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

In this supplementary material, we provide more experimental results (Sec. A), experimental details (Sec. B), implementation details (Sec. C), and limitation & future work (Sec. D) about our proposed PCF-Lift method.

A Experimental Results

A.1 Experiments on More Indoor Datasets

We conduct experiments on more indoor datasets (*i.e.*, Replica [12] and HyperSim [10]). Specifically, we follow the instructions in Panoptic Lifting's official GitHub page[1]to process the datasets for fair comparisons. Tab. 1 shows that our method consistently and significantly outperforms the two SOTA methods, and the visual comparisons in Fig. 1 further indicate our accurate segmentation results.

Table 1: Quantitative comparisons on the Replica and HyperSim datasets. Results of prior works are sourced from their papers, and SQ^{scene} and RQ^{scene} are not reported in the Contrastive Lift paper.

Dataset	Method	SQ^{scene}(%) ↑	RQ^{scene}(%) ↑	**PQ^{scene}(%)** ↑
Replica	Panoptic Lifting [11]	69.1	63.6	57.9
Replica	Contrastive Lift [3]	-	-	59.1
Replica	Ours	**73.4**	**64.6**	**62.0**
HyperSim	Panoptic Lifting [11]	70.4	64.3	60.1
HyperSim	Contrastive Lift [3]	-	-	62.3
HyperSim	Ours	**79.4**	**64.5**	**63.8**

A.2 More Visual Results

We provide more visual results in Fig. 2, Fig. 3, and Fig. 4. The results further verify that our proposed PCF-Lift could generate accurate and consistent segmentation results.

[1] https://github.com/nihalsid/panoptic-lifting/tree/main.

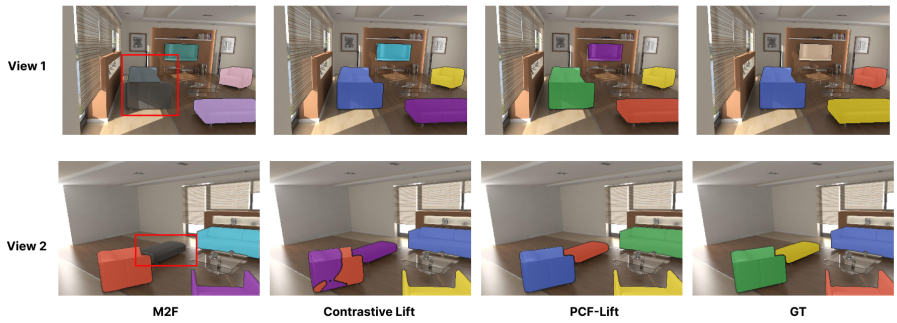

Fig. 1: Visual comparison on HyperSim dataset [10].

A.3 Ablation Study

Cross-view Constraint. To further demonstrate the effectiveness of the proposed cross-view constraint, we present visual comparisons in Fig. 5 (a). The results verify that the proposed cross-view constraint further improves the quality of learned feature space, while the artifacts are significantly reduced. In our main experiments, we apply a predefined threshold of $\tau = 0.9$ to define the cross-view positive pairs. To further explore the impact of this threshold value, we conduct additional experiments using two different values (*i.e.*, $\tau = 0.85$ and 0.95). As shown in Tab. 2, the results are practically stable given a moderate change of threshold τ.

Table 2: Quantitative comparisons of different threshold τ values for the cross-view constraint.

τ	0.85	0.90 (default)	0.95
SQ^{scene}	82.3	82.2	82.2
RQ^{scene}	86.3	86.9	86.5
$\mathbf{PQ^{scene}}$	73.1	73.4	73.2

Multi-view Object Association Algorithm. To validate the effectiveness of the Multi-view object association (MVOA) algorithm, we present visual comparisons in Figure 5 (b). The results illustrate that MVOA achieves superior segmentation accuracy by effectively identifying the prototype of the underlying 3D object given probabilistic feature similarities. In contrast, substituting MVOA with the HDBSCAN [9] algorithm results in further challenges, such as the misidentification of small objects and the unexpected generation of artifacts.

