# Supplementary Materials for "MonoInstance: Enhancing Monocular Priors via Multi-view Instance Alignment for Neural Rendering and Reconstruction"

Wenyuan Zhang[1], Yixiao Yang[1], Han Huang[1], Liang Han[1], Kanle Shi[2],
Yu-Shen Liu[1*], Zhizhong Han[3]
School of Software, Tsinghua University, Beijing, China[1]
Kuaishou Technology, Beijing, China[2]
Department of Computer Science, Wayne State University, Detroit, USA[3]
{zhangwen21,yangyixi21,h-huang20,hanl23}@mails.tsinghua.edu.cn
shikanle@kuaishou.com, liuyushen@tsinghua.edu.cn, h312h@wayne.edu

## 1. Overview

In the main paper, we propose MonoInstance, which is a general approach that explores the uncertainty of monocular depths to provide enhanced geometric priors for neural rendering. Our approach can be applied upon different multi-view neural rendering and reconstruction methods to enhance the monocular priors for better neural representation learning. This supplementary material provides implementation details, additional ablation studies, discussions and additional visualization results of reconstruction and novel view synthesis. All the sections are organized as follows:

- Section 2 provides implementation details of multi-view consistent segmentation, background identification, selection of nearby views and the silhouette loss in 3D Gaussians.
- Section 3 provides additional ablation studies on the weight of instance mask constraint and the usage of normal priors.
- Section 4 discusses the limitations and future works of our method.
- Section 5 provides additional visual comparisons on various benchmarks.

## 2. Implementation Details

**Multi-view consistent segmentation.** Inspired by MaskClustering [20], we utilize a graph-based clustering algorithm to achieve multi-view consistent instance segmentation. Specifically, we first obtain instance segmentation on each image using [13]. And then we use Chamfer Distance (CD) of the point clouds from back-projected instance depths, to measure the similarity between instances from different views. We initialize each instance from all views as a graph node, and consider the CD between two instances as the edge weight between two nodes. Iterative graph clustering [15] is then applied to partition the graph nodes into instance clusters, thereby achieving multi-view consistent instance segmentation. The key difference between MaskClustering and our implementation lies in the use of monocular depths, rather than ground truth depths, for back-projecting instances into world coordinate 3D space. The monocular depths are pre-aligned with the rendered depths. Although the monocular depths may not be 100% accurate, we find that the similarity calculation based on CD is robust to handle noise in the point clouds and does not significantly affect the segmentation results. The visualization of segmentation on individual images and the multi-view consistent segmentation results are shown in Fig. 1, which indicates our ability of accurately segmenting instances of varying scales within the scene. We also visualize two segmented instance masks in Fig. 2 to demonstrate the effectiveness of our multi-view consistent segmentation algorithm.

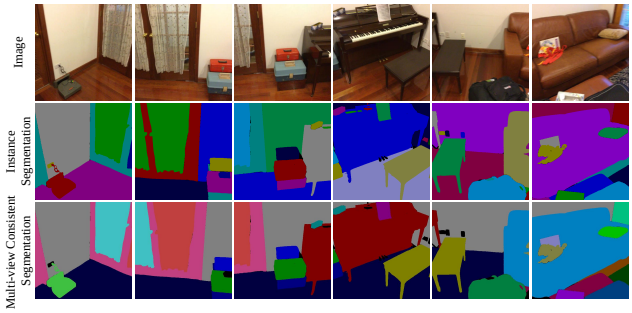


Figure 1. Visualization of instance segmentation. The second row is the raw segmentation on individual images, while the third row is the segmentation after applying multi-view consistent segmentation algorithm.

**Background segmentation.** Existing methods [12, 17, 18] generally assume that the background areas of indoor scenes,

*The corresponding author is Yu-Shen Liu.


