PGT-NeuS: Progressive-Growing Tri-Plane Representation for Neural Surface Reconstruction

Xue-Kun Xiang, Yu-Jie Yuan, Wen-Bo Hu, Yu-Tao Liu, Yue-Wen Ma, and Lin Gao, *Member, IEEE*

***Abstract*—3D reconstruction from multi-view images is a long-standing problem in computer graphic. Neural 3D reconstruction, especially NeuS and its variants, has improved reconstruction quality compared to traditional methods. However, it is still a challenge for these methods to reconstruct fine-grained geometric details since the spherical harmonic positional encoding lacks the ability to express high-frequency signals. In this paper, we propose a multi-resolution tri-plane feature encoding that leverages the detail reconstruction capabilities of high-resolution tri-plane while using the smoothness of low-resolution tri-plane to suppress high-frequency artifacts. Additionally, a progressive training strategy is introduced, gradually merging scene details from coarse to fine granularity, enhancing reconstruction quality while maintaining training stability and reducing difficulty. Furthermore, to address reconstruction challenges arising from sparse viewpoints and inconsistent lighting in image datasets, we introduce normal priors as supervision and propose consistency verification for multi-view normal priors, which assesses the accuracy of normal priors and effectively supervise the reconstructed surfaces. Moreover, we propose a perturbing and fine-tuning strategy on regions of unreliable normal priors to further improve the quality of geometric surface reconstruction.**

***Index Terms*—Neural radiance field, surface reconstruction, progressive learning, normal priors verification.**

I. Introduction

THE proposal of Neural Radiance Fields (NeRF) [1] provides a novel solution for novel view synthesis tasks. Its implicit scene representation not only enables photorealistic image rendering but can also be directly applied to 3D reconstruction tasks. The original NeRF employs volume density as geometric representation, supporting the use of Marching Cubes [2] to extract explicit geometric surfaces. However, the volume density representation lacks explicit definition of geometric surfaces, where different extraction thresholds lead to varying reconstruction results and quality discrepancies. To address this issue, NeuS [3] utilizes a multi-layer perceptron (MLP) to predict signed distance functions (SDF) and derives the transformation from SDF to density values. By employing differentiable volume rendering [4] to compute pixel values, it uses 2D images as supervision to simultaneously optimize both rendering and reconstruction. Since SDF provides an explicit definition of geometric surfaces (SDF=0), NeuS achieves significantly improved surface reconstruction quality compared to NeRF.

However, NeuS [3] adopts the spherical harmonic feature encoding from Neural Radiance Fields (NeRF) and relies solely on MLPs to learn the scene, often resulting in overly smoothed reconstructed surfaces lacking detailed modeling. Subsequent works make improvements to enhance reconstruction accuracy. HF-NeuS [5] introduces an additional offset network to model high-frequency details. Some approaches employ explicit feature encoding to improve model representational capacity, such as the tri-plane encoding [6] in PET-NeuS [7], or the multi-resolution hash grid in NeuS2 [8] and Neuralangelo [9]. However, these explicit encodings either introduce high-frequency artifacts or suffer from hash collision limitations, still exhibit deficiencies in reconstruction accuracy and quality.

On the other hand, the sparse supervision viewpoints and inconsistent lighting in image datasets make the reconstruction of certain surfaces challenging. Some works employ prior supervision to improve reconstruction quality, such as coarse geometry reconstructed by SfM methods [10], depth information [11], symmetry [12], etc. Normal priors are one of the most widely used priors, as monocular image-predicted normals offer high accuracy and low cost. NeuRIS [13] and MonoSDF [14] leverage pre-trained normal estimation networks to predict normal priors, providing supervision for reconstruction, which improves the quality of some surfaces. However, monocular image-predicted normal priors prove unreliable, and incorrect normal priors can degrade the reconstruction quality of corresponding surfaces. Although NeuRIS [13] uses a multi-view normal prior consistency check method to filter normal priors, this approach has limited ability to judge the accuracy of normal priors when dealing with surfaces with complex textures or inconsistent lighting.

In this paper, we propose a multi-resolution tri-plane feature encoding that leverages the detail reconstruction capabilities of high-resolution tri-plane while using the smoothness

Received 31 December 2024; revised 5 July 2025; accepted 8 July 2025. Date of publication 25 July 2025; date of current version 5 September 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62322210, in part by Beijing Municipal Science and Technology Commission under Grant Z231100005923031, and in part by Innovation Funding of ICT, CAS under Grant E461020. Recommended for acceptance by R. Hu. *(Corresponding author: Lin Gao.)*

Xue-Kun Xiang, Yu-Jie Yuan, Yu-Tao Liu, and Lin Gao are with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: xiangxuekun22@mails.ucas.ac.cn; yuanyujie66@gmail.com; liuyutao17@mails.ucas.ac.cn; gaolin@ict.ac.cn).

Wen-Bo Hu and Yue-Wen Ma are with ByteDance Pico, Beijing 100098, China (e-mail: huwenbodut@gmail.com; mayuewen@bytedance.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2025.3590394, provided by the authors.

