MegaSurf: Efficient and Robust Sampling Guided Neural Surface Reconstruction Framework of Scalable Large Scene

ANONYMOUS AUTHOR(S)

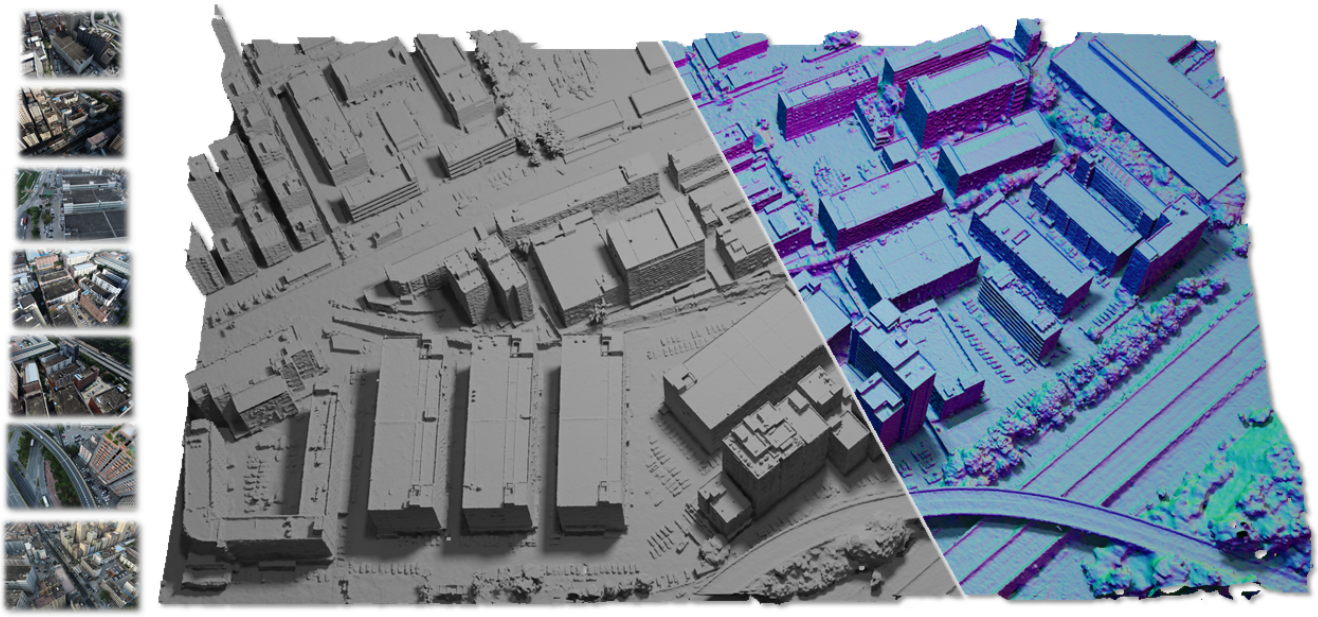


Fig. 1. We present MegaSurf, an efficient and robust neural surface reconstruction framework to reconstruct the 3D large-scale scene from thousands of input RGB images collected by the drone. MegaSurf has both the robustness of the stereo matching and the high-fidelity details of the rendering-based reconstruction methods.

Neural surface reconstruction (NSR) has been shown to have huge potential for 3D reconstruction from multiview images. However, current NSR methods struggle to reconstruct high-quality surfaces due to severe shape-radiance ambiguities when they meet the large-scale scenes captured from aircraft or UAVs, which often contain heavy shadows, illumination variations, and low texture areas. We present MegaSurf, which efficiently and robustly integrates Multiview Stereo (MVS) priors to solve large geometric errors due to the intrinsic shape-radiance ambiguity while preserving high-precision details. Specifically, we propose a lightweight MVS module to rapidly diffuse high-confidence planar geometric information from structure-from-motion (SFM) points, where ambiguities often occur, to guide the NSR. Further, we propose a two-stage sampling-guided NSR approach. We pre-train a sampling proposal network using MVS priors to indicate the next stage sampling position and let these positions represent the scene first at the next stage of training. This strategy helps to overcome large geometric errors due to ambiguity while preserving the high-fidelity details. Our MegaSurf improves the speed of prior acquisition by more than four times that of the SOTA MVS methods and achieves the best reconstruction accuracy on large-scale datasets compared to previous methods.

CCS Concepts: • **Computing methodologies** → **Reconstruction**.

Additional Key Words and Phrases: Neural Surface Reconstruction, Large Scale Scenes, Multiview Reconstruction

1 INTRODUCTION

With airborne oblique imaging, large-scale 3D surface reconstruction has shown tremendous value in urban planning, virtual reality, and navigation. In recent decades, multiview stereo matching [4, 13, 24] has been widely used for large-scale reconstruction. Due to the nature of window-based photo-consistency matching, it is hard to reconstruct fine structures. Recently, neural surface reconstruction (NSR) [8, 18, 26, 27], extended from neural radiance field (NeRF) [10], have significantly promoted the development of surface reconstruction concerning their reconstruction of high fidelity details on small objects or scenes, which conventional MVS methods can hardly achieve. Although NeRF has attempted to apply on the large-scale datasets for novel view synthesis [15, 16, 22, 23], the researches on NSR are limited. Only some works [6, 8] extend the hash table-based NSR approach to limited-scale outdoor scenes since traditional MLP-based methods usually oversmooth the outdoor scenes.

