{cross}} = -\frac{1}{|\mathcal{P}|} \sum_{(\mathcal{F}_r, \mathcal{F}_s) \in \mathcal{P}} \log\left(K_\rho(\mathcal{F}_r, \mathcal{F}_s)\right). \tag{5}$$

In practice, we use the cross-view constraint only in a later optimization stage (*i.e.*, the last few epochs), where the feature space is sufficiently meaningful to provide reliable positive pairs. The overall loss is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{contra}} + w_{\text{cross}}\mathcal{L}_{\text{cross}} + w_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{6}$$

where w_{cross} and w_{reg} are weight hyper-parameters.

3.3 Multi-view Object Association

During the inference phase, to facilitate consistent panoptic segmentations, we introduce a novel clustering algorithm, named the multi-view object association (MVOA) algorithm. The algorithm aims to extract the prototype feature set from the learned instance field for the assignment of instance labels. In general, through volume rendering, we obtain a large-scale rendered per-pixel probabilistic feature set $\{\mathcal{F}\}$ of the training views, as the input of our algorithm. Considering the efficiency issue, we first conduct an instance grouping operation to gather a smaller feature set $\mathcal{C}$, with the help of inconsistent instance segmentation masks $\{\mathcal{K}\}$ obtained from training views. Then, we construct the probabilistic similarity graph based on $\mathcal{C}$ and design a multi-view matching process to collect the prototype feature set $\mathcal{D}$. The details of instance grouping and multi-view matching operations are introduced as follows.

Instance Grouping. We group the feature set $\{\mathcal{F}\}$ from each individual view. For the view of l-th image, we collect the pixels that belong to instance p into the set $\{i : \mathcal{K}_l^i = p\}$, and the corresponding probabilistic feature set is denoted as

$\{\mathcal{F}_l^i : \mathcal{K}_l^i = p\}$. Then, we group $\{\mathcal{F}_l^i : \mathcal{K}_l^i = p\}$ to a single probabilistic feature by calculating the average: $\mathcal{C}_l^p = \sum_i \mathcal{F}_l^i / \big|\{i : \mathcal{K}_l^i = p\}\big|$. Concurrently, we model the score quantity $\mathcal{S}_l^p$ for $\mathcal{C}_l^p$ as an indicator of feature concentration, which aids in the subsequent multi-view matching process. The indicator function $\Phi(\mathcal{C}_l^p)$ averages the PP kernel similarities between $\mathcal{C}_l^p$ and all features in $\{\mathcal{F}_l^i : \mathcal{K}_l^i = p\}$. Mathematically, $\Phi(\mathcal{C}_l^p)$ follows: $\mathcal{S}_l^p = \Phi(\mathcal{C}_l^p) = \sum_{\{i:\mathcal{K}_l^i=p\}} K_\rho(\mathcal{C}_l^p, \mathcal{F}_l^i) / \big|\{i : \mathcal{K}_l^i = p\}\big|$. By applying this across all observed images, we obtain the grouped feature set $\mathcal{C}$ and the corresponding score set $\mathcal{S}$.

Multi-view Matching. To extract the structural relationships among the features, we construct an unoriented similarity graph $G = (\mathcal{C}, E)$, where the grouped feature set $\mathcal{C}$ represents the nodes and E encompasses the edges that indicate quantitative feature similarities measured by the PP Kernel: $E_{<g,h>} = E_{<h,g>} = K_\rho(\mathcal{C}_g, \mathcal{C}_h)$, where $\mathcal{C}_g, \mathcal{C}_h \in \mathcal{C}$. Then, we extract the prototype feature set $\mathcal{D}$ from the graph G and score set $\mathcal{S}$ by a greedy procedure, which is akin to the classical greedy algorithm known as non-maximum suppression (NMS). The pseudo-code of our proposed MVOA algorithm is presented in Algo. 1. In practice, the selection of hyper-parameter $\mathcal{T}$ is based on the average similarity computed across grouped feature pairs identified within each view.

