Multi-Scale Implicit Surfaces Reconstruction for Outdoor Scenes

Tong Xu[1], Ruhao Wang[2], Fei Luo[1], and Chunxia Xiao[1(✉)]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
ggboard35421@gmail.com, {rhwang,luofei,cxxiao}@whu.edu.cn

Abstract. The images used to reconstruct 3D models in outdoor scenes are generally captured at different scales. Accurately reconstructing geometry from multi-scale images has not been extensively addressed. This work proposes a new volume rendering method that combines cone sampling and implicit surface representation to better model geometries from multi-scale images. Besides, to address the problem of unsmooth gradient of different frequency bands in general position encoding, we propose a dynamic position encoding strategy. Meanwhile, we propose an adaptive sampling strategy in image space to focus on the important regions near the reconstructed surfaces and the difficult regions where rendering error exists. Experiments are conducted on the Tanks and Temples dataset as well as the aerial photography dataset. The results show that, compared to the state-of-the-art methods, our work can produce competitive or better high-quality surface reconstruction, especially for scenes with multi-scale images and complex geometric structure.

Keywords: 3D Reconstruction · Implicit Surface Reconstruction · Neural Radiance Field · Multiple Scales · Photometric Consistency

1 Introduction

Reconstructing 3D geometry from multi-view scene images is one major task in computer vision. Traditional methods, like Structure from Motion (SfM) [18], have achieved relatively accurate 3D reconstruction results. The camera poses obtained through sparse reconstruction can be further used for dense reconstruction in Multi-view stereo (MVS) [6]. However, MVS methods usually use image-matching algorithms to perform dense matching and restore 3D point clouds. These methods perform well on Lambert Surfaces but may fail in weak texture areas or reflective areas.

Recently, some implicit surface reconstruction methods based on the neural radiance field have been proposed. Their main idea is to model the scene using multilayer perceptrons (MLPs) and then reduce the difference between the real images and the reconstructed images through differentiable rendering. There

F. Luo and C. Xiao – These authors are co-corresponding authors.

is a certain gap between the predicted geometry and the real geometry due to the shape radiance ambiguity. In addition, when reconstructing the complex geometry of object surfaces, specular reflection scattering on their surfaces makes it difficult for the network to fit this view-dependent effect.

We aim to reconstruct high-quality implicit surfaces from multi-scale images captured in outdoor scenes. We propose a new volume rendering method in this paper that combines hybrid cone sampling and SDF surface representation. The method decomposes the color of each conical frustum into view-dependent and view-independent components. Besides, to address the problem of unsmooth gradient of different frequency bands in position encoding, we propose a dynamic position encoding strategy. Meanwhile, we propose an adaptive sampling strategy in the image space to concentrate sampling rays in the reconstruction area and areas with large error. Finally, we introduce segmentation masks to handle dynamic objects and textureless regions.

2 Related Work

The task of 3D reconstruction is to recover the geometry of the real world from the perspective of vision. Traditional methods, such as SFM [18] and MVS [6], complete the transformation from 2D to 3D by extracting and matching image features and reconstructing the 3D geometry.

In recent years, 3D reconstruction through neural networks has gradually become mainstream. These methods can be roughly categorized into two types: surface rendering [1, 9, 10, 15, 22, 24, 26] and volume rendering [2–5, 12, 13, 17, 20, 25, 27, 31]. Besides, recovering geometry, texture, lighting and other information of the scene from the input images is the main task of inverse rendering. So 3D reconstruction can be regarded as a subtask of inverse rendering. That is, by constraining the rendered images and the real images, the parameter representation of the scene can be continuously optimized to make the final rendering effect more realistic.

The surface rendering method assumes that there is only one intersection point between light rays and surface, so the gradient is backpropagated only near the surface. For the surfaces with sharp depth changes, it may produce relatively smooth reconstruction results. As for the volume rendering, NeRF [14] has applied volume rendering technology, and its high quality effects in rendering have attracted wide attention. However, due to the existence of shape radiance ambiguity of NeRF [14], it is difficult to extract an accurate surface from NeRF. In order to solve this ambiguity, a depth prior is proposed in NerfingMVS [28] to constrain the sampling process of NeRF, which not only makes the rendered depth map more accurate, but also makes the rendered image more realistic.