NSR methods employ loss between rendered color and the input image to back-propagate updates for the scene geometry, current NSR methods face a severe shape-radiance ambiguity [3, 19, 29] which would cause large reconstruction defects. Especially in airborne scenes, the ambiguity often occurs in regions like low textures,

heavy shadows, and high illumination variance. Instead, conventional MVS methods use a direct and more robust stereo-matching approach for surface reconstruction and further employ matching cost aggregation strategies [7, 13, 25] to improve the quality in these difficult regions. Therefore, using multi-view geometric information as a prior to guiding the optimization of NSR [19, 29] has been proved to be a simple and effective strategy.

As for prior acquisition in large-scale scenes, PatchMatch stereo matching methods [13, 24] are the good choice due to their scalability and robust performance. However, these methods require high computations on several global PatchMatch sweeps and a depth fusion to improve the completeness and fine structure reconstruction, which is redundant when fusing into an NSR pipeline, where NSR often performs better. This computational overhead cannot be ignored in large scenes as it scales with the number of images captured. [2, 3] introduce a local photo-consistency loss into NSR without large computations on PatchMatch global optimization, which does not perform well in textureless and shadow regions where shape-radiance ambiguity often occurs.

Although the intensive MVS computations are performed, the direct use of the priors to design a geometric loss[19, 29, 32] would also inevitably introduce noises to degrade high-precision details reconstruction from NSR methods. Confidence values of the geometric priors can be introduced to mitigate the effects of noises. Wei et al. [20] guide the sampling near geometric priors while using confidence to define the sampling range. However, the confidence of MVS prior is often difficult to quantify, and how to use the confidence to control the strength of sampling guidance is often case-dependent.

In this paper, we propose **MegaSurf** to overcome the large geometric error resulting from severe shape-radiance ambiguities while preserving the highly detailed structures in large-scale scenes (Fig. 1). MegaSurf adopts the basic Neuralangelo [8] training strategy for reconstructing high detailed structures. We then present a two-step training strategy to use MVS geometric priors to guide NSR efficiently and robustly: **1)** We propose a lightweight PatchMatch MVS module focusing on extracting high confident planar geometries propagated from SFM points where large shape-radiance ambiguity more likely occur. Instead of sweeping several times through all pixels globally from random geometry initialization [13, 24], our local propagation strategy lets every pixel receive the geometric candidates propagated from neighboring high confident SFM points when the pixel is calculated. Each pixel is only calculated once if the pixel is in a planar geometry similar to SFM points. The other regions, like the small details and trees, are reconstructed mainly relying on NSR. **2)** We propose a two-stage sampling guided training approach to integrate geometric priors to NSR instead of directly adding geometric loss into NSR optimization. In the first stage, we train a cascaded proposal net using geometric priors, which naturally transforms geometric priors into the sampling priors of NSR. In the second phase, we train the rendering net using the well-trained proposal net to control the sampling position while conducting a non-occupancy loss to prevent the correct prior information learned by the proposal net from being destroyed by the shape radiance ambiguity.

In summary, our main contributions are the following:

- We present a lightweight MVS module to efficiently obtain high-confidence planar geometric priors over four times improvement in speed where large shape-radiance ambiguities often occur.
- We present a two-stage sampling guided training to robustly integrate geometric prior by pretraining a proposal net using the geometric priors, which overcomes severe shape-radiance ambiguity while preserving high-fidelity details.
- Our method outperforms previous SOTA NSR and MVS methods on several large scene airborne datasets.

2 RELATED WORK

Multiview stereo matching. Multiview Stereo methods rely on the photo-consistency matching among multiview images to estimate depth maps, which are used to fuse into dense point clouds. The performance of local photo-consistency matching is easily reduced in regions with low textures, shadows, and non-Lambertian materials. Several global matching aggregation methods are applied to improve these regions' reconstruction quality, including semi-global optimization [7], PatchMatch [13], and 3D convolution regularization [25]. Among all the MVS methods, Patchmatch-based MVS methods [13, 24], with their efficient parallelization structure and robust performance, are more suitable and knowledgeable for large-scale scene reconstruction. Even though the learning-based MVS methods [5, 25, 30] show their advantages of reconstruction in difficult regions, their application on large-scale airborne datasets is limited due to the lack of various 3D training sets, which are often expensive to acquire.

Neural surface reconstruction. Recently, rendering-based neural surface reconstruction methods [18, 26, 27] have become a promising way to promote the development of 3D reconstruction due to their high-quality reconstruction results especially on accurate details. The multi-resolution hash encoding [11] provides a compact high-resolution feature representation to promote the training speed and show its potential for high-fidelity reconstruction for large scenes. Li et al [8] introduce a progressive training on the multi-resolution hash encoding representation, and a numerical calculation of normals, extending the high reconstruction accuracy to large outdoor scenes. However, on large-scale scenes, specially acquired from airborne equipment and under the non-object-centric setting, large geometrical errors often occur due to the shape-radiance ambiguity, which is deteriorated by the heavy shadows, low textures, and illumination variations.