Generating Panoptic Segmentation Masks. We first follow the MVOA algorithm to collect the prototype feature set $\mathcal{D}$. For each view, including the novel view, we then generate the semantic labels from the learned semantic field to differentiate between the background and foreground. Coupled with a feature map rendered from the probabilistic instance field, for each foreground pixel, we finally determine its corresponding instance label by assigning the candidate index from the set $\mathcal{D}$ that exhibits the highest similarity to the rendered feature. In practice, the MVOA algorithm needs to be conducted only once, while the extracted prototype features $\mathcal{D}$ will be used for generating the panoptic segmentation masks of all test views.

3.4 Theoretical Analysis

Corollary 1 *If the covariances of the given Gaussian distributions are isotropic and fixed,* i.e.*, $\Sigma_i = \Sigma_j = \sigma \boldsymbol{I}$, where σ is a constant scalar, the probability product kernel can be simplified to an RBF kernel.*

The primary distinction between our probabilistic method and the prior deterministic method [3] is the choice of a similarity kernel. Specifically, our probabilistic method exploits the PP kernel, which offers higher flexibility and stronger expressive capability by adjusting the Gaussian mean and covariance compared to the RBF kernel employed in the deterministic method, as demonstrated in Fig 3. The RBF kernel can be considered as a degenerate PP kernel when applied to a Gaussian distribution with an isotropic and fixed covariance, as stated in Corollary 1. From this perspective, the proposed probabilistic method is a

Algorithm 1: Multi-view object association algorithm (MVOA)

Data: Inconsistent instance mask $\{\mathcal{K}\}$,input feature set $\{\mathcal{F}\}$, Threshold $\mathcal{T}$
Result: Prototype features set $\mathcal{D}$

```
// Instance Grouping
C = {}, S = {}
for K_l in {K} do
    IDs = {unique_ID(K_l)}
    for p in IDs do
        C_l^p = Σ_i F_l^i / |{i : K_l^i = p}|
        S_l^p = Φ(C_l^p)
        C ← C ∪ C_l^p; S ← S ∪ S_l^p;
    end
end
// Multi-view matching
D ← {}
while S ≠ empty do
    m ← argmax S
    D ← D ∪ C_m
    C ← C − C_m; S ← S − S_m
    for C_i in C do
        if K_ρ(C_m, C_i) ≥ T then
            C ← C − C_i; S ← S − S_i
        end
    end
end
```

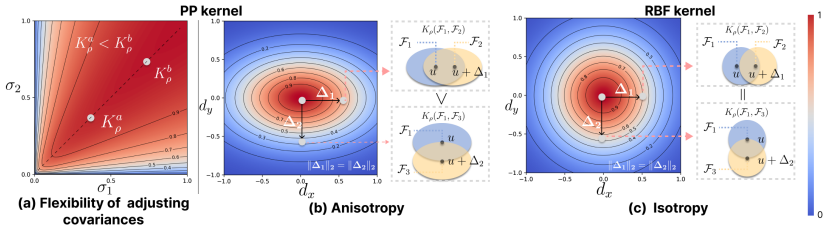

Fig. 3: (a) Flexibility of adjusting covariances. Contour plot of the PP kernel similarity for two Gaussians with different covariance values (σ_1 and σ_2) and fixed mean values in a 1-dimensional case. (b) Anisotropy. Contour plot of the PP kernel similarity for two Gaussians with different Gaussian mean offsets (d_x and d_y) and fixed covariances in a 2-dimensional case. (c) Isotropy. Contour plot of the RBF kernel similarity for two deterministic features with different offsets (d_x and d_y) in a 2-dimensional case.

more general framework, while the deterministic method is a subclass within the broader probabilistic paradigm. The experimental results also verify the effectiveness of our probabilistic solution in the ablation studies of Sec. 4.3.

4 Experiment

4.1 Experimental Settings

Implementation Details. For fair comparison, we adopt the same architecture of TensoRF [6] together with the same layer parameters used in previous works [3,40]. Particularly, our instance field is constructed following the slow-fast architecture of Contrastive Lift [3], where a shallow 5-layer MLP is applied to predict an probabilistic feature embedding for a 3D coordinate input. In practice, the dimension N for the probabilistic feature embedding is 3 in our experiments.