Digital Object Identifier 10.1109/TVCG.2025.3590394

Fig. 1. We present PGT-NeuS, a neural reconstruction method that performs exceptionally well in multi-views reconstruction of exquisite geometric details. Leveraging the proposed progressive-growing tri-plane representation, our method can capture delicate surface details across different objects, achieving high-precision reconstruction. We highly recommend readers to zoom in and observe the details.

of low-resolution tri-plane to suppress high-frequency artifacts. Additionally, we introduce a progressive training strategy, which gradually merging scene details from coarse to fine granularity, enhancing reconstruction quality while maintaining training stability and reducing difficulty. Furthermore, to address reconstruction challenges arising from sparse viewpoints and inconsistent lighting in image datasets, we introduce normal priors as supervision and propose consistency verification for multi-view normal priors, which assesses the accuracy of normal priors and effectively supervise the reconstructed surfaces. Moreover, we propose a perturbing and refining strategy on regions of unreliable normal priors to further improve the quality of geometric surface reconstruction.

Finally, our method achieves high-precision, high-quality geometric surface reconstruction, as shown in Fig. 1.

- We propose a progressive 3D reconstruction method based on multi-resolution tri-plane encoding, which hierarchically integrates scene details from coarse to fine granularity, achieving high-precision 3D surface reconstruction.
- We propose a consistency verification method of multi-view normal priors and a perturbing and refining strategy on regions of unreliable normal priors, which effectively utilizes normal priors to supervise surface reconstruction while specifically addressing challenging-to-reconstruct surfaces to improve their reconstruction quality.
- Compared to current NeRF-based geometric reconstruction methods, our method, PGT-NeuS, can achieve more accurate geometric reconstruction while ensuring rendering quality.

II. Related Work

Neural Radiance Fields. Recently, neural rendering [15], especially NeRF [1], has attracted lots of attentions. NeRF uses multi-layer perception (MLP) to represent the radiance and density of the 3D scene, thereby achieving high-fidelity scene rendering. Explorations have been conducted in various fields around NeRF, including dynamic scene reconstruction [16], [17], [18], [19], [20], digital human modeling [21], [22], [23], [24], large-scale scene modeling [25], [26], [27], [28], editing [29], [30], [31], [32], [33], training and rendering acceleration [34], [35], [36], [37], and 3D generation [38], [39], [40], [41], which has gained widespread attention recently. However, these methods may encounter aliasing issues when rendering images of different scales. So MipNeRF [42] and MipNeRF360 [43] propose cone-casting and non-linear scene parameterization to overcome this challenge. Zip-NeRF [44] further accelerates the training and rendering with a combination of super sampling and iNGP's [37] hierarchical grids. TriMipRF [45] introduces the idea of MipMap into the tri-plane representation [6], which formulates multi-scale feature planes supporting efficient anti-aliasing training and rendering. Zip-NeRF [44] leverages ideas of super sampling, statistics, and signal processing to integrate iNGP's [37] hierarchical grids into MipNeRF360 framework which has the ability to fix zipper-like aliasing artifacts. Our method draws inspiration from TriMipRF [45] and further explores fine-grained geometric reconstruction based on its unexpectedly good geometric reconstruction results.

Surface Reconstruction from NeRF. As a leading work, NeuS [3] proposes a neural surface reconstruction based on SDF representation in the framework of NeRF. Various priors [10], [11], [12], [46], [47] have been introduced to improve the accuracy of geometric reconstruction. Normal prior [13], [14] is one of them, which is estimated by a pre-trained network and provides additional supervision to the SDF prediction. The same idea can be extended to indoor scenes, where the Manhattan World hypothesis constrains the normal direction to follow certain rules [48]. HF-NeuS [5] proposes an additional offset network to model high-frequency details and designs an adaptive optimization strategy that makes the training process focus on improving the regions near the surface. PET-NeuS [7] employs a positional encoding strategy in conjunction with tri-plane features [6] to mitigate noise interference. Neuralangelo [9] introduces SDF into iNGP method [37] and solves the problem of surface roughness using the numerical gradient. NeuS2 [8] proposes an efficient method for calculating spatial gradients of mixed representations, making it possible to apply multi-scale hash encoding to SDF-based volume rendering processes. More recently, LoD-NeuS [49] proposes a multi-scale and multi-convoluted tri-plane representation for fine-grained geometric reconstruction, but tri-planes with different scales are optimized simultaneously, so heavy optimization may lead to a decrease in reconstruction quality. Our method proposes the progressive-growing tri-plane representation which can ease the training burden for accurate reconstruction, with compatible progressive blending and supervision.