Geometric prior guided neural surface reconstruction. Geometric prior guided NSR methods are well-studied in indoor scenes. The intrinsic shape-radiance ambiguity is a serious problem for indoor scene NSR due to the textureless walls, reflective windows, and floors. In the indoor environment, geometry prior is often derived from depth sensors[1], monodepth estimation networks [12, 21], volumetric learning networks [14, 17] as MVS methods also fail to reconstruct these geometries. These learning approaches to derive geometric priors can only be applicable when the 3D training sets are available. Instead of extracting explicit geometric priors, several methods use local multiview photo-consistency [2, 3] to improve the surface reconstruction for laboratory datasets where the multiview images are well captured. To our knowledge, we can

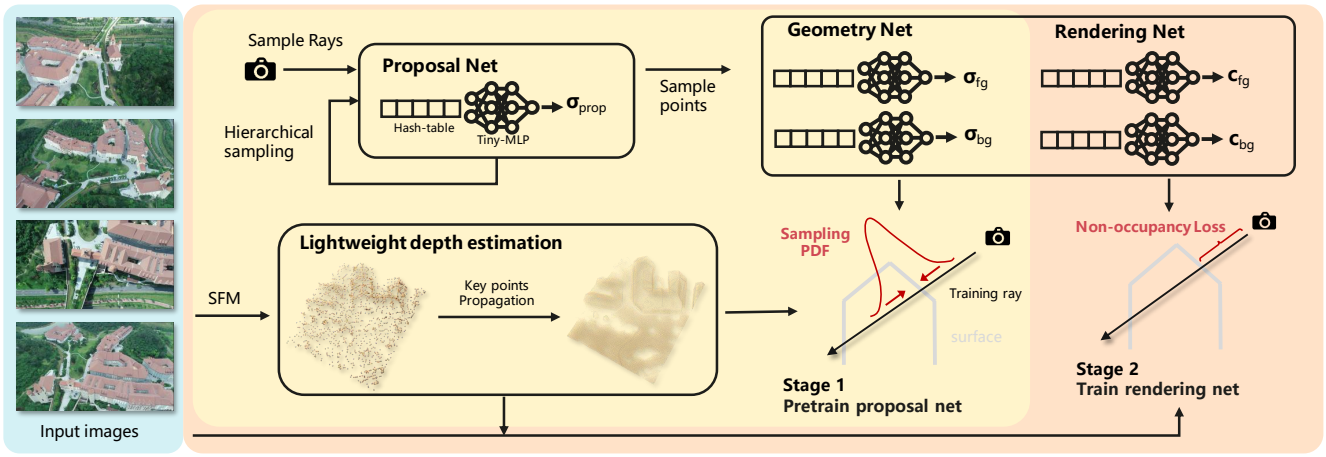


Fig. 2. Method overview. We propose a lightweight MVS module to rapidly propagate SFM points to obtain high-confidence geometry to guide the training of NSR. Then, we propose a two-stage sampling guided training strategy using the geometric prior to alleviate the shape-radiance ambiguity while preserving the details. Stage 1 only trains the proposal and geometry net to encourage sampling probability density function (PDF) concentrating on the space near geometric priors. Stage 2 adds the rendering net and trains the whole structure. Besides, a non-occupancy loss is added to suppress the density between the camera center and geometric priors.

not find any geometric prior guided NSR method applied for large airborne datasets. One reason could be that MVS introduces a large computational overhead which scales with huge number of images captured in large scenes. We propose a lightweight MVS module to reduce the computational overhead and integrate the module to the NSR pipeline, while focusing on deriving high confident planar geometries where large ambiguity issues are more likely to occur. In addition, as most methods apply [6, 19, 29], simply adding geometric loss between the prior and then estimate surface directly introduces MVS noises to over smooth the detailed structures. The same problem happens to the method of Zhang et al. [32], which employs geometric prior to supervising zero-level set from NSR. Wei et al. [20] restrain the sampling near geometric priors while defining their confidence to define the sampling range around the prior to deal with noises. However, our proposed sampling guided approach by pretraining a proposal net, which lets the network itself learn which prior to believing, thus is more robust to noises.

3 METHOD

As shown in Figure 2, Megasurf proposes a lightweight MVS module to rapidly derive confident geometries to guide the place of NSR where large shape-radiance ambiguities more likely occur. Then, a two-stage sampling guided training approach using the geometric priors is provided to improve the robustness of NSR to preserve highly detailed structures when the inevitable noises are introduced in the geometric priors.

3.1 Preliminary

Neural radiance field. By sampling 3D points from camera rays, NeRF [10] employs coordinate MLP networks to learn the density and color fields of the 3D scene. It uses volume rendering to supervise the network, which integrates the color of sampled points along the ray to render each pixel:

$$\mathbf{C}(r) = \sum_i \omega_i \mathbf{c}_i, \qquad \omega_i = T_i \alpha_i, \tag{1}$$

where $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$ is the opacity of the segment i, σ_i is the density, $\delta_i = t_i - t_{i-1}$, and t is the distance from sampling points to the ray center. $T_i = \prod_{j=1}^{i-1}(1-\sigma_j)$ is the accumulated transmittance. As the geometry of NeRF is represented by density, extracting surfaces from densities often leads to noisy results.