There also are several methods for outdoor scene 3D reconstruction. Sun *et al.* [19] proposed a hybrid-voxel and surface-guided sampling technique, which can derive a better sampling area than sphere. It also utilizes the segmentation mask and appearance embedding. Other methods primarily aimed at novel view

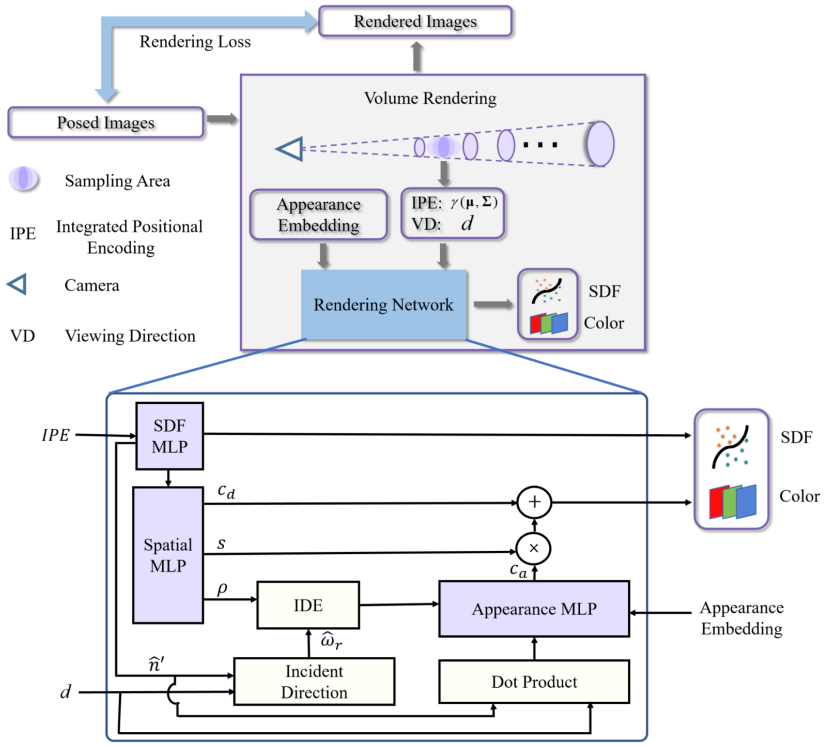

Fig. 1: Overview of our method. We project a cone from each pixel of the input posed image. The SDF value and the color of each conical frustum, which represents a specific sampling region, are generated by feeding the IPE into a rendering network together with the viewing direction d and the appearance embedding. We use the rendering loss to optimize the parameters of the network.

synthesis in large outdoor scenes [21, 30]. These large scenes have a significant amount of data, which is often multi-scale in nature.

3 Method

Fig. 1 is the overview of our method. Firstly, we project a cone from each pixel of the input posed image. The projected cone is divided into a series of conical frustums, each representing a specific sampling region. Then, an integrated positional encoding (IPE) representation is constructed. The SDF value and the color of each conical frustum are generated by feeding the IPE into a neural radiance field network along with the camera viewing direction and the appearance embedding. Next, the SDF values and the colors of all conical frustums are combined using the optimized volume rendering method to calculate the color value of the pixel. Specifically, the optimized volume rendering method first pre-

dicts the depth value of the pixel by a pre-trained depth estimation network and then designs an attenuation coefficient term to multiply with the original weight formula. Finally, the parameters of the network are optimized by iteratively training to generate accurate geometry and appearance representations of the current scene.

3.1 Multi-Scale Rendering with SDF Representation

Multi-Scale Sampling Strategy. We divide the cone region into a set of intervals [2]. For each interval, we calculate the mean μ and covariance Σ of the conical frustum and characterize these values by integrated positional encoding(IPE). The specific form of IPE is shown as:

$$\gamma(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left\{ \begin{bmatrix} \sin\left(2^{\ell}\boldsymbol{\mu}\right) \exp\left(-2^{2\ell-1}\operatorname{diag}(\boldsymbol{\Sigma})\right) \\ \cos\left(2^{\ell}\boldsymbol{\mu}\right) \exp\left(-2^{2\ell-1}\operatorname{diag}(\boldsymbol{\Sigma})\right) \end{bmatrix} \right\}_{\ell=0}^{L-1}, \tag{1}$$

where $\exp\left(-2^{2\ell-1}\operatorname{diag}(\boldsymbol{\Sigma})\right)$ represents the attenuation term and L is the higher dimension of position encoding.

Color Representation. Our network decomposes the color of a conical frustum into the view-independent color, like diffuse reflection, and the view-dependent color, like specular reflection [22]. View-independent color can be directly output by spatial MLP. In order to represent the view-dependent color, we use the reflection direction as the input of the network and parameterize it as a function of the normal vector and viewing direction. The specific form of the reflection direction is shown as:

$$\hat{\omega}_r = 2\left(\hat{\omega}_0 \cdot \hat{n}\right)\hat{n} - \hat{\omega}_0, \tag{2}$$

where $\hat{\omega}_o = -\hat{d}$ is a unit vector from a point in space to the center of the camera, opposite to the view direction $\hat{d}$, and $\hat{n}$ is a normal vector at the point, which is the gradient of the SDF values.