Furthermore, we explore the impact of varying similarity threshold $\mathcal{T}$ values on the MVOA algorithm performance. In practice, the selection of hyper-parameter $\mathcal{T}$ is based on the average similarity computed across grouped feature pairs identified within each view in our experiments. We conduct additional experiments by applying perturbations of -0.05 and $+0.05$ to the default $\mathcal{T}$ value. The experiment results in Tab. 3 indicate that the MVOA algorithm achieves stable performance within a reasonable range of threshold values.

Table 3: Quantitative comparisons of different similarity threshold $\mathcal{T}$ values used in the multi-view object association algorithm.

Perturbation	-0.05	0.0 (default)	$+0.05$
SQ^{scene}	81.7	82.2	82.1
RQ^{scene}	86.9	86.9	86.5
$\mathbf{PQ^{scene}}$	72.8	73.4	73.1

Concentrate Loss in Main Paper Eq.3. We employ the concentrate loss, based on the 2D masks in each view, to encourage features of the same instance to maximally converge towards a similarity of one. This facilitates the subsequent probabilistic clustering process. Further, we reformulate the calculation of concentration loss using the PP kernel, incorporating it as a component of the probabilistic contrastive loss in Eq. (4) in main paper. We perform an additional ablation study on this loss, showing that the average $\{\mathbf{PQ^{scene}(\%)}, \text{SQ}^{\text{scene}}(\%), \text{RQ}^{\text{scene}}(\%)\}$ drops from $\{\mathbf{73.4}, 82.2, 86.9\}$ to $\{\mathbf{70.3}, 78.7, 86.7\}$ on the Messy Room dataset [3], if we drop this loss term.

Quality of Semantic Field. In our main paper, we utilize scene-level Panoptic Quality metric to assess the quality and consistency of the rendered panoptic maps. Note that, the semantic field is adopted from the previous works (*i.e.*, Panoptic Lifting [11] and Contrastive Lift [3]) to ensure a fair comparison. For completeness, we include an additional quantitative comparison of the rendered semantic maps in Tab. 4. Following the previous methods, we report the mean Intersection over Union (mIoU) metric. As Tab. 4 shows, the semantic performances of different methods are practically similar as expected. This fact further proves that the notable improvements in scene-level PQ mainly result from the enhanced quality of the learned instance fields, attributed to our proposed probabilistic design.

A.4 3D Representation Choice

For fair comparisons, we follow the prior baselines (Panoptic Lifting [11] & Contrastive Lift [3]) to adopt the TensoRF [4] representation. Yet, our method is

Table 4: Quantitative comparisons of rendered semantic maps. We report the mean Intersection over Union (mIoU) metric. As expected, the performance is almost the same, further indicating that the improved final performance in PQ^{scene} is mainly influenced by the higher quality of the learned instance fields with the proposed probabilistic design.

Method	Messy Rooms dataset [3]		ScanNet [5]	
	PQ^{scene}	mIoU	PQ^{scene}	mIoU
Panoptic Lifting [11]	63.2	91.8	58.9	65.2
Contrastive Lift [3]	69.0	91.8	62.0	65.2
Ours	**73.4**	91.7	**63.5**	65.2

Table 5: Quantitative comparisons (PQ^{scene}) using different 3D representations on the Messy Room dataset [3].

Scene ID	TensoRF [4]		ZipNeRF [2]	
	Contrastive Lift [3]	Ours	Contrastive Lift [3]	Ours
large_corridor_25	76.5	**81.0**	77.3* (*re-implemented*)	**80.7**
old_room_25	78.9	**80.9**	79.0* (*re-implemented*)	**81.7**

agnostic to the 3D representation choice, since the proposed probabilistic feature branch is separate from the other branches (*i.e.*, color and density) for 3D reconstruction representations. Further, we conducted an experiment on two scenes in the Messy Room Dataset using the ZipNeRF [2] representation. From Tab. 5, we can see that our probabilistic method consistently outperforms the deterministic method.