Neural surface reconstruction. Most NSR methods use SDF as the geometric representation instead of density in NeRFs, as the surface can be represented by the zero-level set of the signed distance function (SDF), $S = \{\mathbf{x} : f(\mathbf{x}) = 0\}$. To use volume rendering, VolSDF [26] defines the volume density function τ to map the signed distance $f(x)$ to volume density σ:

$$\tau(\mathbf{x}) = \beta^{-1} \Psi_\beta(f(\mathbf{x})), \tag{2}$$

where $\beta > 0$ is a scheduling parameters and approaches 0 during optimization,$\tau(\mathbf{x})$ is the cumulative distribution funcion (CDF) of the zero-mean Laplace distribution with scale β. Manually controlling the β allows different reconstructed cases to have the same β, so that the surface details of different cases are consistent.

Neuralangelo. Recently, multi-resolution hash encoding proposed by Muller et al. [11] is a compact feature representation that can represent large-scale scenes in unprecedented detail. Neuralangelo [8] designs a coarse-to-fine optimization scheme to reconstruct the surfaces with progressive levels of detail:

$$\gamma_l = [F_0, F_1, ..., F_{start+l}], \quad l_{start} < l < l_{max}, \tag{3}$$

where γ represents the features from hash grids, F is the features of each level of hash grid, and the coarse to fine resolution spans from level l_{start} to level l_{max}. Another important contribution is the

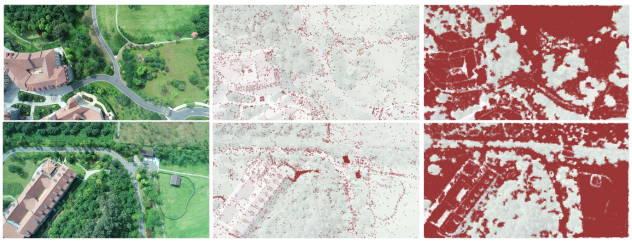
design of a numerical gradient computation to distribute the back-propagation updates to wider neighboring hash grids to improve the smoothness of surface reconstruction:

$$\nabla_x f(x) = \frac{f\left(\gamma\left(x+\epsilon_x\right)\right) - f\left(\gamma\left(x-\epsilon_x\right)\right)}{2\epsilon}, \tag{4}$$

where ϵ is the step size away from x for sampling points to calculate gradient numerically.

However, when applying it to large-scale airborne datasets, severe shape radiance often happens in the areas of heavy shadows, low textures, and illumination variations.

3.2 Lightweight PatchMatch MVS Module



(a) Images (b) SFM points (c) Confidence mask

Fig. 3. The illustration of the high-confidence region acquired by our lightweight MVS module. (a) The input images. (b) Sparse SFM points. (c) The high-confidence position which we used as the geometric prior during our NSR training.

Prelimiary of heavy PatchMatch MVS module. Commonly used PatchMatch MVS module starts from randomly initializing geometry on each pixel, and every pixel uses PatchMatch optimization to select the best geometric candidate with the smallest photo-consistency loss $\boldsymbol{E}_{NCC}$ from all the candidates propagated from its neighboring pixels. The candidates are often chosen from a window area, e.g. 11×11, centered in the pixel. Each pixel will continuously update its geometry from neighbors until it receives its accurate geometry. Due to the random initialization, pixels often require several global patch-matching optimizations to get the accurate candidate to converge.

Lightweight local propagation from SFM points. Instead, we start from high confident SFM points in each image as activate key points p_{act} to propagate the information to surrounding neighbors. We randomly select eight neighboring pixels for each p_{act} within a 11*11 pixel area as candidate key points p_{cand}. Next, we perform PatchMatch operation on the p_{cand} and corresponding neighbor pixels p_{nbr} with a distance of 3 pixels. The p_{cand} become a new p_{act} when they satisfy: 1) The depth difference between the p_{cand} and corresponding p_{act} is less than the given reconstruction accuracy. 2) The mean depth difference between the p_{cand} and its p_{nbr} is less than the given reconstruction accuracy. In this way, if the candidate key point is in the similar plane with the SFM points, it will immediately receive the accurate candidate which will be most likely selected from all the candidates with a minimal photo-consistency loss.

When the activated key point is determined, a 5*5 pixel neighbor mask is generated. The area within the mask is not sampled, which means no new key point is generated within the mask. This is not only to mitigate the incorrect propagation to the outside of the plane across the edge, but more importantly, increase the speed of propagation. When no p_{act} exists, we perform the PatchMatch operation for all pixels that are not sampled. Finally, our MVS module outputs a depth map with a mask indicating the high-confidence geometries propagated from our approach. Note that we do not require a depth fusion step to filter the depth noises, further indicating our high efficiency.

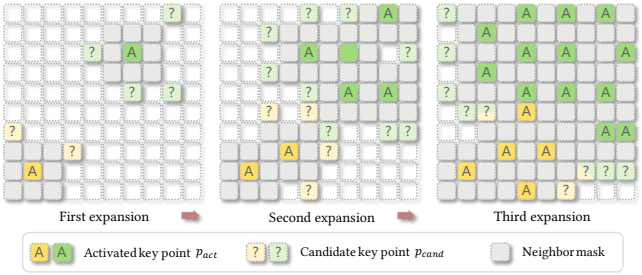


Fig. 4. The propagation strategy of our lightweight MVS module. The high-confidence geometric information is progressively propagated to its surrounding area.