With spatially varying materials in complex scenes, the view-dependent color cannot be represented only as a function of reflection direction. Hence, we introduce Integrated Directional Encoding (IDE) [22]. IDE encodes the distribution of reflection direction $\hat{\omega}$ using a set of spherical harmonics Y_l^m under von Mises-Fisher (vMF) distribution [22] with mean $\hat{\omega}_r$ and concentration parameter $k = 1/\rho$. The roughness ρ is output by the spatial MLP. The final integrated direction encoding function is defined as:

$$\begin{aligned} \mathbf{IDE}(\hat{\omega}_r, k) &= \left\{\mathbb{E}_{\hat{\omega}\sim\mathbf{vMF}(\hat{\omega}_r,k)}\left[Y_l^m(\hat{\omega})\right]\right\}, \\ \text{with} \quad (l,m) &\in \left\{(l,m) | l = 2^0, 2^1, ..., 2^L; m = 0, 1, ..., l\right\}. \end{aligned} \tag{3}$$

The IDE expression can be further simplified as:

$$\begin{aligned} \mathbf{IDE}(\hat{\omega}_r, k) &= A_l(k) Y_l^m(\hat{\omega}_r), \\ \text{with} \quad A_l(k) &\approx \mathbf{exp}(-\frac{l(l+1)}{2k}). \end{aligned} \tag{4}$$

In outdoor scenes with uncontrollable lighting, the lighting conditions can change. Therefore, we defines an appearance embedding that encodes each image into a vector as the input of appearance MLP. The color of a conical frustum can be synthesized by combining the view-independent color c_d output from the spatial MLP and the view-dependent color c_a output from the appearance MLP. The blending coefficient of the viewpoint-dependent color is S, which is output by the spatial MLP. The color of a conical frustum is shown as:

$$color = c_d + S \times c_a. \tag{5}$$

Depth-Guided Volume Rendering. Before introducing the optimized volume rendering strategy in our work, we first review the weight formula used in NeuS [23]. The specific form of the weight formula is defined as:

$$w(t) = \lambda(t)T(t)\rho(t), \tag{6}$$

where $T(t)$ corresponds to the accumulated transmittance along the ray and $\rho(t)$ corresponds to the opacity in NeRF. These two functions can be respectively expressed as:

$$T(t) = \Phi_s(f(\mathbf{p}(t))), \tag{7}$$

$$\rho(t) = \max\left(\frac{-\frac{\mathrm{d}\Phi_s}{\mathrm{d}t}(f(\mathbf{p}(t)))}{\Phi_s(f(\mathbf{p}(t)))}, 0\right), \tag{8}$$

where $\Phi_s(x) = (1 + e^{-sx})^{-1}$, $\mathbf{p}(t)$ represents pixel rays and f represents SDF MLP. This weight function can achieve a local maximum at the point where the SDF value is 0. When a ray penetrates through multiple surfaces, the weight of the first surface is greater than the second surface. However, points near the second surface penetrated by the ray still contribute to the final color.

To address the issue of weight ambiguity, we propose a depth-guided volume rendering strategy. Specifically, we predict depth in multi-view images through a pre-trained depth estimation network. For regions closer to the predicted depth, their weights remain the same as the original weight values. For regions further away from the predicted depth, a decay factor λ is introduced in the weight formula to penalize the weights. The specific form of decay factor λ is shown as:

$$\lambda(t) = \begin{cases} 1 & 0 < |Dp - t| < \alpha \\ \exp(\beta(\alpha - |Dp - t|)) & |Dp - t| \geq \alpha \end{cases}, \tag{9}$$

where Dp represents the depth predicted by the pre-trained depth estimation network, t is the depth of the sampled point, α and β are pre-set hyperparameters that control the confidence range and decay rate in the weight function. The final weight function is shown as:

$$w_{new}(t) = \lambda(t)w(t). \tag{10}$$

Reference Image Predicted Mask Sampling Range

Fig. 2: From left to right: Original image, the foreground mask predicted by the network, final sampling range after image processing. For not generating artifacts in the edge area of the reconstructed object, the sampling range is larger than the predicted mask.

3.2 Dynamic Position Encoding

NeRF [14] uses a strategy to encode the coordinates of three-dimensional points in space. It encodes each coordinate of the point x and the observation direction d to $2L$ dimensions, so that the network can fit the high-frequency part of the scene. The coding formula is defined as:

$$p(x) = \begin{cases} \sin(2^0\pi x), \cos(2^0\pi x) \\ \sin(2^1\pi x), \cos(2^1\pi x) \\ \cdots \\ \sin(2^{L-1}\pi x), \cos(2^{L-1}\pi x) \end{cases} \tag{11}$$

According to Eq. (11), by making derivative of x, the scale coefficient will gradually increase with the number of layers.