For the training of the instance field, we sample rays from two different views in the last few epochs for optimization (Sec. 3.2), using the cross-view constraint $\mathcal{L}_{\text{cross}}$ with a predefined threshold value $\tau = 0.9$ and weight $w_{\text{cross}} = 0.05$; while in other training epochs, we set the weight w_{cross} to 0 and only sample the rays from single views. Besides, w_{reg} in Eq. (6) is set to 0.001 throughout the whole training process. For the training of the color, density and semantic fields, we adopt the same loss terms and training strategies from previous works [3, 40]. More implementation details are provided in the supplementary material.

Metrics. Since our method is proposed to solve the panoptic-lifting task, we employ scene-level Panoptic Quality (PQ^{scene}) for evaluation, which was introduced in Panoptic Lifting [40] and widely used in the related works [3,40]. Unlike the standard PQ metric [23], this metric particularly considers the consistency of instance IDs across multiple views. By merging predictions and ground truths with consistent instance IDs into subsets for PQ^{scene}-based evaluation, we determine a match of subset pair when the intersection over union (IoU) exceeds 0.5, in line with prior baselines [3, 40]. Since PQ^{scene} is a product of scene-level segmentation quality (SQ^{scene}) and recognition quality (RQ^{scene}), we also provide the SQ^{scene} and RQ^{scene} metrics for more detailed comparisons.

Baselines. We mainly compare the proposed method with the current state-of-the-art methods that target lifting 2D panoptic predictions to 3D, *i.e.*, Contrastive Lift [3] and Panoptic Lifting [40]. Moreover, we compare our method with the recent NeRF-based 3D panoptic segmentation approaches, *i.e.*, Panoptic Neural Fields [26] (PNF) and DM-NeRF [44].

Datasets. Following the previous works [3, 40], we conduct experiments on two public datasets, the ScanNet dataset [12] and the Messy Room dataset [3], for both quantitative and qualitative evaluations. For fair comparisons on ScanNet, following the other state-of-the-art approaches [3, 40], we adopt Mask2Former (M2F) [9] to generate 2D panoptic segmentation predictions, coupled with the same protocol in [40] that maps the COCO [28] vocabulary to 21 classes. For the experiments on Messy Rooms, we use the LVIS [16] vocabulary and the Detic [50] 2D panoptic segmentations, which are also utilized in Contrastive Lift [3].

4.2 Main Experiments

ScanNet Dataset. To verify the performance on real data, we conduct experiments on the ScanNet dataset [12] of 12 scenes. Quantitative comparisons provided in Tab. 1 demonstrate that our method consistently outperforms previous baselines, including both the 3D panoptic segmentation approaches [26, 44] and the recent state-of-the-art methods for lifting 2D panoptic segmentation [3, 40]. Moreover, visual comparisons between our method and the latest state-of-the-art method [3] are presented in Fig. 4, which exhibits our method's capability of achieving a consistent and accurate 3D panoptic segmentation.

Table 1: Results on the ScanNet dataset. We report the PQ^{scene}, SQ^{scene}, and RQ^{scene} metrics. Since Contrastive Lift [3] does not report the performance using the SQ^{scene} and RQ^{scene} metrics, we apply the officially-released pre-trained model and re-run the clustering algorithm to obtain the values reported in this table. For the other metric values, we directly report the ones in previous papers [3, 40].

Method	Venue	Type	SQ^{scene}(%)	RQ^{scene}(%)	**PQ^{scene}(%)**
DM-NeRF [44]	ICLR'23	3D panoptic segmentation	53.3	46.1	41.7
PNF [26]	CVPR'22	3D panoptic segmentation	63.0	50.7	48.3
PNF [26] + GT BBoxes	CVPR'22	3D panoptic segmentation	70.0	55.9	54.3
Panoptic Lifting [40]	CVPR'23	2D panoptic Lifting	73.5	65.0	58.9
Contrastive Lift [3]	NeurIPS'23	2D panoptic Lifting	75.7	63.6	62.0
Ours	-	2D panoptic Lifting	78.5	65.4	63.5

Table 2: Results on the Messy Rooms dataset [3]. Following [3], the PQ^{scene} metric is reported on both "old room" and "large corridor" environments with an increasing number of objects in the scene $(25, 50, 100, 500)$.