3.3 Sampling-guided surface reconstruction

We propose a two-stage sampling-guided training approach. In the first stage, we train the proposal net and foreground geometry net to transform the prior geometries from our lightweight MVS module to the sampling for largely reducing ambiguities. In the second stage, we further design a non-occupancy loss using the prior geometry to reduce the ambiguity between the camera center and prior geometries when the rendering net is added for training. Here, we first explain the sampling strategy for airborne scenes combined with a proposal network.

Illustration of our sampling strategy using proposal network. As our large airborne scenes are all unbounded cases, we follow NeRF++ [31] to subdivide the scene into a foreground and background region, which employs a uniform sampling and an inverse distance sampling, respectively. We regard the minimum bounding box of the region of interest as foreground and rescale the space into a cube with a range of $[-1, 1]$. As the foreground region does not cover our cameras in this way, we start our uniform sampling within the foreground of each ray from the position where the ray intersects with the cube.

After the initial sampling, we provide a cascade proposal net $Prop$, which provides a two level hierarchical sampling procedure [10] to sample finer queries more efficiently. Our proposal net adopts multi-resolution hash encoding following a tiny MLP for more efficient feature presentation. Compared to the occupancy grid proposed by Muller et al. [11], our training approach has better stability for large complex scenes.

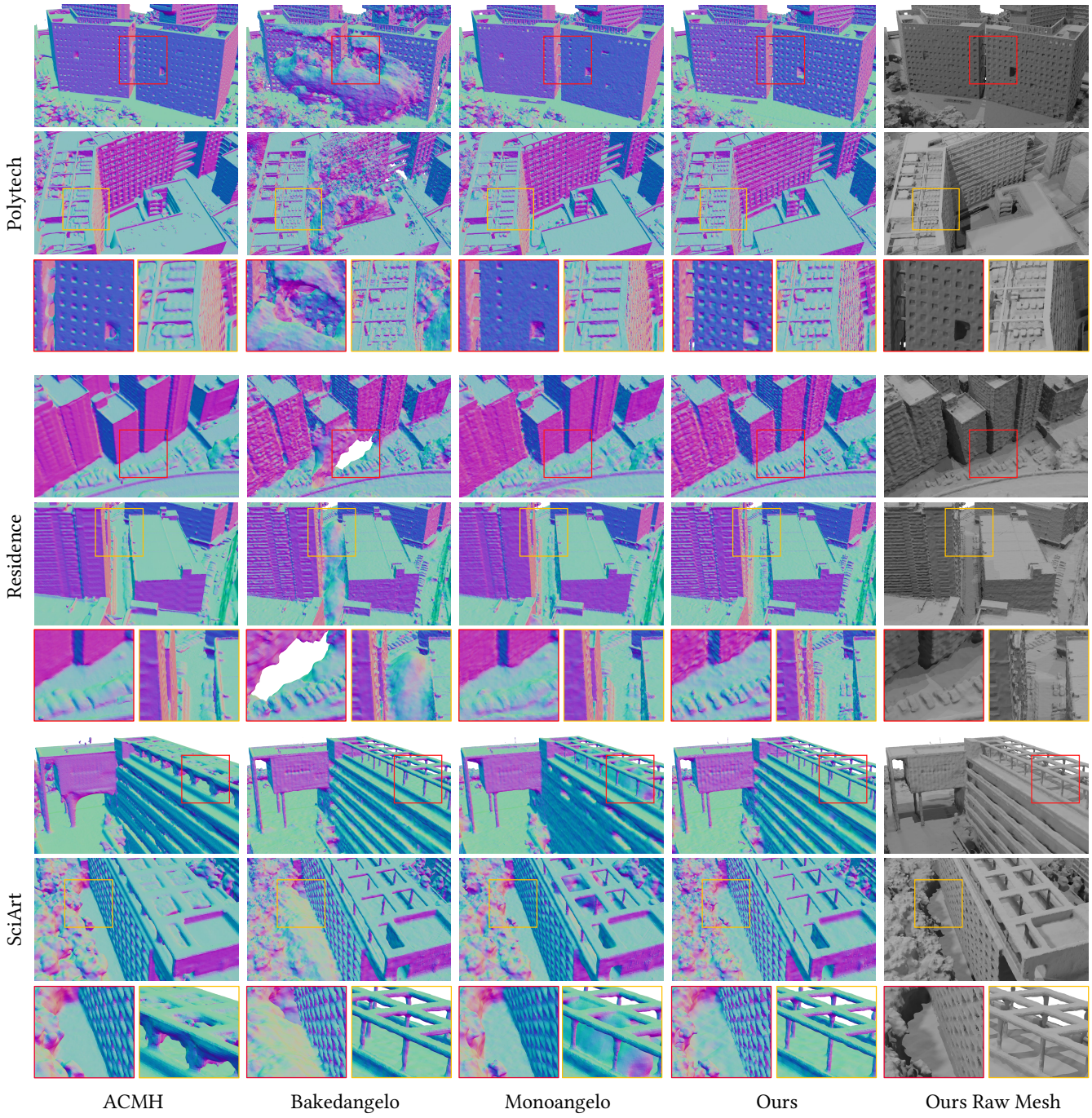


Fig. 5. Qualitative results on the Urbanscene3D dataset. MegaSurf both have the robustness to the severe shape-radiance ambiguity and preserve high-fidelity details. The meshes in the first four columns use vertex normal as its vertex color.