A.5 Efficiency

Table 6: Comparisons of the training speed in iterations per second. Overall, our PCF-Lift significantly enhances performance and achieves comparable efficiency, compared to baselines [3, 11].

Method	Ours	Contrastive Lift [3]	Panoptic Lifting [11]
Training speed	21.74	23.16	21.87

We compare the training speed in iterations per second, measured on an NVIDIA 3090 RTX GPU. Overall, our PCF-Lift significantly enhances performance and achieves comparable efficiency, compared to the previous methods, as shown in Tab 6. Besides, the panoptic segmentation rendering efficiency of our

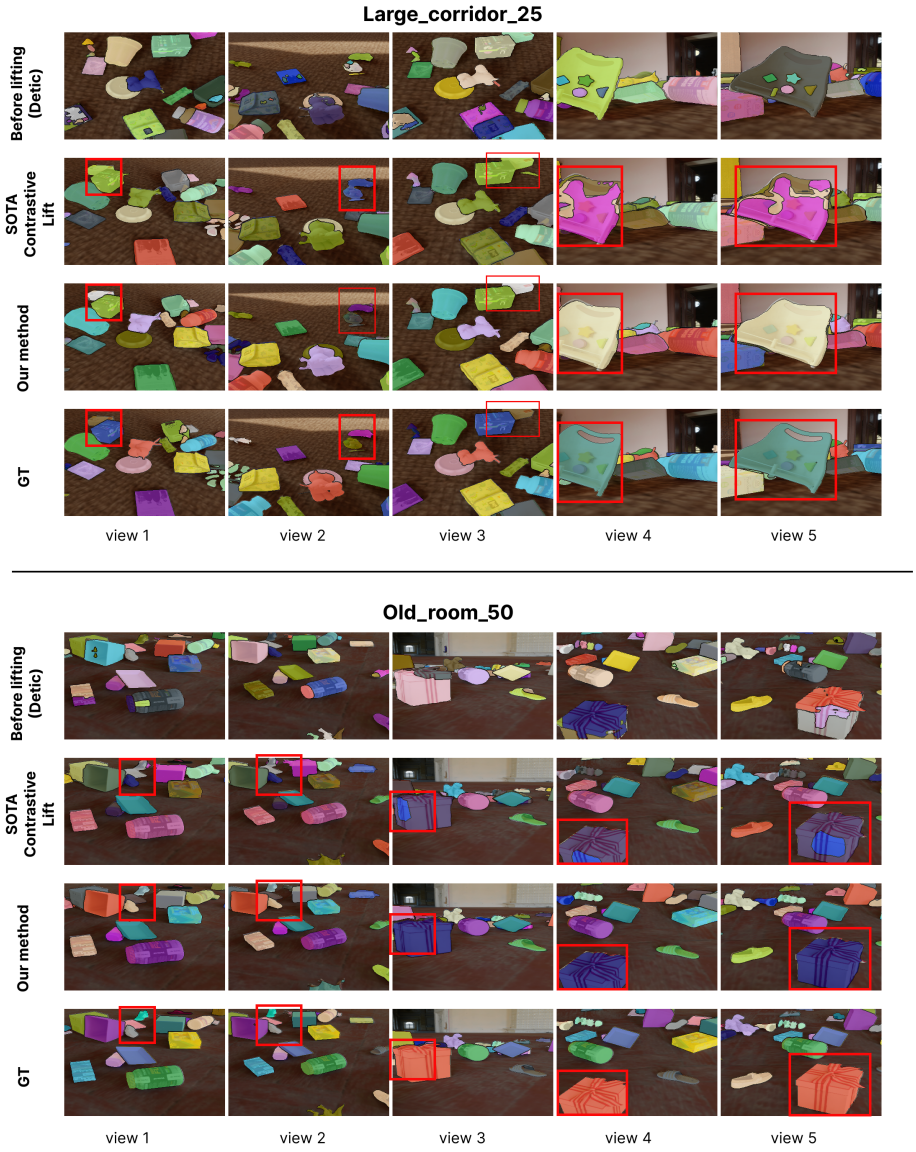

Fig. 2: Visual comparisons on the Messy Room dataset [3].