Method/ Number	Old Room Environment (%)				Large Corridor Environment(%)				**Mean(%)**
	25	50	100	500	25	50	100	500	
Panoptic Lifting [40]	73.2	69.9	64.3	51.0	65.5	71.0	61.8	49.0	63.2
Contrastive Lift [3]	78.9	75.8	69.1	55.0	76.5	75.5	68.7	52.5	69.0
Ours	80.9	78.3	74.8	60.3	81.0	79.4	74.0	58.8	73.4

Messy Rooms Dataset. We also conduct experiments on the challenging Messy Room dataset [3], which is provided by Contrastive Lift [3], containing up to 500 objects in each scene. We present the quantitative results in Tab. 2, showing that our method (73.4%) achieves significant improvements in terms of the mean PQ^{scene} results compared to the current state-of-the-art methods of Contrastive Lift [3] (69.0%) and Panoptic Lifting [40] (63.2%). The overall performance highlights the advantages of our probabilistic approach over the previous deterministic methods, particularly in segmenting complex scenes with hundreds of objects. Further, our method achieves a mean SQ^{scene} of 82.2% and a mean RQ^{scene} of 86.9%, surpassing the SQ^{scene} of 77.7% and RQ^{scene} of 86.6% obtained by Contrastive Lift [3]. Moreover, the visual comparisons in Fig. 4 further indicate that our method can accurately segment the small objects in the red frames, whereas Contrastive Lift [3] struggles to distinguish such instances.

4.3 Ablation Study

Effectiveness of Each Component. We conduct an ablation study on the Messy Room dataset [3], which contains more challenging scenes with hundreds of instances. As shown in Tab. 3 (b) and (d), the proposed probabilistic feature embeddings greatly benefit our method, while replacing it with deterministic one leads to significant performance drop. Furthermore, the effectiveness of our proposed cross-view constraint is verified in Tab. 3 (e) and (f). For the proposed multi-view object association (MVOA) algorithm, the results in Tab. 3 (c), (d), (e), and (f) show that it particularly benefits the clustering of our proposed

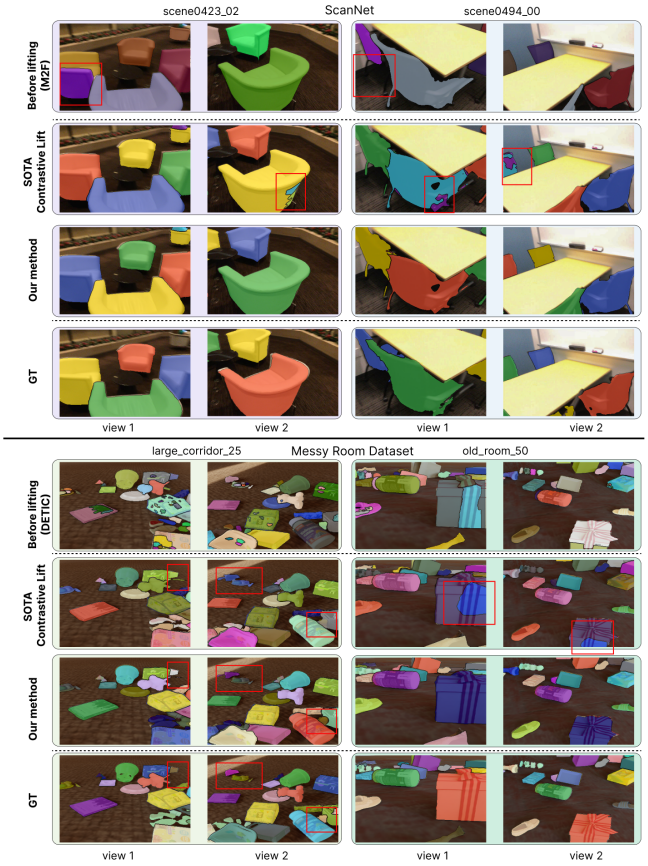

Fig. 4: Visual comparison of the latest state-of-the-art method Contrastive Lift [3] and our method on the ScanNet [12] dataset and the Messy Room [3] dataset.