Figure 2. Visualization of the instance masks and the corresponding images.

such as walls and floors, are reliable. This is because the monocular priors are generated by a pre-trained neural network, which has a strong smooth inductive bias. Furthermore, smooth regions exhibit similar and simple feature patterns, enabling the network to more effectively predict monocular cues. Therefore, we propose to directly apply monocular priors to the background areas of indoor scenes, while estimating uncertainty for objects within the scene. To this end, we utilize state-of-the-art semantic segmentation models [6, 14] to segment the background areas in indoor scenes. Specifically, for each image, we first use Grounding DINO [8] with a pre-defined prompt to obtain bounding boxes for the background areas. The prompt is defined as "room background, ceiling, floor." We then use SAM [6] to obtain masks for the areas within the bounding boxes. Subsequently, for multi-view consistent instance segmentation, we identify those instances that fall within the background masks in most frames, and set the uncertainty of these instances to zero.

**Selection of nearby views.** For high-uncertainty areas ($U(u,v) > 0.8$ in our implementation) where monocular priors are unreliable, we aim to mine more reliable photometric consistency as a remedy. To this end, we warp sampling points on the ray emitted from reference view to the instance mask in the nearby view, and accumulate the filtered projected points to obtain the final color as additional constraint. Selecting appropriate nearby views for the reference view is vital in this process. Fortunately, the multi-view consistent instance segmentation provides us with strong prior information to achieve this, as illustrated in Fig. 3. Specifically, for an instance mask area $S_r^i$ with label $i$ in the reference view $I_r$, we emit a ray $O_r$ towards the center pixel of the bounding box (BBX) of the instance mask area $S_r^i$. For all other views $\{I_j\}_{j=1}^N$ which contain the same instance $i$, we then emit rays $\{O_j\}_{j=1}^N$ towards the center pixel of the BBX of the mask area $S_j^i$ in each view. We then calculate the distances $\{\delta_j\}_{j=1}^N$ and the angles $\{\theta_j\}_{j=1}^N$ between $O_r$ and all $\{O_j\}_{j=1}^N$. The appropriate nearby views are selected based on whether the distances and angles are within specified thresholds,

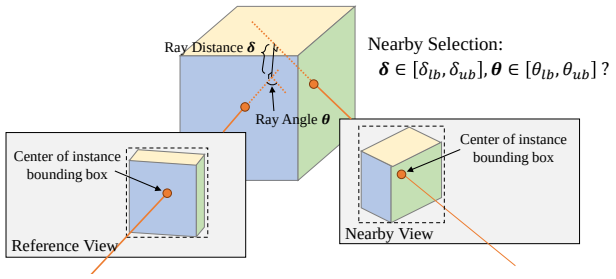


Figure 3. Illustration of selecting nearby views.

$$\Omega(I_j) = \begin{cases} 1 & \delta_j \in [\delta_{lb}, \delta_{ub}] \wedge \theta_j \in [\theta_{lb}, \theta_{ub}] \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

In our practice, we set $\delta_{lb} = 10°, \delta_{ub} = 60°, \theta_{lb} = 0, \theta_{ub} = 0.04$. Comparing to existing methods which rely on visibility or image features [2, 3, 9], our strategy provides a more robust and effective solution for exploring photometric consistency cues in multi-view instances.

**$\mathcal{L}_{sil}$ in 3D Gaussians.** Since 3D Gaussians can be directly splatted onto the image plane without any sampled points in 3D space, we design a variant of our instance mask constraint (Sec. 3.3 in the original paper), encouraging the projected instance depth points on the nearby view to move towards the instance mask in the nearby view, as illustrated in Fig. 4. Specifically, for a pixel $(u, v)$ in the instance mask area $S_r^i$ with high uncertainty, we back-project it into 3D space using the rendered depth, and then project it onto the nearby view$\{I_j\}_{j=1}^N$. The loss function is defined as the shortest distance from the projected point to the instance mask silhouette $\partial S_j^i$,

$$\mathcal{L}_{sil}(u,v) = \begin{cases} 0 & \pi_j(p(u,v)) \in S_j^i \\ d(\pi_j(p(u,v)), \partial S_j^i) & \pi_j(p(u,v)) \notin S_j^i \end{cases}, \quad (2)$$

where $p(u,v)$ is the back-projected point of pixel $(u,v)$ and $\pi_j(p(u,v))$ is the projected pixel of $p(u,v)$ on the nearby view $I_j$. $d(\cdot,\cdot)$ denotes the Euclidean distance between two pixels on the image.