In order to solve this problem, a new position encoding strategy is proposed in BARF [11]. Based on the original position encoding, it adds a weight coefficient to gradually release the network's response to high-frequency information during the training process. We use this strategy to define the weight coefficient of the position encoding, the k-th weight ω_k is defined as:

$$\omega_k(\beta) = \begin{cases} 0 & \text{if } \beta < k \\ \frac{1-\cos((\beta-k)\pi)}{2} & \text{if } 0 \leq \beta - k < 1, \\ 1 & \text{if } \beta - k \geq 1 \end{cases} \tag{12}$$

where $k = 0, 1, 2, ..., L-1$ and β is trainable in the optimization process, gradually increasing from 0 to L. As the training gradually iterates, the position encoding formula is closer to the Eq. (11). Thus, the final dynamic position encoding fomula is defined as:

$$\gamma(x) = \omega_k(\beta)p(x). \tag{13}$$

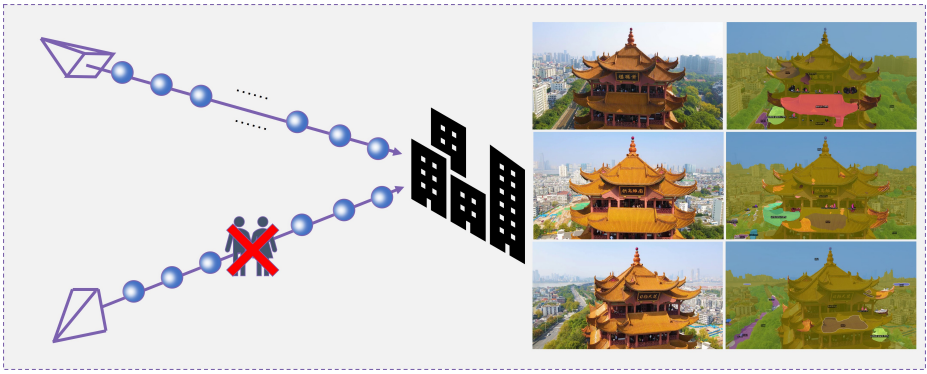

Fig. 3: Ray sampling algorithm. We add a segmentation mask algorithm Detectron2 [29] to exclude the ray belonging to transient objects from the training. We let the ray weight of the poorly textured area (*e.g.* sky) approach 0 to avoid reconstruction included in the spherical shell.

3.3 Adaptive Sampling Strategy in Image Space

Defining which pixel requires centralized sampling not only can accelerate the convergence speed of the network, but also improve the geometry accuracy. In this section, we introduce two adaptive sampling strategies proposed in our work.

Sampling Focusing Strategy. One idea for speeding up network convergence is to reduce the sampling range in the reconstruction area. In Fig. 2, we use NeRF++ [32] to model the background of the scene so that the foreground mask can be extracted after a certain stage of training. Because the mask image predicted by the network is not accurate enough, we still need to do some image processing to eliminate noise. Besides, if we only use the denoised mask as the sampling range, some artifacts will appear on the borders of the predicted mask due to the lack of constraint for empty space. To enhance the constraints on empty space, we expand the denoised mask to get a wider range of the mask.

Confidence-Based Sampling Strategy. To further optimize the geometry structure of the scene, the confidence score between the rendered color and the real color is calculated to assign lower confidence values to the positions where the color estimation is inaccurate. This method can make the network pay more attention to areas with inaccurate color estimation, thereby improving the accuracy of scene geometry. The specific form of confidence S_j^i is defined as:

$$S_j^i = 1 - \frac{1}{3}\left|\left|C_{gt}^i(j) - C_{render}^i(j)\right|\right|_1, \tag{14}$$

where $C_{gt}^i(j)$ is the ground truth color of the j-th pixel in the i-th image, $C_{render}^i(j)$ is the rendered color of the j-th pixel in the i-th image. $||.||_1$ repre-

sents the $L1$ norm, calculating the absolute value of the color difference between the two pixels.

3.4 More Details

When 3D reconstruction is carried out for outdoor scenes, different semantic parts of the scene have different effects on 3D reconstruction, so we use semantic segmentation to process different parts of the scene. Fig. 3 shows our segmentation results and ray sampling strategy. In our algorithm, the panoramic segmentation model pretrained in detectron2 [29] is used to get the segmentation mask. we use the scene segmentation to constrain the ray pointing to the sky so that the rays in these places are in the weight sum area 0 in the network. Additionally, we only sample the static part of the scene when sampling, to ensure that there is no occlusion problem for the same surface point to be recovered.