Explicit Representations in NeRF. The original NeRF is a pure implicit representation while subsequent work has explored incorporating explicit feature representation to improve the model capabilities and accelerate training and inference. The most

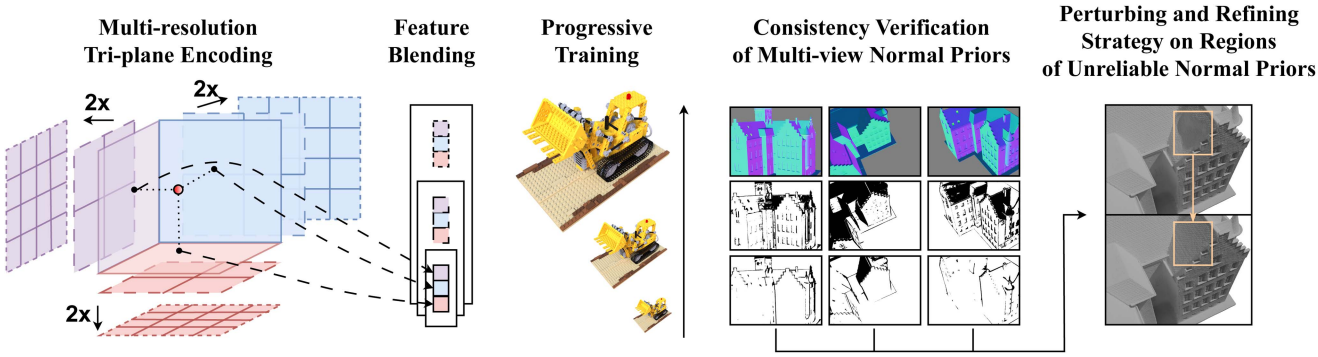


Fig. 2. Pipeline of our method. We present a multi-resolution tri-plane encoding that combines high-resolution detail reconstruction with low-resolution artifact suppression. Our progressive training strategy gradually refines scene details from coarse to fine. We also introduce normal-prior supervision with multi-view consistency verification to assess and enhance normal accuracy. A perturbing and refining strategy further improves geometry in unreliable normal regions.

common choices are the octree of voxel [36], the multi-resolution hash table [37] and the tri-planes [6]. Recently, some more compact explicit representations have been proposed, such as vector-matrix representation [50] and even tri-vector [51]. We further propose a progressive-growing tri-plane representation to achieve the reconstruction of fine geometric details.

III. Method

The overall pipeline of our method is shown in Fig. 2. We propose a multi-resolution tri-plane feature encoding that leverages the detail reconstruction capabilities of high-resolution tri-plane while using the smoothness of low-resolution tri-plane to suppress high-frequency artifacts. Additionally, we introduce a progressive training strategy, which gradually merging scene details from coarse to fine granularity, enhancing reconstruction quality while maintaining training stability and reducing difficulty. Furthermore, to address reconstruction challenges arising from sparse viewpoints and inconsistent lighting in image datasets, we introduce normal priors as supervision and propose consistency verification for multi-view normal priors, which to assess the accuracy of normal priors and effectively supervise the reconstructed surfaces. Moreover, we propose a perturbing and refining strategy on regions of unreliable normal priors to further improve the quality of geometric surface reconstruction.

A. Preliminaries

We first briefly review the preliminaries of Neural Radiance Fields(NeRF) [1], Neural Implicit Surfaces(NeuS) [3], and tri-plane encoding [6].

NeRF. NeRF [1] use a fully connected network to reconstruct a scene, where the input are the sampled points in the scene and the view direction, and the output is the volumetric density and color values. NeRF samples three-dimensional spatial points in the scene by emitting rays from the camera center to the image pixels and accumulate the outputs of each point into pixel values using volumetric rendering [4]. The formula for volumetric rendering is:

$$C(\mathbf{r}) = \sum_{i=1}^{N} \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \left(1 - \exp\left(-\sigma_i \delta_i\right)\right) \mathbf{c}_i, \quad (1)$$

where δ is the distance between adjacent sampling points, σ and $\mathbf{c}$ are the predicted density and color, and $C(r)$ is the accumulated color of the ray $\mathbf{r}$, corresponding to the pixel's color value. The training of Neural Radiance Fields is supervised by the difference between the accumulated color of the pixel and the true color.

NeuS. NeuS [3] introduces a Signed Distance Function (SDF) to improve the quality of surface reconstruction based on Neural Radiance Fields. NeuS derives the conversion formula from SDF to volume density, incorporating the SDF as an optimizable parameter in volumetric rendering. The conversion formula is as follows:

$$\rho(t) = -\frac{d\Phi_s(S)}{dt} / \Phi_s(S) \quad (2)$$

where Φ_s is a sigmoid function with coefficient s, S is the SDF value predicted by model, and ρ is the opaque density proposed by NeuS, equivalent to the density in Neural Radiance Fields. NeuS outperforms Neural Radiance Fields in terms of mesh reconstruction quality due to the clear definition for surfaces(SDF=0).