PCF-Lift is almost identical to Contrastive Lift (0.457 vs. 0.456 sec. per image), tested on an NVIDIA 3090 RTX GPU using the Messy Room Dataset.

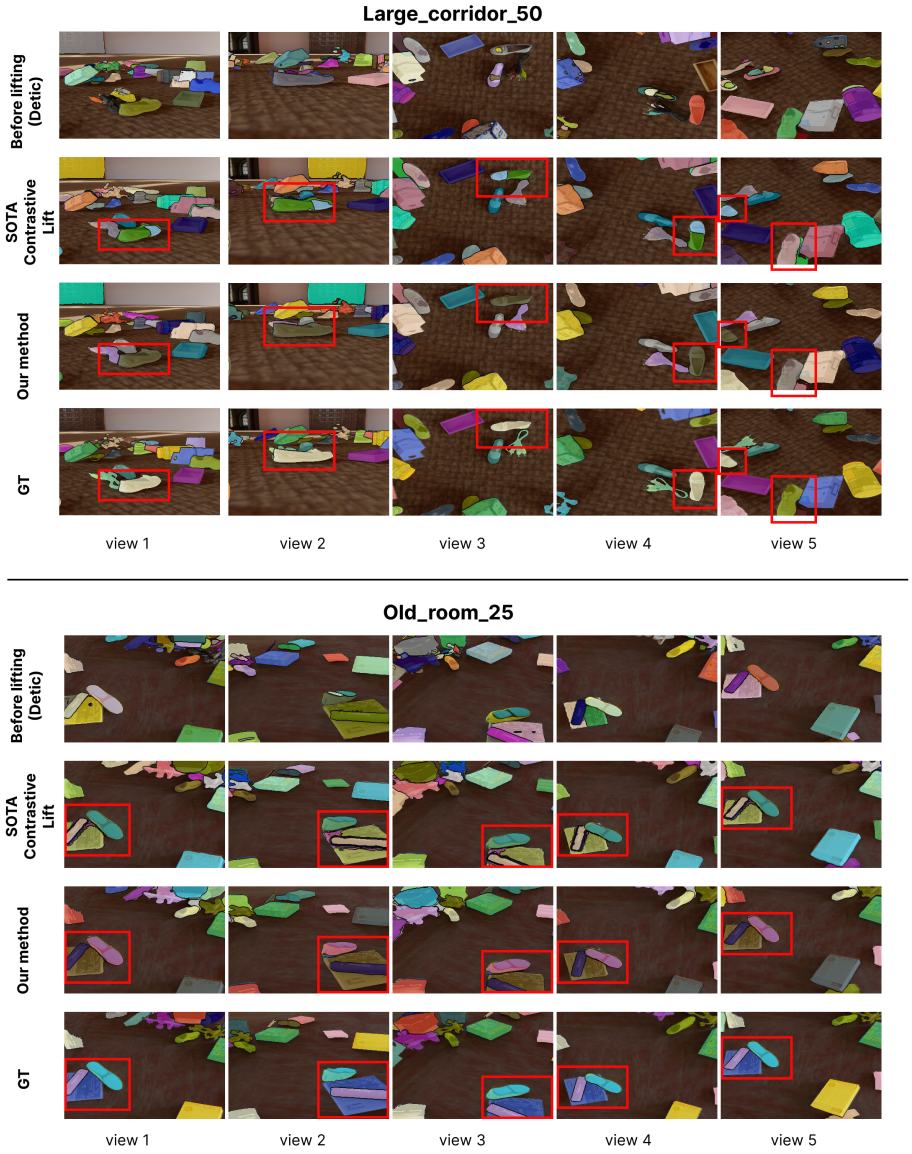

Fig. 3: Visual comparisons on the Messy Room dataset [3].