3.5 Loss Function

In general, we randomly sample the pixels of the input image to get $\{C_k, M_k, o_k, v_k\}$, where C_k represents the color corresponding to the pixel, $M_k \in \{0, 1\}$ represents whether the pixel belongs to the sky area. o_k and v_k represent the origin points and their directions of responding sampling rays respectively. The final loss function is defined as follows:

$$\mathcal{L}_{sum} = \mathcal{L}_{color} + \mathcal{L}_{sky} + \mathcal{L}_{reg}. \tag{15}$$

The color loss is defined as:

$$\mathcal{L}_{color} = \frac{1}{m} \sum_{k} \mathcal{R}(\hat{C}_{i,k}, C_{i,k}), \tag{16}$$

where $\hat{C}_{i,k}$ represents the pixel color rendered by volume rendering, $C_{i,k}$ represents the ground-truth color and $\mathcal{R}$ represents $L1$ loss.

The sky mask loss is actually a variant of the mask loss. It can separate the sky from the reconstructed foreground. The mask loss of sky is defined as:

$$\mathcal{L}_{sky} = BCE(M_k, \hat{M}_k), \tag{17}$$

where $\hat{M}_k$ represents the sum of sample point weights on the ray predicted by the network.

To regularize the SDF predicted by network f_θ, we also use the eikonal term [7] to normalize the sampling points on the ray:

$$\mathcal{L}_{reg} = \frac{1}{mn} \sum_{k,j} (||\nabla f(\hat{p}_{k,j})||_2 - 1)^2. \tag{18}$$

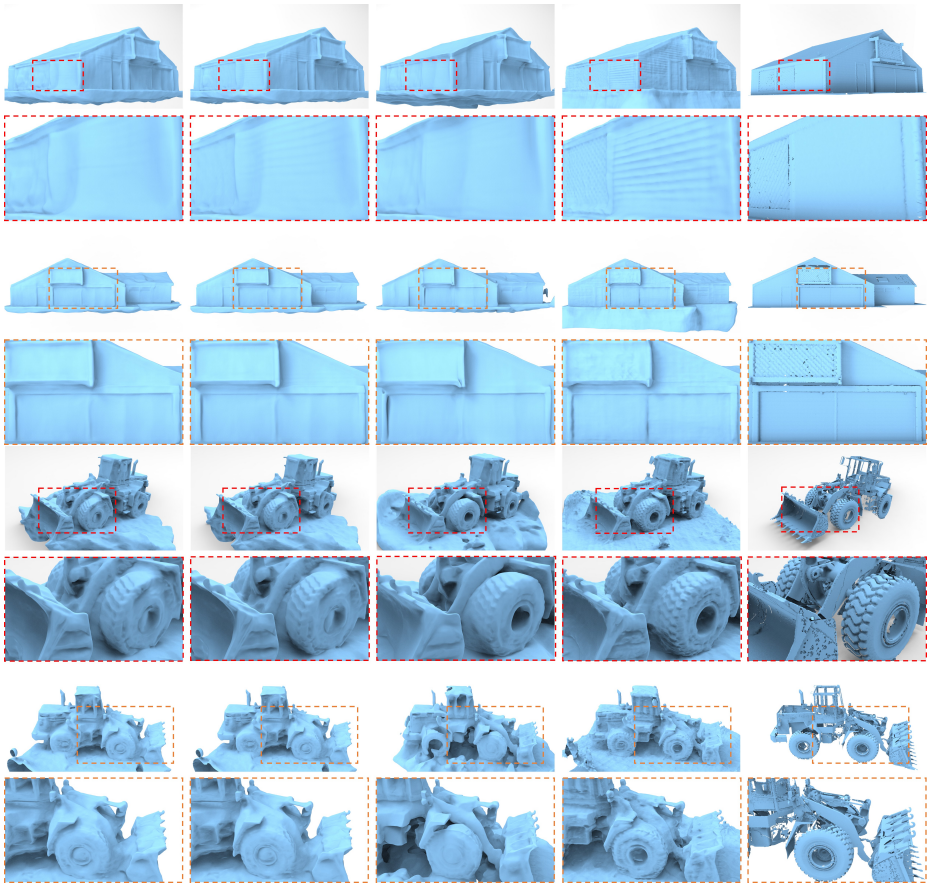

Fig. 4: Qualitative comparisons on the Tanks and Temples. (a) UNISURF [16]; (b) NeuS [23]; (c) Sun *et al.* [19]; (d) Ours; (e) Ground Truth.

4 Experiments

4.1 Implementation Details

We assume the reconstructed area to be inside a unit sphere and use two MLP networks to represent the geometric part and color part of the scene respectively. We use 8 layers with 256 hidden units for the geometry MLP and 4 layers with 512 hidden units for the color MLP. For spatial MLP, we use 8 layers with 256 hidden units. The number of sampled pixels is equal to the batch size, and we set it to 1024. The initial learning rate is set to 5e-4 and dynamically changes during the optimization process. The α value is set to 0.05. The learning rate lr gradually decreases from $(1-\alpha) \times \text{lr}$ to $\alpha \times \text{lr}$ during the network optimization process. We realize our method on a single NVIDIA RTX3090Ti GPU.

Table 1: Quantitative results on the Tanks and Temples.