## 3. Ablations

**Weight of instance mask constraint term.** We adaptively set hyperparameters according to scene specifications such as size, which works well without a need of manual tuning in different datasets. For example, in the confidence computation step, the query ball radius is linearly related to the bounding box size of the fused point cloud, as depicted in

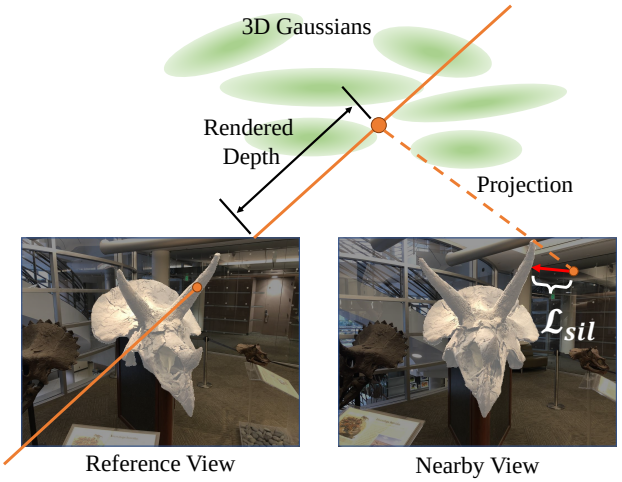


Figure 4. Illustration of instance mask constraint in 3D Gaussians.


Figure 5. Effect of the mask constraint weight $\lambda_2$ in Eq. (10).

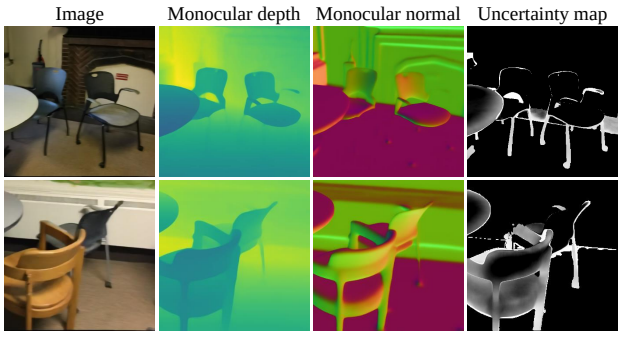


Figure 6. Visualization of monocular depth, monocular normal and our estimated uncertainty map. Monocular depths and normals tend to produce similar patterns on the same structures, therefore the uncertainty map can be universally applied to both depth and normal priors.

Table 1. Ablation study of adaptive depth and normal prior loss on all scenes of ScanNet dataset.

| Adaptive Depth | Adaptive Normal | CD↓ | F-score↑ |
|---|---|---|---|
| | | 0.041 | 0.749 |
| ✓ | | 0.038 | 0.763 |
| | ✓ | 0.039 | 0.772 |
| ✓ | ✓ | **0.037** | **0.786** |

Eq. (6) in the original paper. As for the loss terms, the values of $\lambda_1, \lambda_3, \lambda_4$ are consistent with baseline. Here we evaluate the different choice of the weight $\lambda_2$ in Eq. (10) in the original paper, as illustrated in Fig. 5. When $\lambda_2$ is small, the optimization relies solely on photometric loss, which is insufficient to recover geometric details such as chair legs and the lamp on the piano. When $\lambda_2$ is too large, the overfitting of colors from nearby views misleads the geometry, such as the piano legs.

**Normal priors.** Although we use monocular depth to measure the uncertainty, the evaluated uncertainty maps can also be applied to adaptive normal prior loss. This is because depth and normal priors are homologous dual outputs from the same foundational model, thus exhibiting similar patterns on the same structures. Fig. 6 provides an illustration, where the monocular depth map and normal map show similar degeneration on the thin structures, such as the chair legs. We further conduct an ablation study to demonstrate the effectiveness of the uncertainty map on depth and normal priors, as reported in Tab. 1. "Adaptive Depth" denotes employing the uncertainty maps as weights on the depth loss, and similarly for "Adaptive Normal". Numerical comparisons indicate that our uncertainty maps can improve the quality of both depth and normal priors.