A.6 Robustness Experiments

For robustness experiments, we present more visual results in Fig. 6. The results verify PCF-Lift's capability to maintain performance consistency under various conditions and showcase its robustness.

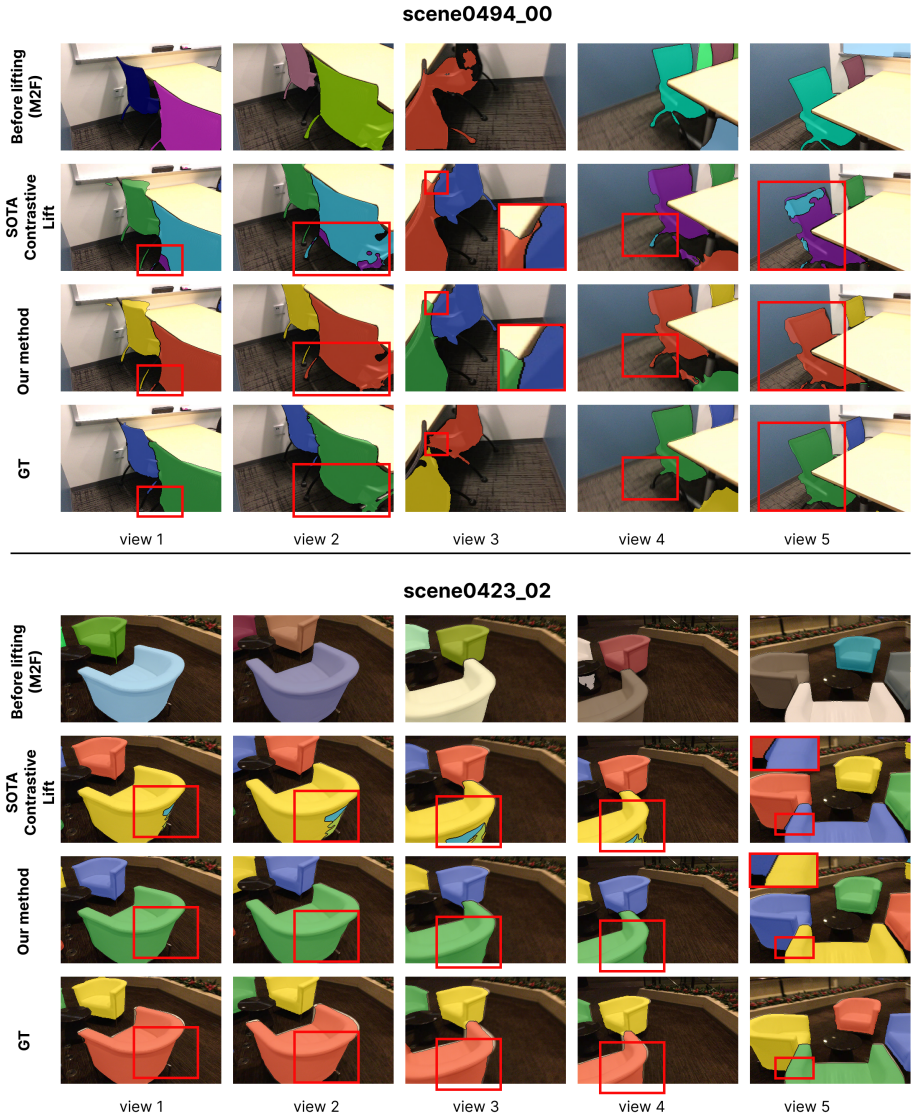

Fig. 4: Visual comparisons on ScanNet dataset [5].