Method	Chamfer Dist. ↓	Precision ↑	Recall ↑
UNISURF [16]	1.14	0.52	0.51
NeuS [23]	0.83	0.70	0.56
Sun *et al.* [19]	0.67	**0.73**	0.64
Ours	**0.52**	0.71	**0.82**

Table 2: Quantitative results on the aerial photography dataset.

Method	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
UNISURF [16]	64.21	21.48	0.69	0.54
NeuS [23]	63.59	23.33	0.71	0.51
Sun *et al.* [19]	62.76	22.77	**0.73**	0.49
Ours	**56.62**	**23.88**	**0.73**	**0.48**

4.2 Qualitative and Quantitative Comparisons

We validate the effectiveness of our method using the Tanks and Temples dataset [8] as well as the aerial photography dataset. The aerial photography dataset derives from videos taken circularly around the reconstructed objects. We extract frames and pre-process them to generate the data format. We compare our method with the current new representative methods, including UNISURF [16], NeuS [23], and Sun *et al.* [19].

Qualitative results on Tanks and Temples with ground truth. Fig. 4 shows the qualitative results on the Tanks and Temples with ground truth. Since the ground truth of this dataset is point cloud rather than mesh, we perform a qualitative comparison by first downsampling the point cloud data of the ground truth to 300,000 points. Then, we exploit the downsampled points for surface reconstruction using the Ball Pivoting algorithm. In terms of qualitative results, the proposed method in this study demonstrates superior performance in overall structure and details.

Quantitative results. Table 1 and Table 2 show the quantitative results. The results show that our method has achieved the best or the second-best quantitative indicators in all aspects.

Qualitative results on the Tanks and Temples without ground truth. Fig. 5 shows the qualitative results on the Tanks and Temples without ground truth. In the family scene, other methods exhibit blurry details in the characters, and the details of the arms and folds on the clothes are not well restored. Our method can recover these details more effectively and outperforms Sun *et al.* [19] in terms of overall structure. In the Francis scene, our method produces smoother reconstruction results in the staircase area.

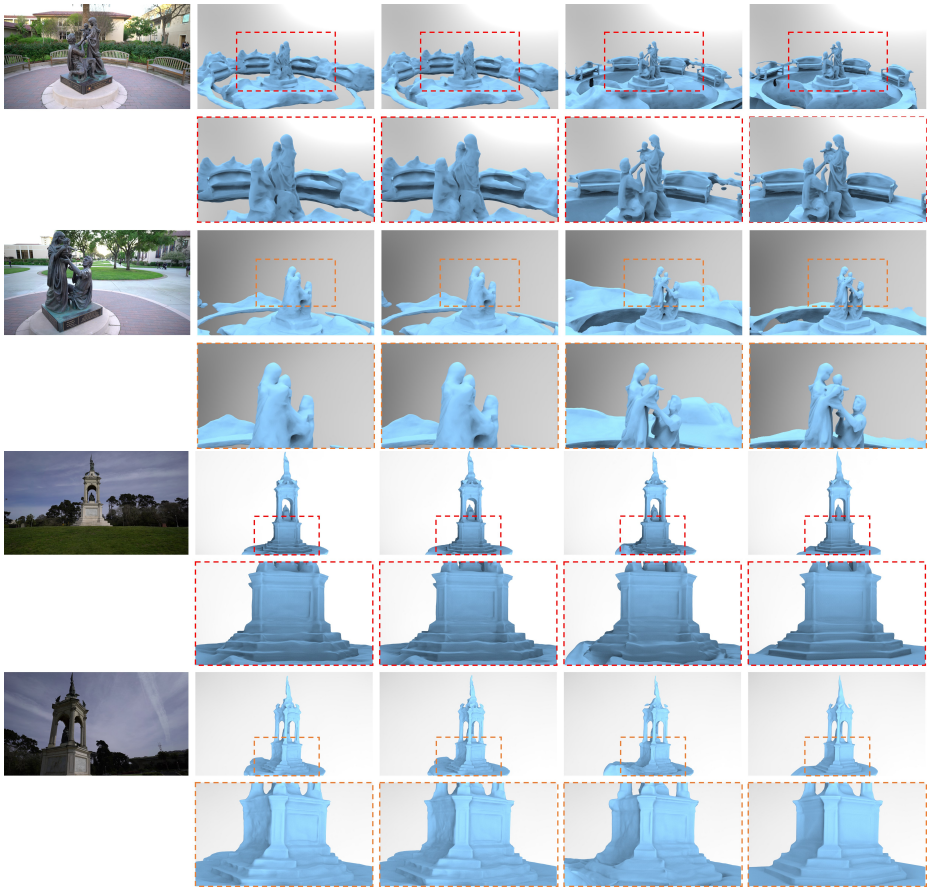

Fig. 5: Qualitative comparisons on the Tanks and Temples. (a) Reference Image; (b) UNISURF [16]; (c) NeuS [23]; (d) Sun *et al.* [19]; (e) Ours.