Stage1: pretrain proposal net using geometric prior. Different from previous work using the geometric prior to forming a depth loss in NSR, we use geometric cues to pre-train the cascade proposal net for better resistance to the geometric errors of these priors. Due to the memory and computational efficiency, the proposal net is designed with a limited resolution of hash encoding. Therefore, we jointly pretrain the geometry net, which has a much higher feature resolution, to best transform the dense geometric prior to

the sampling in the NSR. We design a cascaded proposal loss for each level of proposal net (often two levels, $Prop_0$ and $Prop_1$) and the geometric net f_g by maximize the sampling weights within the range $[t_{prior} - \epsilon, t_{prior} + \epsilon]$ of all three sampling levels mentioned above:

$$\begin{aligned} L_{prop} =& 1 - \sum_{i \in \Lambda} \omega_i^h, \\ \Lambda : \{i : t_{prior} - \epsilon < t_i <& t_{prior} + \epsilon\}, h \in Prop_0, Prop_1, f_g. \end{aligned} \tag{5}$$

We further add a curvature loss to improve the smoothness of sampling to address the noise and incompleteness of the geometric priors:

$$\mathcal{L}_{\text{curv}} = \frac{1}{N} \sum_{i=1}^{N} \left| \nabla^2 f(\mathbf{x}_i) \right|, \tag{6}$$

The overall loss designed at stage 1 is:

$$L_{stage1} = L_{curv} + L_{prop}. \tag{7}$$

Stage2: train rendering net with non-occupancy loss. When adding the rendering net into the pipeline in stage 2, we propose a non-occupancy loss to prevent sampling information carried by the pretrained proposal net from being damaged. The non-occupancy loss aims to limit the color contribution of the samples between camera center and the geometric prior indicated position. Note that there still leaves a large sampling space exceeding the prior position for the rendering net to recover high details. As shown in Figure 2, it is robust to the noise given by geometric priors, as the real surface can still be sampled. The non-occupancy loss is:

$$\begin{aligned} L_{nocc} &= || \sum_{i \in \Gamma} \omega_i c_i ||_1, \\ \Gamma :& \{i : t_i < t_{prior} - \epsilon\}, \end{aligned} \tag{8}$$

Similar to Equation 5, we also add an epsilon buffer according to accuracy of geometric priors.

Our MegaSurf also uses color, eikonal, and curvature loss as the basic loss functions, as common NSR methods often adopt. Therefore, the stage 2 training loss is defined as:

$$L_{stage2} = L_{color} + L_{curv} + L_{eikonal} + L_{nocc}. \tag{9}$$

4 EXPERIMENTS

4.1 Experimental Setup

Baselines. Our experiments are conducted on Urbanscene3D [9] dataset and the Songshanhu which is collected by our drone. Their areas are between $60000m^2 (300m \times 200m)$ and $150000m^2 (300m \times 500m)$. We divided the whole scenes into several blocks and each block covers a 150m x 150m ground region. We compare MegaSurf with ACMH [24], a traditional reconstruction method, and two NSR methods: Bakedangelo [28] and Monoangelo. Bakedangelo combines BakedSDF [27] with Neuralangelo [8] settings and has a better background modeling, which is more efficient than Neuralangelo. We migrate the key ideas of MonoSDF [29] to Bakedangelo which called Monoangelo, as the results obtained by MonoSDF are generally oversmooth.

Training and evaluation. We train MegaSurf for 200k iterations per block. The memory consumption is about 22G. After NSR training, we extract the mesh from the SDF by Marching Cube. We compared the reconstruction results of SciArt and Polytech with the LiDAR ground truth following the official evaluation protocol. It is worth noting that because only the main building has LiDAR information, the result of the numerical comparison is only used to measure the reconstruction quality of the building in the scene.

4.2 Comparisons

We developed our lightweight MVS module on ACMH software [24], which claims the equal quality, but three time speed than another popular open source software, COLMAP [13]. We project the high-confidence geometric prior obtained by our lightweight MVS module to the space to form a point cloud and compare it with ACMH.

Table 1. Quantitative results of generating the priors of our lightweight MVS module vs ACMH.