Tri-plane encoding. Tri-plane encoding is a popular explicit feature representation introduced in EG3D [6] and combined with Neural Radiance Fields. The original Neural Radiance Fields uses spherical harmonic positional encoding along with a multilayer perceptron to query the attributes of sampled points, which not only results in inefficient training and inference but also lack the ability to model high-frequency details. Tri-plane Encoding explicitly stores features in three orthogonal planes, projecting spatial points onto each plane and obtaining features through bilinear interpolation. The interpolated features are then aggregated, and a small multilayer perceptron is used to obtain the attributes of the sampled points. By explicitly storing features, Tri-plane Encoding possesses strong detail representation capabilities.

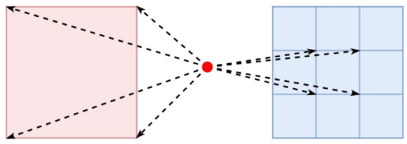


Fig. 3. For each sampled point, the unit grids where its projected point is located on different resolution planes (left is low, right is high) will be optimized in the same training step. This implies that local spaces at different ranges will jointly influence the sampled points.

B. Multi-Resolution Tri-Plane Feature Encoding

Increasing the resolution of the tri-planes is a straightforward way to improve reconstruction quality since higher-resolution tri-planes can model finer-scale spatial features.

However, high-resolution tri-plane feature sampling focuses on smaller-scale spatial regions, causing the model's fitting to become overly localized, which introduces high-frequency artifacts in the reconstructed surfaces, degrading the quality of smoother regions. Conversely, low-resolution tri-planes do not produce such high-frequency artifacts due to their modeling of larger local regions, but their weaker representational capacity results in reconstructed geometry lacking finer details.

To fully leverage the advantages of both, we propose a multi-resolution tri-plane encoding method, as illustrated in Fig. 3. We project spatial points simultaneously onto tri-plane feature planes of different resolutions, query explicit features across varying local scales, and blend them as input for subsequent networks. This feature sampling method integrates information from different spatial scales, not only enhancing detail reconstruction but also suppressing the occurrence of high-frequency artifacts.

C. Progressive Growing Training

Using multi-resolution tri-planes for joint training directly will lead to model convergence difficulties, mainly because: 1) Simultaneous optimization of multi-resolution tri-planes will significantly increase model parameters, causing exponential growth in optimization space complexity; 2) Gradient competition between different resolution features will interfere with optimization direction, making it difficult for the model to coordinate balanced development across different levels. Meanwhile, if high-resolution images are used directly for supervision, dense pixel sampling will cause the model to focus excessively on local detail features during early training stages, resulting in insufficient learning of macro structures and suboptimal geometric reconstruction results.

To address these issues, we propose a progressive training strategy. We first use low-resolution image supervision and corresponding low-resolution tri-planes for training to quickly construct the macro structure of the scene. As training progresses, higher-resolution tri-planes are initialized by upsampling features from the currently trained highest-resolution tri-plane to ensure effective inheritance of scene information, while gradually introducing higher-resolution image supervision. During this process, we design a progressive blending mechanism controlled by blending weight α, enabling newly added high-resolution tri-plane features to integrate smoothly into training, which both maintain the stability of learned scene structures and achieve progressive detail enhancement from low to high resolution. Let n denote the currently added tri-plane in training, where the minimum-resolution tri-plane is defined as $n = 0$. When $n = 1$, representing the first blending of high- and low-resolution tri-planes, the feature blending formula is as follows:

$$T_1 = t_0(1 - \alpha) + t_1\alpha \tag{3}$$

where T_n denotes the sum of feature vectors from all currently blended tri-planes up to the nth layer (in the current formula $n = 1$), and t_n indicates the feature vector of the nth tri-plane. The blending weight α gradually increases from 0 to 0.5, progressively incorporating new high-frequency features into the model training. When the $n = 1$ tri-plane is added, the numerical range of feature vectors maintains T_1 comparable to T_0, ensuring the network input remains stable without significant fluctuations. For cases where $n > 1$, the blending formula is as follows:

$$T_n = T_{n-2} + t_{n-1}(1 - \alpha) + t_n\alpha \tag{4}$$

Each newly added high-resolution tri-plane must undergo weighted blending with the highest-resolution tri-plane from the existing combination, since its feature vectors are initialized through upsampling of the latter. This blending strategy enhances model stability and progressively improves learning capability.

D. Consistency Verification of Multi-View Normal Priors

In neural implicit surface reconstruction, 2D image supervision as the primary loss term struggles to reconstruct surfaces with sparse viewpoints or inconsistent lighting. While introducing normal priors can improve such surfaces' quality, their effectiveness depends on the accuracy of normal priors: even state-of-the-art methods like StableNormal [52] exhibit multi-view inconsistencies due to their monocular-based prediction and tend to produce erroneous normals in regions with similar textures but distinct true normals. Directly using such error-containing normal priors degrades reconstruction quality, while their over-smoothed nature further harms geometric detail recovery.