Qualitative results on the aerial photography dataset. Fig. 6 shows the qualitative results on the aerial photography dataset. Our method presents certain degree of improvement compared to other methods. Compared to the baseline method and the method proposed by Sun *et al.* [19], our method exhibits noticeable improvements.

4.3 Ablation Study

To evaluate the impact of different components in our proposed method, we design six variants: (a) *w/o Multi-Scale Sampling*; (b) *Only viewing direction as input*; (c) *w/o Depth-Guided Volume Rendering*; (d) *w/o Dynamic Positional Encoding*; (e) *w/o Adaptive Sampling Strategy*; (f) *w/o Segmentation Mask*.

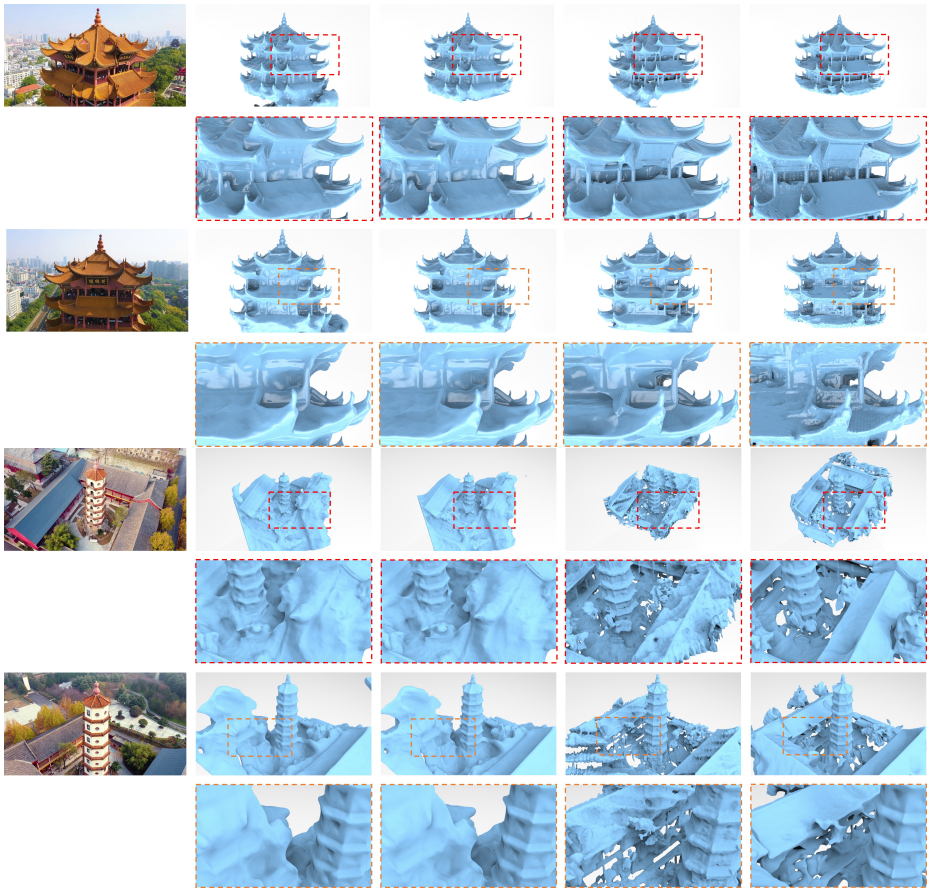

Fig. 6: Qualitative comparisons on the aerial photography dataset. (a) Reference Image; (b) UNISURF [16]; (c) NeuS [23]; (d) Sun *et al.* [19]; (e) Ours.

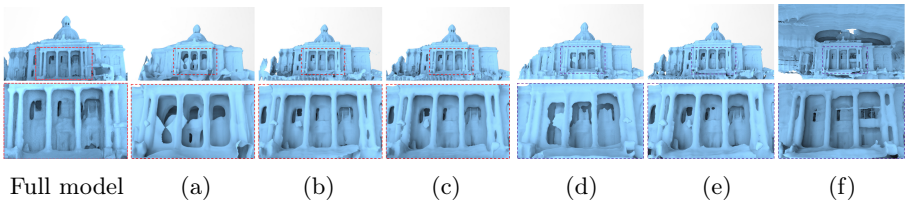

Fig. 7: Ablation study results. The full model is on the left. (a) w/o Multi-Scale Sampling; (b) Only viewing direction as input; (c) w/o Depth-guided Volume Rendering; (d) w/o Dynamic Positional Encoding; (e) w/o Adaptive Sampling Strategy; (f) w/o Segmentation Mask.

Table 3: Quantitative comparison results of Ablation Study.