Table 3: Ablation study on the Messy Room dataset [3]. The model (a) corresponds to Contrastive Lift [3] and the model (f) corresponds to our full method (PCF).

Model	Feature space	Clustering	SQ^{scene}(%)	RQ^{scene} (%)	PQ^{scene}(%)
(a)	Deterministic [3]	HDBSCAN [30]	77.7	86.6	69.0
(b)	Deterministic [3]	MVOA	79.3	86.2	70.4
(c)	Learned Gaussian distribution	HDBSCAN [30]	78.0	86.6	69.6
(d)	Learned Gaussian distribution	MVOA	81.3	86.8	72.3
(e)	Learned Gaussian distribution (+ Cross-view constraint)	HDBSCAN [30]	78.8	86.9	70.4
(f)	Learned Gaussian distribution (+ Cross-view constraint)	MVOA	82.2	86.9	73.4

probabilistic representation. Also, MVOA can be effectively employed to the deterministic method [3] for a performance boost, as Tab. 3 (a) and (b) show.

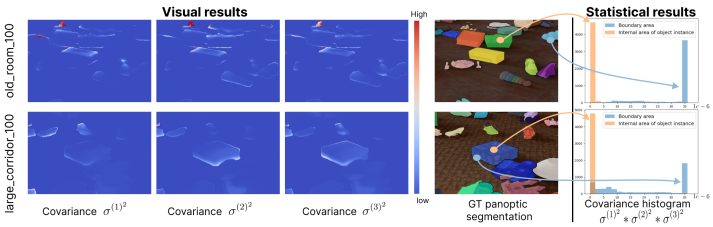

Fig. 5: The visualization results of learned covariance components and the statistical results of covariances in two scenes of the Messy Room dataset [3]. For the histograms, the horizontal axis denotes the range of covariance magnitudes, while the vertical axis corresponds to the frequency statistics for those magnitudes. We calculate and plot the covariance magnitudes ($\sigma^{(1)^2} * \sigma^{(2)^2} * \sigma^{(3)^2}$) of two distance image regions, the boundary areas and the internal areas of object instances, across all observed views.

Uncertainty Analysis. To verify whether our method could provide meaningful modeling for uncertainty, we provide the rendered uncertainty maps on two scenes of the Messy Room dataset [3] in Fig. 5. The figure clearly illustrates that the regions of high covariance are mainly located near the boundaries of the object instances, due to the *inconsistent segmentation* issue that leaves severe ambiguity around the instance boundaries. Moreover, we statistically analyze the learned covariances in two distinct image regions, namely the boundary areas and the internal areas of object instances, across all observed views. These results further demonstrate our method's ability to model high uncertainty in areas where the instance boundaries are ambiguous. We provide more details in the supplementary material.

4.4 Robustness Experiments

2D Backbones. In practice, the quality of panoptic segmentation generated by different 2D models may vary a lot. To study the robustness of our probabilistic method and the deterministic method [3] when incorporating with different 2D models, we select four different models from the model zoo of Detic [3] and utilize the LVIS [16] vocabulary to generate the 2D panoptic segmentation predictions for lifting, as shown in Fig. 6 (a). The quantitative results are presented in Fig. 6 (b), where our probabilistic methods consistently outperform the baseline method by a large margin. Particularly, we observe that the proposed MVOA algorithm can consistently boost the performance of a deterministic method [3].