NeuRIS [13] filters unreliable normal priors through multi-view consistency verification: it maps depth and normals from the current view to neighboring views via homography transformation, then evaluates prior reliability using Normalized Cross-Correlation (NCC) of corresponding pixels in the original 2D images. However, this method suffers from local texture variations, approximate homography mapping, and depth errors—even correct normals may be misjudged due to multi-view pixel value discrepancies. Although NeuRIS applies grayscale conversion and denoising to reduce texture interference, lighting variations still disrupt NCC consistency, limiting its reliability assessment in scenes with complex textures or inconsistent illumination.

We adopt normal priors images instead of real images or NeuRIS's grayscale denoised versions for pixel queries, offering two key advantages:

1) *Smoothness robustness:* The inherent continuity of normal priors avoids texture sampling errors in homography transformations.
2) *Physically-grounded thresholding:* By replacing NCC evaluation with multi-view Normal Angle Loss (NAL), we transform threshold determination into an angular constraint on normals (e.g., angular difference between adjacent views $<$ threshold). This eliminates texture/illumination-dependent empirical thresholds (e.g., NeuRIS's 0.66) and simplifies verification procedures.

For a given camera view i and its adjacent view j, the NAL calculation formula for any pixel patch P between the two cameras can be expressed as:

$$\mathrm{NAL}_{ij}(P,n,d)=\cos\left(\frac{\sum_{q\in P}N_i(q)}{N_P},\frac{\sum_{q\in P}N_j(H_{n,d}(q))}{N_P}\right) \tag{5}$$

where n and d are the rendered normal and depth of pixel q in the ith camera, N denotes normal prior images, N_P is the number of pixels in the pixel block P, and H represents the homography transformation from camera i to camera j, which can be calculated using the following formula:

$$H_{n,d}=K_j\left(R_jR_i^{-1}-\frac{(t_i-t_j)n^T}{dv^Tn}\right)K_i^{-1} \tag{6}$$

where K, R, and t are the camera's intrinsic parameter matrix, rotation matrix, and translation vector, respectively.

We evaluate normal reliability through NAL, the angular difference between predicted normals for the same surface region across adjacent views. Specifically, for a pixel patch P in a certain view, if its NAL calculated with any adjacent view is below threshold θ, the normal priors are considered consistent between these two views. If P satisfies this condition with most adjacent views in a given camera, it is marked as reliable. All normal priors are filtered through this criterion to generate local reliability masks, ensuring only regions of reliable normal priors supervise surface reconstruction and suppressing erroneous priors. This mechanism leverages the physical property of normal consistency in the world coordinate system, transforming multi-view geometric constraints into quantifiable angular thresholds without relying on image textures or empirical parameters.

To mitigate the adverse effects of normal priors smoothness on high-precision detail reconstruction, we apply normal supervision only during the early and middle training stages to accelerate macro-structure learning. As the progressive strategy in Section III-C enhances the network's representational capacity and strengthens high-frequency image signals, the weight of normal priors supervision gradually decays to reduce smoothing constraints on detail optimization. As demonstrated in Fig. 4, our method effectively overcomes smoothing interference and achieves high-precision geometric detail reconstruction.

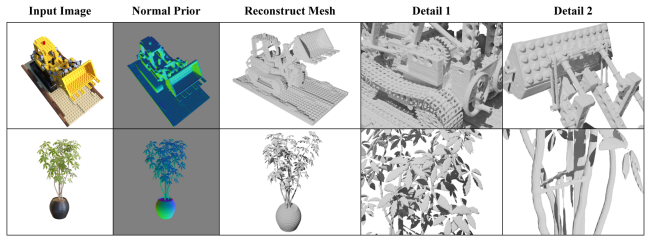


Fig. 4. Our method effectively overcomes smoothing interference and achieves high-precision geometric detail reconstruction.

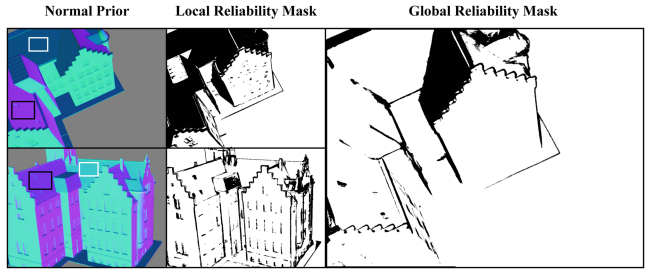


Fig. 5. Local and global reliability mask. If a region is identified as unreliable in the local reliability masks of all cameras, it will be marked (black pixels) in the global reliability mask.

E. Perturbing and Refining Strategy on Regions of Unreliable Normal Priors

While multi-view normal consistency verification provides reliable supervision for under-reconstructed regions, underfitting persists in areas with unreliable normal priors. Experiments reveal that insufficient viewpoint coverage leads to concave surface artifacts due to supervision scarcity, which become irreversible when normal priors simultaneously fail in these regions. To address this, we propose a perturbing and refining strategy targeting regions of unreliable normal prior, actively guiding the model to re-optimize these erroneous geometric structures through controlled intervention.