We notice that the latest multi-view stereo method, DUSt3R [19], can estimate consistent depth maps and confidence maps for each depth from a given set of unconstrained images using data-driven priors. However, DUSt3R can not directly predict uncertainty according to multi-view depth consistency, which is merely based on single views. This limits the uncertainty estimation on unseen scenes in which data-driven prior may not generalize well.

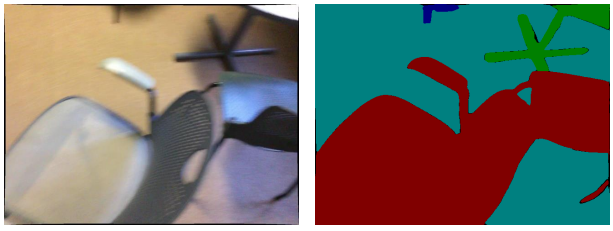


Figure 7. Failure case. The instance segmentation erroneously segment the two chairs into one instance.

## 4. Discussion

**Limitations.** We utilize multi-view consistent instance segmentation to evaluate the uncertainty maps, hence the performance of uncertainty maps might be impacted by the accuracy of segmentation. Existing instance segmentation algorithms sometimes suffer from issues such as part missing and instance merging [7, 11]. As shown in Fig. 7, the two

chairs are erroneously segmented into one instance, which was identified by the multi-view consistent instance segmentation algorithm. In such cases, the affected instance in this frame is excluded in the multi-view segmentation results, and the uncertainty of this instance is set as 0.5. However, this failure case does not significantly harm the final results, because it merely weakens the constraints of the monocular priors and the multi-view photometric consistency.

**Future works.** One of the future works is to explore the optimization of instance segmentation along with the training process, to correct the erroneous in instance segmentation, similar as ManhattanSDF [4]. Another future work is to explore the integration of powerful generative models with multi-view neural rendering frameworks to improve the reconstruction in areas with little viewpoints coverage.

# 5. Results

## 5.1. Comparisons

We provide additional visual comparisons across various multi-view neural rendering benchmarks. Fig. 8 displays the comparisons in dense-view reconstruction task on ScanNet [1] and Replica [16] datasets. Fig. 9 provides comparisons in sparse-view reconstruction task on DTU [5] dataset, and Fig. 10 shows comparisons in sparse novel view synthesis task on LLFF [10] dataset. The visual results further demonstrate the versatility and superiority of our method across different neural rendering tasks and datasets.

## 5.2. Video Display

We made a video in the supplementary materials to provide additional examples of the results of 3D reconstruction and novel view synthesis. In the first part of the video, we circle around the indoor scenes from the dense-view reconstruction experiment on "scene0616_00" in ScanNet [1] dataset and "office0" in Replica [16] dataset, respectively. In the second part, we showcase the objects from the sparse-view reconstruction experiment on "scan40" and "scan105" in DTU [5] dataset. In the third part, we record surround RGB and depth videos from sparse-input novel view synthesis on "leaves" and "room" in LLFF [10] dataset. The results showcase the superiority performance of our method across different tasks, demonstrating the effectiveness and universality of our proposed MonoInstance. Please refer to the the video for more details.

## 5.3. Training Time

We report the training time of each module in our method, as shown in Tab. 2, which demonstrates that our algorithm does not introduce significant additional overhead. Since our instance segmentation and uncertainty estimation processes require aggregating information from all views, the computational cost is related to the number of viewpoints. Furthermore, since the uncertainty map provides strong knowledge for multi-view geometry inference, we achieve faster convergence speed than MonoSDF.