B Experimental details

Uncertainty Analysis. To verify whether our method could provide meaningful modeling for uncertainty, we conduct a statistical analysis on the learned covariances within two distinct regions of images: the boundary areas and the internal areas of object instances, across all observed views. For identifying boundary areas, we apply the Canny edge detection algorithm followed by a dilation operation on the ground-truth segmentation map for each view. The internal ar-

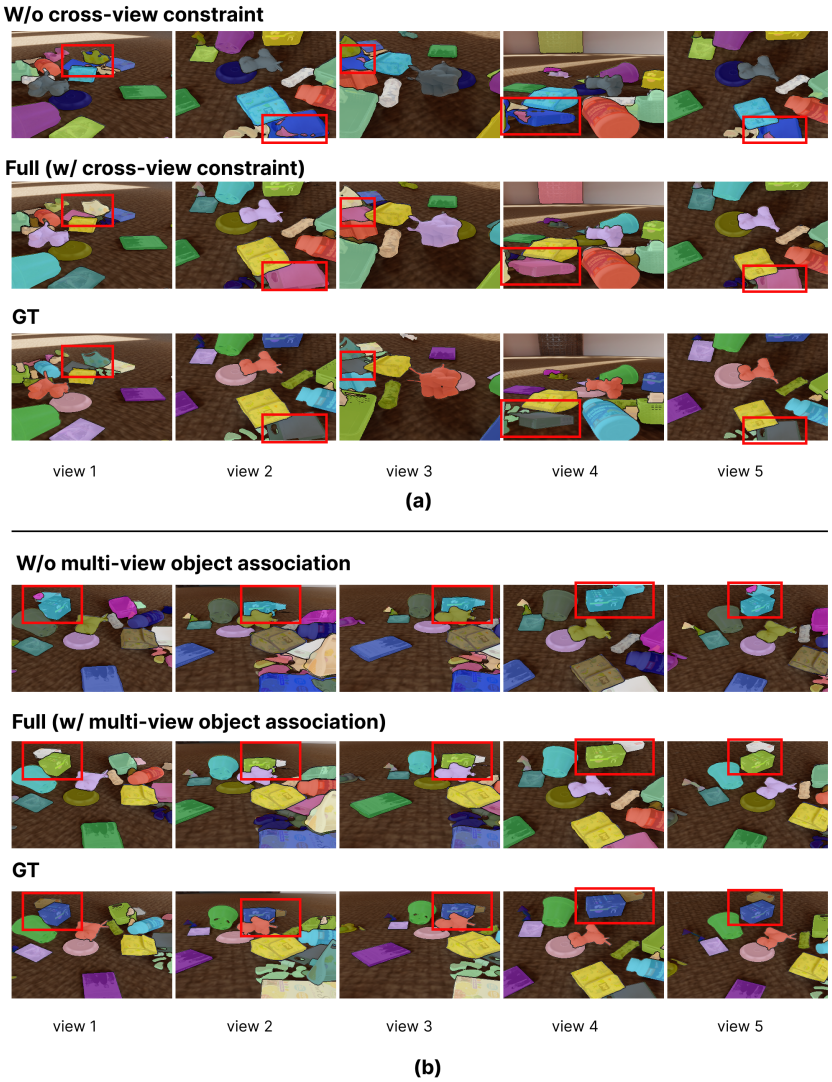

Fig. 5: Visual comparisons for ablation study.

eas of object instances are identified as the remaining foreground pixels outside the boundary areas. We then select the top-10 (K=10) covariance samples from both the boundary and internal areas for each view and plot their histograms separately to analyze the variance patterns.

Selected Models in Robustness Experiments. To study the robustness of our probabilistic method when incorporating different 2D models, we choose the