Hand-crafted Noise. We study the robustness by adding hand-crafted noise. Specifically, for all given panoptic segmentation masks, we randomly select 100 pixels as anchors. Then, for each anchor, we randomly choose a pixel within a $W * W$ window centered on the anchor and assign its instance ID to all pixels within this window, to simulate inaccurate segmentation predictions around object boundaries. As shown in Fig. 6 (c), as W gradually increases, the instance

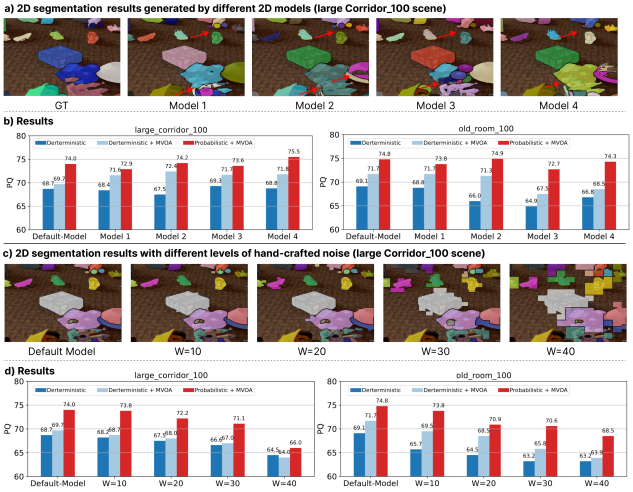

Fig. 6: Quantitative comparisons when using different 2D models or adding various levels of hand-crafted noise. We compare our probabilistic method ("Probabilistic + MVOA" in red) with the deterministic method [3] ("Deterministic" in dark blue) and a variant of deterministic method using our proposed MVOA algorithm ("Deterministic + MVOA" in light blue). Besides, the four tested models are from the model zoo of Detic [50][1], and the two tested scenes are from the Messy Room dataset [3].

boundaries are continuously blurred. Our probabilistic approach consistently achieves the best performance as presented in Fig. 6 (d) and the results demonstrate that the proposed MVOA algorithm benefits the deterministic method [3].

5 Conclusion

We present PCF-Lift for the panoptic lifting task. First, we propose to learn the probabilistic feature embeddings through a multivariate Gaussian distribution for instance representation. For training, we reformulate the contrastive loss with the Probability Product kernel and propose a novel cross-view constraint to enhance the feature consistency across different views. During the inference phase, we propose a novel multi-view object association algorithm to effectively identify the prototype features representing underlying 3D object instances. We verify the effectiveness and robustness of PCF-Lift by extensive experiments.

[1] We use the official pre-trained models provided in https://github.com/facebookresearch/Detic/tree/main. Please refer to our supplementary material for the details of the four models.

Acknowledgements

This work is supported by the InnoHK of the Government of Hong Kong via the Hong Kong Centre for Logistics Robotics, the Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, and the CUHK T Stone Robotics Institute.

References

1. Bae, G., Budvytis, I., Cipolla, R.: Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13137–13146 (2021)
2. 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)
3. Bhalgat, Y., Laina, I., Henriques, J.F., Zisserman, A., Vedaldi, A.: Contrastive Lift: 3D object instance segmentation by slow-fast contrastive fusion. arXiv preprint arXiv:2306.04633 (2023)
4. Cai, K., Lu, C.X., Huang, X.: Uncertainty estimation for 3D dense prediction via cross-point embeddings. IEEE Robotics and Automation Letters **8**(5), 2558–2565 (2023)
5. Cen, J., Zhou, Z., Fang, J., Shen, W., Xie, L., Jiang, D., Zhang, X., Tian, Q., et al.: Segment anything in 3D with NeRFs. Advances in Neural Information Processing Systems **36** (2024)
6. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision. pp. 333–350. Springer (2022)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12475–12485 (2020)
9. 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. pp. 1290–1299 (2022)
10. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
11. Dahnert, M., Hou, J., Nießner, M., Dai, A.: Panoptic 3D scene reconstruction from a single rgb image. Advances in Neural Information Processing Systems **34**, 8282–8293 (2021)
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 conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
13. Der Kiureghian, A., Ditlevsen, O.: Aleatory or epistemic? does it matter? Structural safety **31**(2), 105–112 (2009)