# References


[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 4

[2] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6260–6269, 2022. 2

[3] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-NeuS: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 2

[4] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4

[5] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014. 4

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[7] Xiangtai Li, Henghui Ding, Haobo Yuan, Wenwei Zhang, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *European Conference on Computer Vision*, 2024. 2

[9] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. SparseNeuS: Fast Generalizable Neural Surface Reconstruction from Sparse Views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022. 2

[10] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. 4

[11] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions*

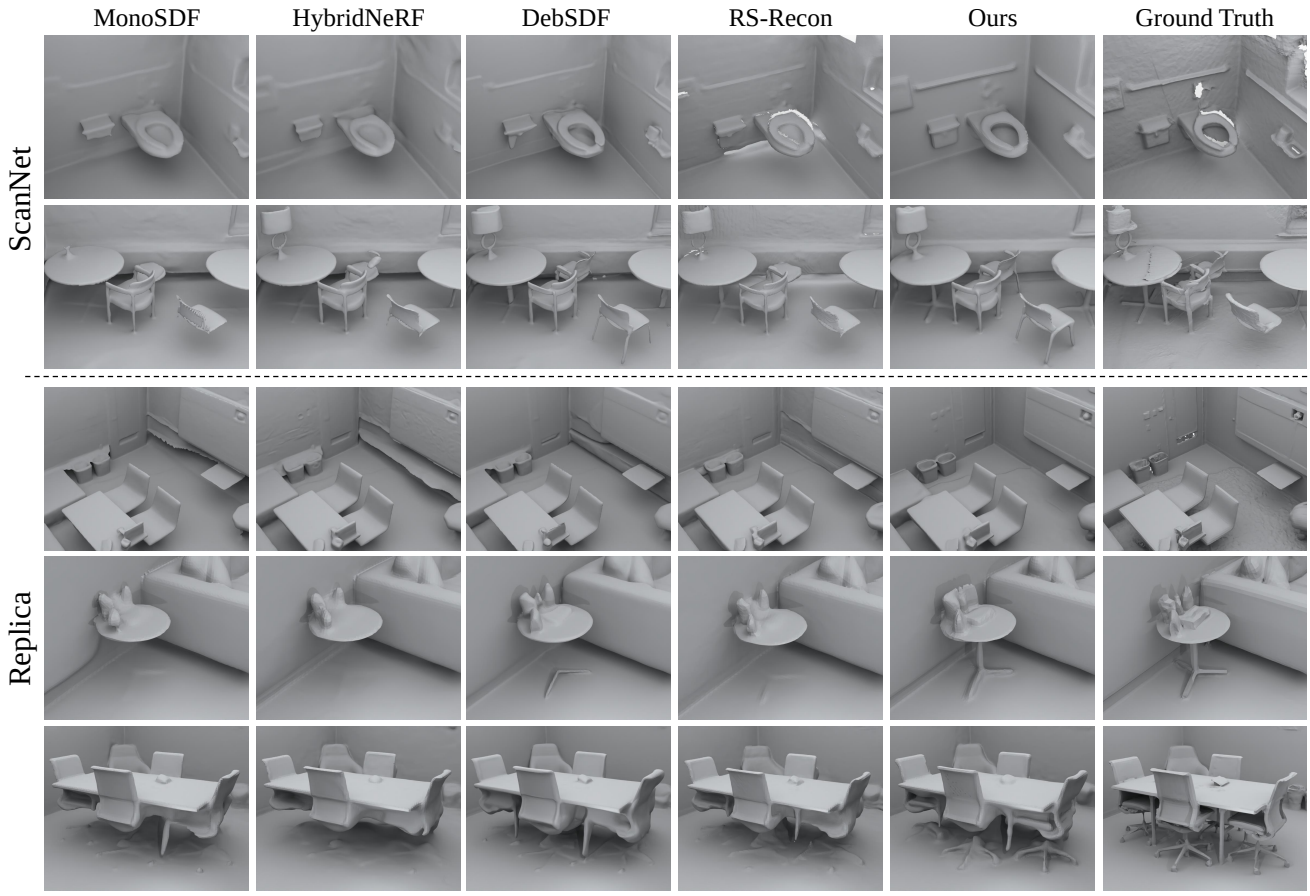


Figure 8. More visual comparisons of dense-view reconstruction on ScanNet and Replica datasets.