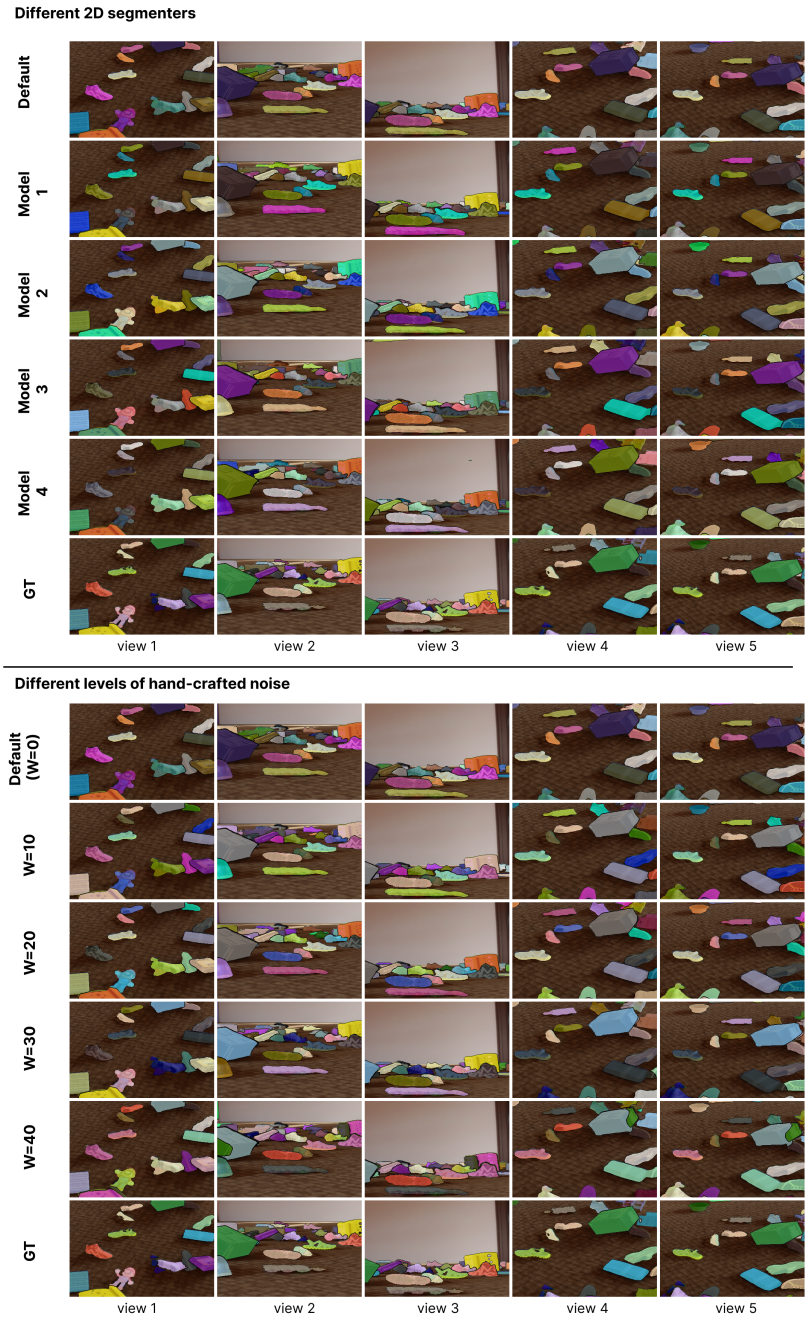

Fig. 6: Visual comparisons of PCF-Lift's results using different 2D segmentation models and incorporating different levels of hand-crafted noise.

4 official models with LVIS [6] vocabulary in Detic [13] to conduct the experiments. Each model differs in terms of its architectural backbone and the type of supervision applied during training. Specifically, Model 1 employs a ResNet50 backbone with box supervision, Model 2 is built on a Swin-B backbone also using box supervision, Model 3 utilizes a ResNet50 backbone but with Detic supervision, and Model 4 is based on the Swin-B backbone, incorporating Detic supervision. This diverse selection allows us to comprehensively assess the performance impact of different 2D models on our method's robustness. The pretrained models can be downloaded from https://github.com/facebookresearch/Detic.