Local regions of unreliable normal priors are labeled as unreliable(black) in the local reliability mask, and the criterion for determining global regions of unreliable normal priors is that a region is labeled as unreliable in all local reliability masks across viewpoints. For instance, in Fig. 5, the sloped roof regions (black and white boxes) are marked as unreliable (black pixels) in some local reliability masks(top-center) but may be labeled as reliable (white pixels) in others (bottom-center). The global reliability mask identifies regions that are consistently unreliable across all views (marked black) by aggregating local reliability masks from all viewpoints, thereby preventing global errors caused by single-view misjudgments.

Global regions of unreliable normal priors lack valid normal supervision across all cameras. When such regions suffer reconstruction difficulties due to sparse viewpoints or lighting variations, conventional optimization struggles to correct them, leading to underfitting. To address this, our method extracts unreliable pixel block P_u from the global reliability mask (black pixels) and applies a perturbation loss to them:

$$L_{pert}(P_u)=\sum_{q\in P_u}F(x(q,d))G_{P_u}(q)+ \tag{7}$$

$$\sum_{q\in P_u}\left(|C_r(q)-\bar{C}_r(P_u)|^2+|N_r(q)-\bar{N}_r(P_u)|^2\right) \tag{8}$$

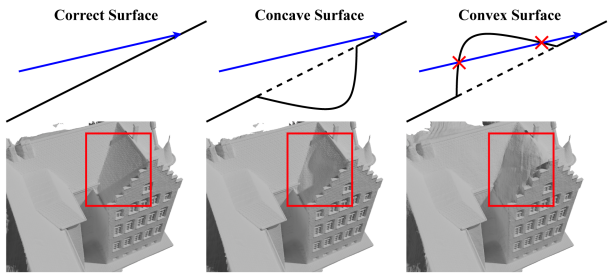


Fig. 6. The illustration of concave and convex surfaces. Concave surfaces fail to establish occlusion relationships with camera rays, making them difficult to fit using only 2D image supervision; whereas convex surfaces can effectively occlude rays, generating substantial corrective gradients that accelerate convergence in these regions.

where C_r and N_r represent the rendered color and normal of pixel q in pixel block P_u, F denotes the signed distance field function, and $x(q, d)$ represents the spatial point corresponding to the rendering depth d of pixel q. The variable G_{P_u} is a Gaussian distribution-based weight with the same size as pixel block P_u, with the maximum weight located at the center of the block.

The first term of the loss penalizes the SDF values of pixel block P_u, pushing the SDF at the corresponding rendering depth toward negative values. This causes the surface of unreliable regions to expand outward and form a convex shape. The Gaussian distribution-based weight G_{P_u} assigns higher weights at the center of the block to control the shape of convex. The second term of the loss is used to constrain the smoothness of color and normals within that pixel block, ensuring that the convex shape remains continuous and smooth.

Perturbation loss L_{pert} forcibly drives regions of unreliable normal priors to protrude outward regardless of their reconstruction quality. Applied alternately with the original training process during later stages to prevent network collapse, L_{pert} is deactivated after multiple perturbing iterations. Training then reverts to the original loss function, focusing optimization on those convex surfaces. During re-optimization, the model increases ray sampling density (targeting the convex surfaces area) to accelerate surface refining.

The rationale for perturbing surfaces in global regions of unreliable normal priors into convex shapes lies in the following: Surfaces in regions of reliable normal priors can be accurately reconstructed with normal supervision, while regions of unreliable normal priors often develop concave artifacts due to lack of supervision (see Fig. 6, middle). The black lines in the legend represent the reconstructed surfaces, and the blue arrows indicate camera rays from the training set. The concave surface fails to occlude training rays (blue arrows with no intersections), making the 2D image loss unable to detect geometric errors. In contrast, the convex surface (right) amplifies image loss gradients by occluding rays (intersections between black curves and blue arrows), forcing the model to actively correct geometry. Although regions of unreliable normal priors may contain correctly reconstructed surfaces, which will be disturbed to convex shapes but can be recovered via refining, while concave erroneous surfaces cannot self-correct due to gradient absence, necessitating forced convex perturbation to activate optimization pathways. This strategy leverages occlusion physics to find non-optimizable surfaces and improve the quality of their reconstruction.

IV. Experiments

A. Experimental Settings

In this section, we will introduce the hyper-parameter settings of the network, the information of datasets, and several comparative methods.

Implementation details. Our implementation is based on NeuS [3] with a batch size of 512, using identical SDF and color network configurations. We employ 4 levels of tri-planes with resolutions ranging from low to high as 200, 400, 800, and 1600 respectively. Mesh extraction is performed at 512 resolution. Our training proceeds for 400 k steps, with tri-plane levels and training image resolution progressively increased at 20 k, 40 k, and 60 k iterations. We employ color loss and eikonal loss proposed in NeuS, and further use our normal prior loss and perturbation loss for supervision. We use normal prior loss for supervision from the beginning, and compute local reliability masks at 100 k iteration then use it to filter normal prior, and gradually reduce the weight of normal prior loss from 1 to 0 to diminish its impact on detail reconstruction from 100 k to 200 k iterations. The threshold θ is set to 0.9848, which is approximately equals to $\cos 10^\circ$, to classify the reliability of normal prior(sample patches having NAL values above θ are considered reliable). We conduct 10 rounds of perturbing and refining for regions of unreliable normal priors starting from 300 k iteration, each comprising 2 k iterations (1 k standard training + 1 k perturbing and refining for regions of unreliable normal priors). The remaining 80 k iterations proceed with standard training, in which pixels in regions of unreliable normal priors has higher probability to be sampled. The entire training process requires approximately 16–20 hours with peak VRAM usage around 23 GB.