14. Fan, Z., Wang, P., Jiang, Y., Gong, X., Xu, D., Wang, Z.: NeRF-SOS: Any-iew self-supervised object segmentation on complex scenes. arXiv preprint arXiv:2209.08776 (2022)
15. Gasperini, S., Mahani, M.A.N., Marcos-Ramiro, A., Navab, N., Tombari, F.: Panoster: End-to-end panoptic segmentation of LiDAR point clouds. IEEE Robotics and Automation Letters **6**(2), 3216–3223 (2021)
16. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019)
17. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. arXiv preprint arXiv:2203.08414 (2022)
18. Jebara, T., Kondor, R., Howard, A.: Probability product kernels. The Journal of Machine Learning Research **5**, 819–844 (2004)
19. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH computer graphics **18**(3), 165–174 (1984)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
21. 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 (2023)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. 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)
24. 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. arXiv preprint arXiv:2304.02643 (2023)
25. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing NeRF for editing via feature field distillation. Advances in Neural Information Processing Systems **35**, 23311–23330 (2022)
26. 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)
27. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. IEEE Access **8**, 193907–193934 (2020)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
29. Liu, L., Gu, J., Zaw Lin, K., Chua, T.S., Theobalt, C.: Neural sparse voxel fields. Advances in Neural Information Processing Systems **33**, 15651–15663 (2020)
30. McInnes, L., Healy, J., Astels, S.: hdbscan: Hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

32. Milioto, A., Behley, J., McCool, C., Stachniss, C.: LiDAR panoptic segmentation for autonomous driving. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8505–8512. IEEE (2020)
33. Narita, G., Seno, T., Ishikawa, T., Kaji, Y.: Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4205–4212. IEEE (2019)
34. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. Advances in neural information processing systems **30** (2017)
35. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J.: Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 55–64 (2020)
36. Park, J., Lee, J., Kim, I.J., Sohn, K.: Probabilistic representations for video contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14711–14721 (2022)
37. Porzi, L., Bulo, S.R., Colovic, A., Kontschieder, P.: Seamless scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8277–8286 (2019)
38. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning. pp. 1278–1286. PMLR (2014)
39. Rosinol, A., Gupta, A., Abate, M., Shi, J., Carlone, L.: 3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans. arXiv preprint arXiv:2002.06289 (2020)
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. pp. 9043–9052 (2023)
41. Sirohi, K., Mohan, R., Büscher, D., Burgard, W., Valada, A.: Efficientlps: Efficient LiDAR panoptic segmentation. IEEE Transactions on Robotics **38**(3), 1894–1914 (2021)
42. Tang, W., Biqi, Y., Li, X., Liu, Y.H., Heng, P.A., Fu, C.W.: Prototypical variational autoencoder for 3D few-shot object detection. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
43. Tschernezki, V., Laina, I., Larlus, D., Vedaldi, A.: Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In: 2022 International Conference on 3D Vision (3DV). pp. 443–453. IEEE (2022)
44. Wang, B., Chen, L., Yang, B.: Dm-NeRF: 3D scene geometry decomposition and manipulation from 2D images. arXiv preprint arXiv:2208.07227 (2022)
45. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O.: Pointcontrast: Unsupervised pre-training for 3D point cloud understanding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 574–591. Springer (2020)
46. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3D scenes. arXiv preprint arXiv:2312.00732 (2023)
47. Zhang, W., Pang, J., Chen, K., Loy, C.C.: K-net: Towards unified image segmentation. Advances in Neural Information Processing Systems **34**, 10326–10338 (2021)
48. Zhang, X., Chen, Z., Wei, F., Tu, Z.: Uni-3D: A universal model for panoptic 3D scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9256–9266 (2023)

49. 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. pp. 15838–15847 (2021)
50. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: European Conference on Computer Vision. pp. 350–368. Springer (2022)
51. Zhou, Z., Zhang, Y., Foroosh, H.: Panoptic-polarnet: Proposal-free LiDAR point cloud panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13194–13203 (2021)