Variants	Chamfer Dist. ↓	Precision ↑	Recall ↑
(a)	1.11	0.72	0.70
(b)	0.75	0.66	**0.80**
(c)	0.73	0.66	0.79
(d)	1.27	**0.73**	0.66
(e)	0.68	0.64	0.76
(f)	**0.56**	0.59	0.50
Full model	0.72	**0.73**	**0.80**

We conduct an ablation analysis of our method in the courthouse scene, and the quantitative results of the experiments are shown in Table 3. The visual effects of the experiments are illustrated in Fig. 7. (a) indicates that the multi-scale sampling strategy has a significant impact on the reconstruction results. (b) suggests that modeling both specular reflections and illumination simultaneously is important. (c) and (e) demonstrate that depth-guided weight formulation and adaptive sampling strategy contribute to improved reconstruction accuracy. (d) indicates that the dynamic position encoding strategy can restore the overall structure of the scene more comprehensively. (f) shows that using a segmentation mask strategy can effectively separate the sky from the reconstructed objects, preventing the inclusion of sky-related artifacts in the reconstruction results.

5 Conclusion

We have proposed a method to reconstruct high-precision implicit surfaces from multi-scale images captured in outdoor scenes. To integrate the multi-scale rendering method into the geometric representation of the scene, we introduce a novel volume rendering approach that combines hybrid cone sampling and implicit surface representation. To address the problem of unsmooth gradient of different frequency bands in general position encoding, we propose a dynamic position encoding strategy. We also introduce an adaptive sampling strategy in the image space to concentrate rays on the reconstruction and error-prone regions. Finally, we introduce the segmentation mask to handle dynamic objects and weak texture regions.

Acknowledgement. This work is partly supported by the National Natural Science Foundation of China (No.61972298 and No.62372336), CAAI-Huawei MindSpore Open Fund and Wuhan University-Huawei GeoInformatices Innovation Lab.

References

1. Akkouche, S., Galin, E.: Implicit surface reconstruction from contours. The Visual Computer **20**(6), 392–401 (2004)

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. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
4. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
5. Fu, Q., Xu, Q., Ong, Y.S., Tao, W.: Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. Advances in Neural Information Processing Systems **35**, 3403–3416 (2022)
6. Goesele, M., Curless, B., Seitz, S.M.: Multi-view stereo revisited. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 2402–2409. IEEE (2006)
7. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. arXiv preprint arXiv:2002.10099 (2020)
8. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)
9. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8456–8465 (2023)
10. Li, Z., Long, X., Wang, Y., Cao, T., Wang, W., Luo, F., Xiao, C.: Neto:neural reconstruction of transparent objects with self-occlusion aware refraction-tracing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18547–18557 (October 2023)
11. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5741–5751 (2021)
12. Lindell, D.B., Martel, J.N., Wetzstein, G.: Autoint: Automatic integration for fast neural volume rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14556–14565 (2021)
13. Luo, F., Zhu, Y., Fu, Y., Zhou, H., Chen, Z., Xiao, C.: Sparse rgb-d images create a real thing: A flexible voxel based 3d reconstruction pipeline for single object. Visual Informatics **7**(1), 66–76 (2023)
14. 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)
15. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3504–3515 (2020)
16. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5589–5599 (2021)

17. Pan, X., Xu, X., Loy, C.C., Theobalt, C., Dai, B.: A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. Advances in Neural Information Processing Systems **34**, 20002–20013 (2021)
18. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
19. Sun, J., Chen, X., Wang, Q., Li, Z., Averbuch-Elor, H., Zhou, X., Snavely, N.: Neural 3d reconstruction in the wild. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–9 (2022)
20. Takikawa, T., Litalien, J., Yin, K., Kreis, K., Loop, C., Nowrouzezahrai, D., Jacobson, A., McGuire, M., Fidler, S.: Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11358–11367 (2021)
21. Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12922–12931 (2022)
22. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5481–5490. IEEE (2022)
23. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
24. Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3295–3306 (2023)
25. Wang, Y., Skorokhodov, I., Wonka, P.: Hf-neus: Improved surface reconstruction using high-frequency details. Advances in Neural Information Processing Systems **35**, 1966–1978 (2022)
26. Wang, Y., Li, Z., Jiang, Y., Zhou, K., Cao, T., Fu, Y., Xiao, C.: Neuralroom: Geometry-constrained neural implicit surfaces for indoor scene reconstruction. ACM Trans. Graph. **41**(6) (nov 2022). https://doi.org/10.1145/3550454.3555514
27. Wang, Z., Luo, F., Long, X., Zhang, W., Xiao, C.: Learning long-range information with dual-scale transformers for indoor scene completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18569–18579 (October 2023)
28. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5610–5619 (2021)
29. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
30. Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D.: Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII. pp. 106–122. Springer (2022)
31. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems **34**, 4805–4815 (2021)
32. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)