C Implementation Details

We ensure a fair comparison by utilizing a similar architecture, specifically TensoRF [4], as used in recent works [3, 11]. Our method initializes the grid of color and density with a resolution of 128^3 and progressively increases this resolution to 192^3 till the end of the training process. For the semantic field predictions, our approach utilizes a five-layer Multilayer Perceptron (MLP) with 256 hidden units. To represent instance information, we use a five-layer MLP with 256 hidden units to predict two crucial components: a 3-dimensional vector representing the Gaussian mean and an additional 3-dimensional vector for the diagonal Gaussian covariance in both slow and fast instance fields. Practically, the MLPs for semantic and instance fields do not incorporate position encoding, while the semantic and instance predictions are directly generated from the input 3D positions. To ensure that the Gaussian covariance values are strictly positive, we employ the following activation function $g(x)$:

$$g(x) = \begin{cases} x + 1, & \text{if } x > 0 \\ \exp(x), & \text{if } x \leq 0 \end{cases}.$$

In our experiments, we train our neural fields for 400k iterations using a learning rate of 5×10^{-4} for all MLPs, a learning rate of 0.01 for the grids, and a batch size of 2048 on all scenes unless otherwise stated. Following baseline works [3, 11], our PCF-Lift employs a multi-stage strategy to ensure a stable and effective model training. In the initial 40K iterations, we only use the RGB reconstruction loss to train the model. This initial step is critical, allowing for a reasonable quality of the density field to support the rendering of instance and semantic fields. After the initial 40K iterations, we introduce the semantic segmentation loss, *i.e.*, cross-entropy loss, to train the semantic field. After 160k iterations, we add the instance loss term to the overall training loss. Besides, we use the additional segment consistency loss (proposed in Panoptic Lift [11]) to train the semantic field from 280K iteration. Throughout the training process, we balance the contributions of the RGB loss, semantic segmentation loss, segmentation consistency loss, and instance loss with weights of 1.0, 0.1, 0.1, and 1.0, respectively. To further enhance the stability of the training, the gradients propagated from the semantic and instance field-related losses are particularly

blocked from influencing the density field. During the inference, we utilize the proposed multi-view object association (MVOA) to extract the prototype features for generating the instance segmentation results. Despite the effectiveness of the MVOA algorithm in identifying relevant prototype features, we observe that applying a score threshold to filter out prototypes with low confidences can further enhance the overall performance. To determine the optimal score threshold, we undertake a hyper-parameter sweep using a partial training set following the common practice used in [3].

D Limitation & Future work

To fairly compare with the state-of-the-art works, we also adopt the TensorRF [4] as the basis of our PCF-Lifting method, while leading to relatively slow training time. To achieve higher efficiency, we will consider investigating the latest 3D reconstruction technique, such as 3D Gaussian Splatting [7], in the future work. Moreover, the current experiments focus on indoor scenes, leaving the challenging task of outdoor scene understanding as an interesting area for future research. To provide a preliminary insight, we present one outdoor scene from the Mip-NeRF 360 dataset [1], as shown in Fig. 7. A more comprehensive exploration of outdoor scenes will be addressed in our future work.

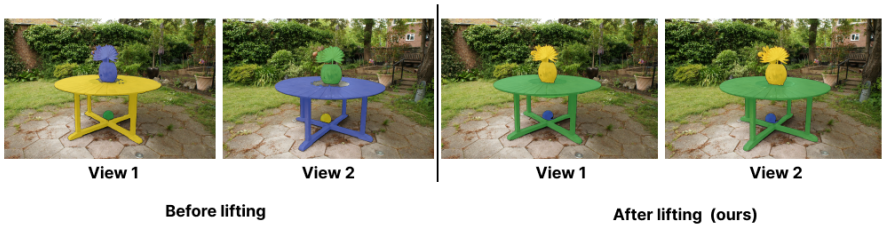

Fig. 7: Lifting an outdoor scene of the Mip-NeRF 360 dataset [1]. Note that, similar to the processing of the Messy Room dataset [8], we categorize all object types into a single "foreground" class.

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)
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)

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. 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)
5. 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)
6. 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)
7. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
8. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. IEEE Access **8**, 193907–193934 (2020)
9. McInnes, L., Healy, J., Astels, S.: hdbscan: Hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
10. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021)
11. 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)
12. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
13. 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)