Method	$Acc_{50} \downarrow$	$Comp_{50} \downarrow$	$Overall_{50} \downarrow$	$Acc_{95} \downarrow$	$Comp_{95} \downarrow$	$Overall_{95} \downarrow$
Artsci						
ACMH	0.1566	0.1432	0.1499	0.2035	0.3663	0.2849
Ours	0.1629	0.1320	0.1475	0.2010	0.3832	0.2921
Polytech						
ACMH	0.1021	0.1043	0.1032	0.1701	0.2300	0.2000
Ours	0.1227	0.1218	0.1222	0.1937	0.2704	0.2320

Table 1 shows that our lightweight MVS module is comparable to ACMH. Only in the PolyTech dataset, our method may has less detailed structures. However, our method is more than four times faster than ACMH when only the PatchMatch step is counted (2). This matches the configuration of ACMH, which applies four times PatchMatch global sweeps on each pixel. Furthermore, ACMH requires a depth fusion step to filter noisy geometries for the final geometric prior. This step is extremely slow when a large number of images are applied due to their naive implementation, which is not counted in our table. Note that we do not need this fusion step and can also get comparable geometries.

Table 2. The time consumption for generating the priors of our lightweight MVS module vs ACMH.

	Residence	SciArt	PolyTech	Songshanhu
Image number	2581	3091	2508	738
Image size	1216×912	1216×912	1500×1000	1368×768
ACMH PatchMatch time	5891s	7274s	7641s	1200s
Ours	1133s	1357s	1460s	291s

We provide qualitative and quantitative comparisons to evaluate the performance of our method. Fig 5 and Table 3 shows the results respectively.

Bakedangelo can generate realistic details, but it suffers inherent shape-radiance ambiguity, often leading to incorrect geometry. Traditional methods such as ACMH are stable in large scene reconstruction. However, due to the large amount of noise in point clouds, the triangulation may incorrectly connect the points and

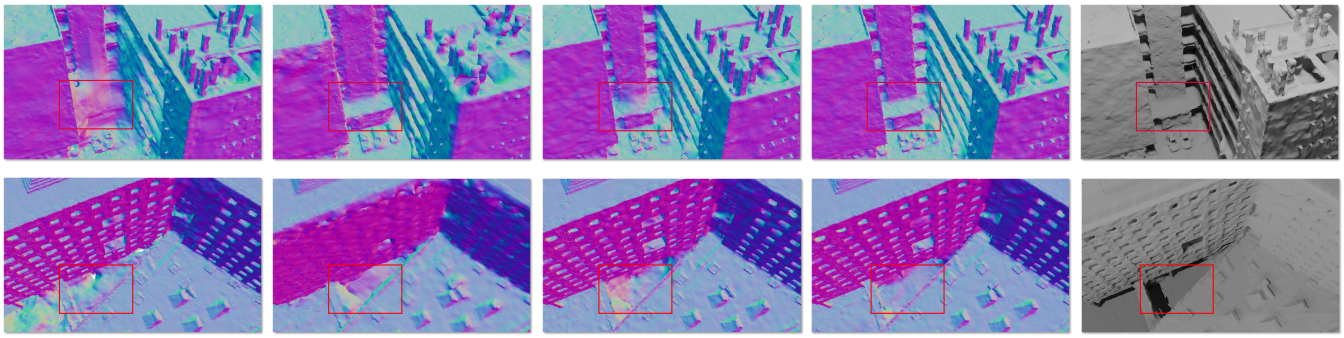


Fig. 6. Visualization results of the ablation study.

cause over-smoothing. Monoangelo takes depth priors as a regular term to guide the NSR optimization. The depth provided by MVS can help Monoangelo overcome shape-radiance ambiguity, but the noise in priors makes it difficult to reconstruct the fine geometric details. MegaSurf uses priors to train a cascade proposal net to guide the sampling position of the stage 2 of NSR optimization. The well-trained proposal net helps the NSR to use the position provided by the prior to representing the scene, thereby helping to overcome the ambiguity. At the same time, the information containing noise is not transferred to the second optimization stage, thus preventing damage to the accurate detail reconstruction. MegaSurf achieves SOTA results on the Urbanscene3D dataset.

Table 3. Quantitative evaluation of reconstruction with existing methods on the Urbanscene3D dataset. Acc_{50} represents the average of the first 50% accuracy, $Comp_{50}$ represents the average of the first 50% completeness, $Overall_{50}$ represents the mean value of Acc_{50} and $Comp_{50}$. The same for the others. MegaSurf achieves the best surface reconstruction performance.

Method	$Acc_{50}\downarrow$	$Comp_{50}\downarrow$	$Overall_{50}\downarrow$	$Acc_{95}\downarrow$	$Comp_{95}\downarrow$	$Overall_{95}\downarrow$
Artsci						
ACMH	0.1261	0.1122	0.1192	0.2958	0.5136	0.4047
Bakedangelo	0.1294	0.1147	0.1221	0.3319	0.5813	0.4566
Monoangelo	0.1380	0.1313	0.1347	0.3778	0.6152	0.4965
Ours	0.1237	0.1052	0.1145	0.2990	0.4138	0.3564
Polytech						
ACMH	0.0640	0.0874	0.0757	0.1588	0.2499	0.2044
Bakedangelo	0.1054	0.0954	0.1004	0.2989	0.3969	0.3479
Monoangelo	0.0686	0.0620	0.0653	0.1810	0.2472	0.2141
Ours	0.0729	0.0646	0.0688	0.1763	0.2086	0.1925

4.3 Ablations

We perform ablation experiments over several MegaSurf training strategies. The experiment was conducted on Urbanscene3D. The qualitative and quantitative evaluation results are shown in Fig 6 and Table 3, respectively.