Table 2. Training time of each part.

| Task | Methods | Time of each part | | | | Total |
|---|---|---|---|---|---|---|
| Dense Recon | MonoSDF | Stage1: 14h | | | | 14h |
| | Ours | Stage1: 1.8h | Segmentation: 19min | Uncertainty: 13min | Stage2: 7.2h | 9.5h |
| Sparse NVS | FSGS | Stage1: 1.5h | | | | 1.5h |
| | Ours | Stage1: 0.3h | Segmentation: 2min | Uncertainty: 3min | Stage2: 1.3h | 1.7h |

*on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. 3

[12] Minyoung Park, Mirae Do, Yeon Jae Shin, Jaeseok Yoo, Jongkwang Hong, Joongrock Kim, and Chul Lee. H2O-SDF: Two-phase learning for 3d indoor reconstruction using object surface fields. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[13] Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4047–4056, 2023. 1

[14] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2

[15] Satu Elisa Schaeffer. Graph clustering. *Computer science review*, 1(1):27–64, 2007. 1

[16] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 4

[17] Ziyu Tang, Weicai Ye, Yifan Wang, Di Huang, Hujun Bao, Tong He, and Guofeng Zhang. ND-SDF: Learning normal

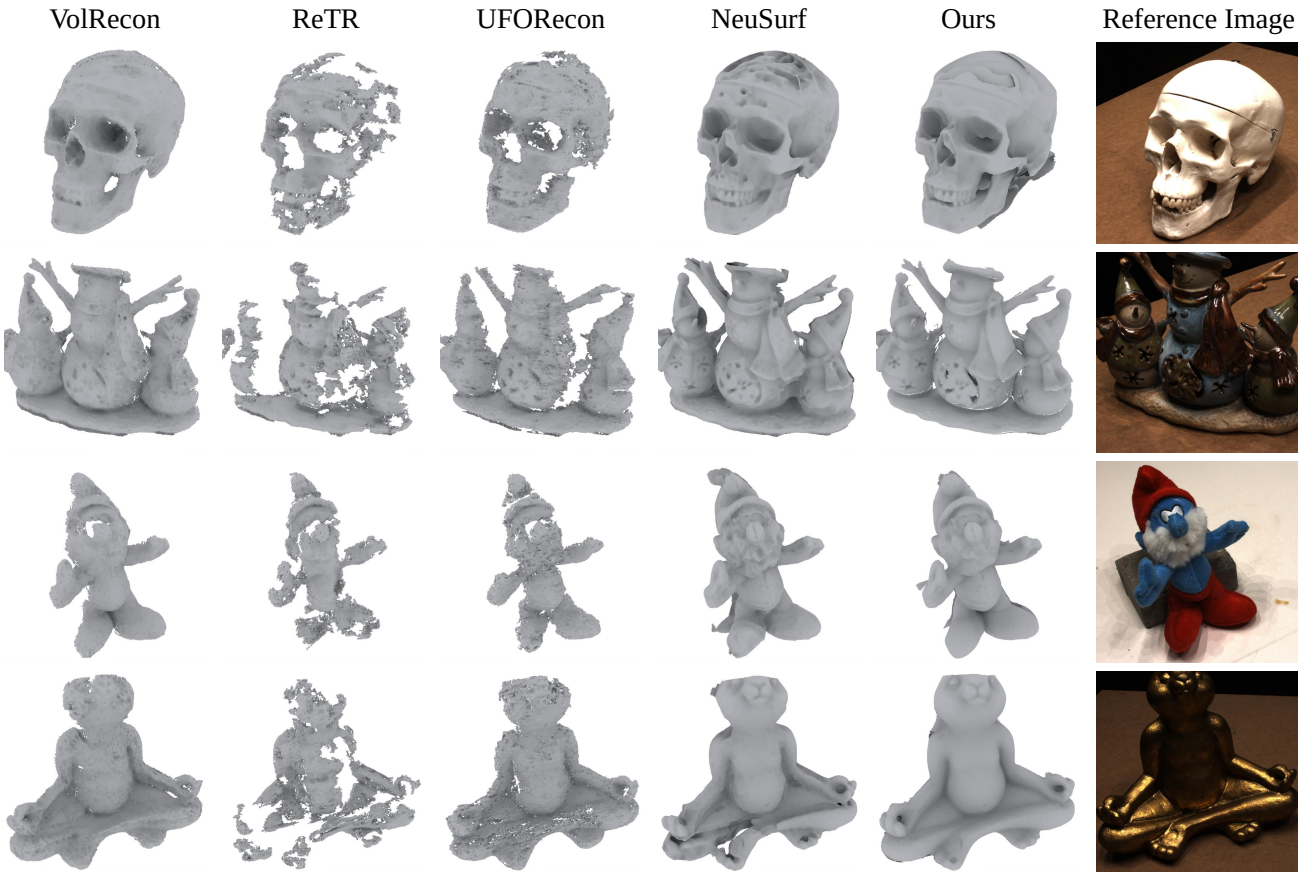


Figure 9. More visual comparisons of sparse-view reconstruction on DTU dataset.

deflection fields for high-fidelity indoor reconstruction. *International Conference on Learning Representations*, 2025. 1

[18] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. NeuRIS: Neural reconstruction of indoor scenes using normal priors. In *European conference on computer vision*, pages 139–155. Springer, 2022. 1

[19] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[20] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28274–28284, 2024. 1

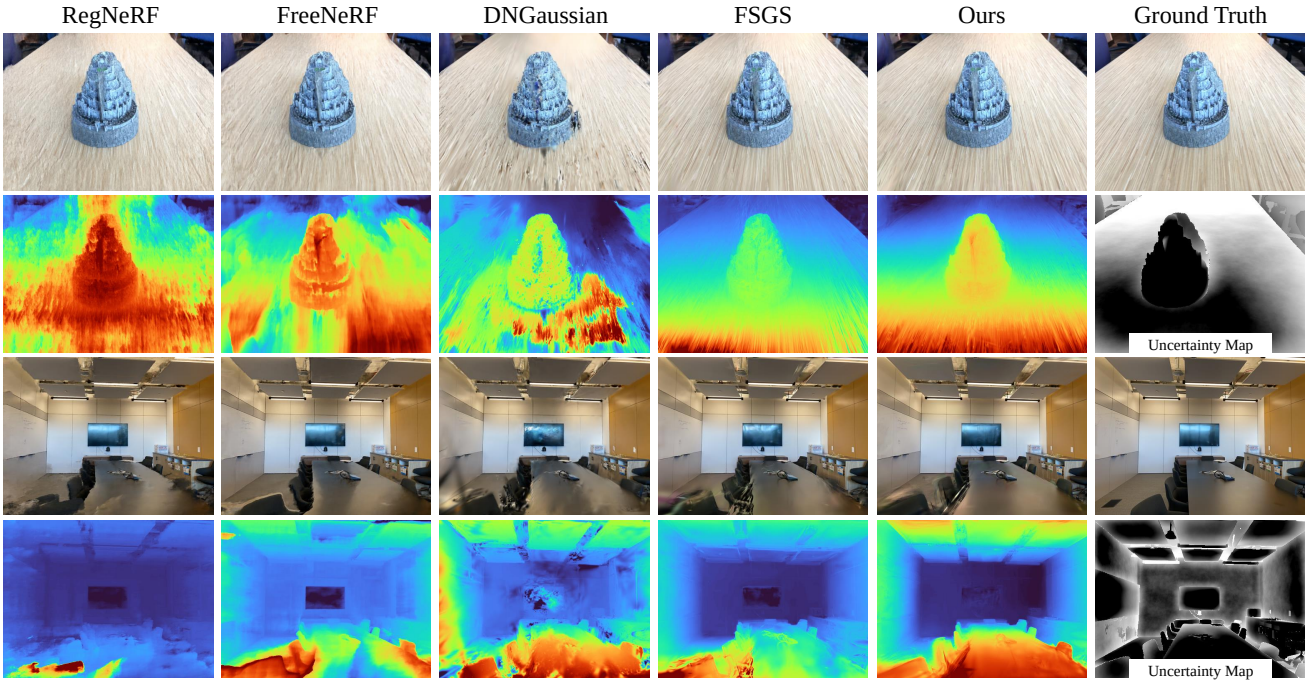


Figure 10. More visual comparisons of sparse-input novel view synthesis on LLFF dataset. Note that we don't evaluate monocular depths for test views, therefore we choose the closest viewpoint from the training views to display the estimated uncertainty maps.