Datasets. We perform experiments on DTU [53] and NeRF-synthetic [1] datasets, each scene of the former has 49 or 64 images with 1200×1600 resolution, and each of the latter has 100 images of 800×800 resolution instead. We construct training images at different resolutions through down-sampling 3 times to match 4 levels of the multi-resolution tri-planes. Each subsequent sampling resolution is half of the original resolution.

Baselines. We choose NeuS [3], NeuS2 [8], PET-NeuS [7], LoD-NeuS [49], NeuRIS [13] and 2DGS [54] methods as the comparative methods. These methods are state-of-the-art neural reconstruction methods based on SDF (or Gaussian Splatting [55]), some of them also employing tri-plane representation. Specifically, NeuRIS is trained using normal priors, which other works do not have. We choose the Peak Signal-to-Noise Ratio (PSNR) as the quantitative metric for evaluating the rendering quality and Chamfer Distance (CD) for the accuracy of the reconstructed mesh.

B. Comparison

In all comparative tables presented in this paper, red and orange highlight the first and second best-performing results, respectively.

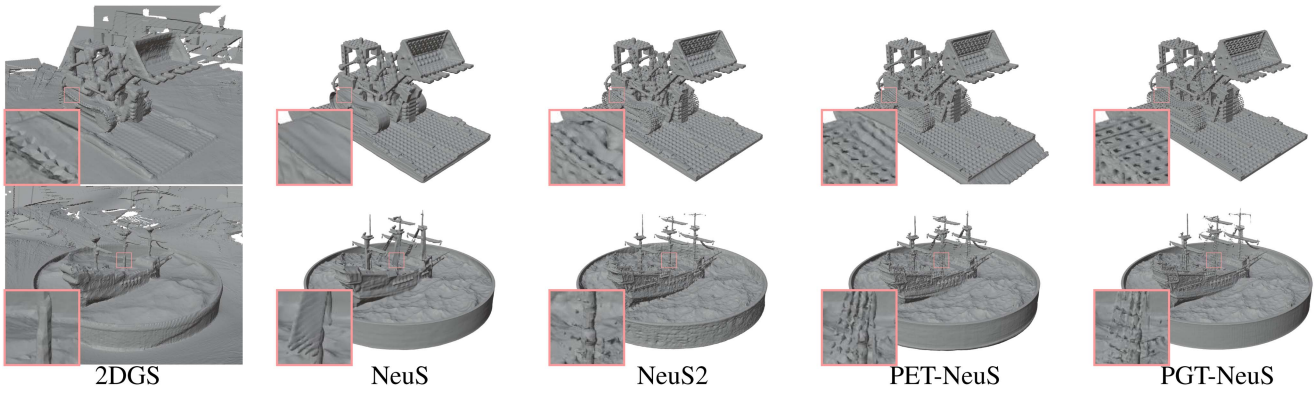


Fig. 7. Visual comparisons on the NeRF-Synthetic dataset. Our method achieves finer geometric modeling capability, accurately reconstructing intricate structures such as the voids in bulldozer tracks and the fine details of sailboat masts.

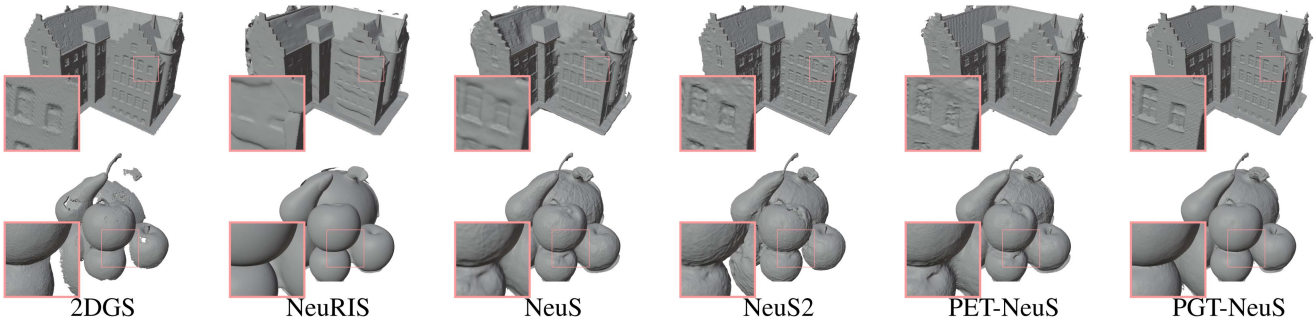


Fig. 8. Visual comparisons on the DTU dataset. Beyond fine geometric reconstruction, our method achieves superior surface smoothness in areas like building facades and fruit surfaces.