Proposal net. We freeze the parameters of the proposal net (Freeze *Prop*) after stage 1 training. When the parameters of the proposal net are not affected by the rendering loss of stage 2, we found that the ambiguity is somewhat alleviated. Because the sampling position cannot change during the optimization, sample points are difficult to focus around the real surface for finer reconstruction, resulting in over-smoothing.

Occupancy loss. L_{nocc} is designed to prevent the new surface from appearing in areas where σ should be smaller according to the reliable prior information when we take rendering loss at optimization stage 2. When L_{nooc} is removed, the scene accuracy increases, but the completeness decreases. From the visualization results, we can see that the scene has some raised surfaces.

Table 4. Quantitative results of the ablation study on the Urbanscene3D dataset.

Method	$Acc_{50}\downarrow$	$Comp_{50}\downarrow$	$Overall_{50}\downarrow$	$Acc_{95}\downarrow$	$Comp_{95}\downarrow$	$Overall_{95}\downarrow$
Artsci						
Base	0.1294	0.1147	0.1221	0.3319	0.5813	0.4566
Freeze *Prop*	0.1486	0.1546	0.1516	0.3736	0.6982	0.5359
No L_{nocc}	0.1220	0.1058	0.1139	0.2980	0.4649	0.3815
Full	0.1237	0.1052	0.1145	0.2990	0.4138	0.3564
Polytech						
Base	0.1054	0.0954	0.1004	0.2989	0.3969	0.3479
Freeze *Prop*	0.0794	0.0734	0.0764	0.2218	0.3182	0.2700
No L_{nocc}	0.0693	0.0608	0.0651	0.1749	0.2120	0.1935
Full	0.0729	0.0646	0.0688	0.1763	0.2086	0.1925

4.4 Conclusion

We introduce MegaSurf, a prior guided neural surface reconstruction approach for reconstructing large-scale scenes. We propose a lightweight MVS module to progressively propagate the SFM information to its surrounding area to obtain high-confidence geometric prior, which is proven to be more than four times faster than the SOTA method. Then, we propose a two-stage sampling-guided training strategy. In the first stage, we pre-train a sampling proposal net using the priors to indicate the next stage sampling position, which helps to overcome the large geometric errors from ambiguity. In the second stage, we train the rendering net to restore the high-fidelity details while conducting a non-occupancy loss to prevent the correct prior information learned by the proposal net from being destroyed. Experiments on large-scale scene datasets show our SOTA performance.

REFERENCES

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

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

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

[4] Yasutaka Furukawa and Jean Ponce. 2010. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 8 (2010), 1362–1376. https://doi.org/10.1109/TPAMI.2009.161

[5] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2495–2504.

[6] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. 2023. StreetSurf: Extending Multi-view Implicit Surface Reconstruction to Street Views. *arXiv preprint arXiv:2306.04988* (2023).

[7] Heiko Hirschmuller. 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 807–814.

[8] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8456–8465.

[9] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. 2022. Capturing, Reconstructing, and Simulating: the UrbanScene3D Dataset. In *ECCV*. 93–109.

[10] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*. Springer, 405–421.

[11] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–15.

[12] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12179–12188.

[13] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 501–518.

[14] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. 2021. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15598–15607.

[15] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8248–8258.

[16] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. 2022. 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*. 12922–12931.

[17] Likang Wang, Yue Gong, Qirui Wang, Kaixuan Zhou, and Lei Chen. 2023. Flora: dual-frequency loss-compensated real-time monocular 3d video reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 1.

[18] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689* (2021).

[19] Yusen Wang, Zongcheng Li, Yu Jiang, Kaixuan Zhou, Tuo Cao, Yanping Fu, and Chunxia Xiao. 2022. NeuralRoom: Geometry-Constrained Neural Implicit Surfaces for Indoor Scene Reconstruction. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–15.

[20] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. 2021. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5610–5619.

[21] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. 2022. Toward practical monocular indoor depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3814–3824.

[22] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. 2022. BungeeNeRF: Progressive Neural Radiance Field for Extreme Multi-scale Scene Rendering. In *The European Conference on Computer Vision (ECCV)*.

[23] Linning Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. 2023. Grid-guided Neural Radiance Fields for Large Urban Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8296–8306.

[24] Qingshan Xu, Weihang Kong, Wenbing Tao, and Marc Pollefeys. 2022. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4945–4963.

[25] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 767–783.

[26] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems* 34 (2021), 4805–4815.

[27] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P. Srinivasan, Richard Szeliski, Jonathan T. Barron, and Ben Mildenhall. 2023. BakedSDF: Meshing Neural SDFs for Real-Time View Synthesis. *arXiv* (2023).

[28] Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. 2022. SDFStudio: A Unified Framework for Surface Reconstruction. https://github.com/autonomousvision/sdfstudio

[29] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems* 35 (2022), 25018–25032.

[30] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. 2020. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928* (2020).

[31] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020).

[32] Yongqiang Zhang, Zhipeng Hu, Haoqian Wu, Minda Zhao, Lincheng Li, Zhengxia Zou, and Changjie Fan. 2023. Towards Unbiased Volume Rendering of Neural Implicit Surfaces With Geometry Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4359–4368.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009