TABLE I
QUANTITATIVE COMPARISONS OF RECONSTRUCTION QUALITY ON DTU DATASET

CD	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
2DGS	0.48	0.91	0.39	0.39	1.01	0.83	0.81	1.36	1.27	0.76	0.70	1.40	0.40	0.76	0.52	0.80
NeuRIS	1.52	1.64	1.78	2.01	2.88	1.48	1.35	2.13	2.63	1.79	1.86	2.11	1.73	1.70	1.42	1.87
NeuS	1.00	1.37	0.93	0.43	1.10	0.65	0.57	1.48	1.09	0.82	0.52	1.20	0.35	0.49	0.54	0.84
NeuS2	0.56	0.76	0.49	0.37	0.92	0.71	0.76	1.22	1.08	0.63	0.59	0.89	0.40	0.48	0.55	0.70
PET-NeuS	0.56	0.75	0.68	0.36	0.87	0.76	0.69	1.33	1.08	0.66	0.51	1.04	0.34	0.51	0.48	0.71
LoD-NeuS	0.65	0.91	0.37	0.48	1.04	0.86	0.82	1.21	0.95	0.69	0.56	1.30	0.41	0.58	0.56	0.76
Ours	0.57	0.83	0.43	0.32	0.86	0.58	0.56	1.35	0.93	0.76	0.49	1.12	0.32	0.47	0.42	0.66

TABLE II
QUANTITATIVE COMPARISONS OF RENDERING QUALITY ON DTU DATASET

PSNR	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
NeuS	28.20	27.10	28.13	28.80	32.05	33.75	30.96	34.47	29.57	32.98	35.07	32.74	31.69	36.97	37.07	31.97
NeuS2	28.44	27.14	29.70	29.67	31.75	27.83	24.84	31.24	26.86	30.57	26.05	28.93	28.98	27.82	32.48	28.82
PET-NeuS	30.15	27.69	29.17	29.55	33.78	33.65	30.95	35.21	29.53	33.43	36.58	33.54	32.34	38.50	37.61	32.79
Ours	35.34	30.67	34.87	33.78	33.51	31.76	29.58	34.13	33.25	36.01	36.69	31.50	33.54	36.69	37.41	33.92

We present visual comparisons of different methods on two scenes from the NeRF-synthetic dataset [1] in Fig. 7. Our method demonstrates superior reconstruction of fine-grained geometric details, benefiting from the multi-resolution tri-plane representation that captures intricate structures while integrating broader local spatial contexts. Further comparisons on the DTU dataset [53] (Fig. 8) reveal our method's capability to reconstruct complex real-world geometries, such as roof tiles and windows, where baseline approaches exhibit varying degrees of detail loss. For instance, NeuS and 2DGS produce oversmoothed surfaces, while NeuRIS—despite using normal priors—fails to preserve fine details due to its inconsistent reliability verification of normal priors, leading to artifacts from erroneous supervision.

Quantitative evaluations in Tables IV and I report Chamfer Distance (CD) values on NeRF-synthetic and DTU, respectively, confirming our method's geometric accuracy. Additionally, Table II shows our approach achieves higher PSNR than competitors on the DTU dataset, as improved geometry directly enhances rendering fidelity.

We further compare our method with LoD-NeuS [49]. Since its code is unreleased, Fig. 9 provides visual comparisons on two NeRF-synthetic examples, where LoD-NeuS meshes

TABLE III
QUANTITATIVE ABLATION STUDY OF CHAMFER DISTANCE ON DTU DATASET

CD	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
ST	0.61	0.97	1.28	0.43	1.00	0.76	0.83	1.35	1.10	0.83	0.53	1.35	0.35	0.57	0.49	0.83
MT	0.61	1.00	1.19	0.42	0.95	0.78	0.80	1.34	1.02	0.79	0.54	1.28	0.33	0.53	0.47	0.80
MT+Pt	0.61	1.00	1.20	0.37	0.92	0.70	0.61	1.35	1.03	0.84	0.51	1.27	0.35	0.52	0.45	0.78
MT+Pt+Pr	0.61	1.00	1.23	0.37	0.91	0.63	0.57	1.36	1.12	0.78	0.51	1.28	0.32	0.47	0.45	0.77
MT+Pt+Pr+MVN(NCC)	0.63	0.87	0.55	0.48	0.94	0.67	0.57	1.45	1.09	0.75	0.54	1.22	0.36	0.56	0.59	0.75
MT+Pt+Pr+MVN(NAL)	0.61	0.83	0.50	0.35	0.88	0.65	0.57	1.46	1.03	0.74	0.51	1.19	0.33	0.49	0.45	0.70
weighted-unr	0.63	0.93	0.48	0.47	1.08	0.75	0.92	1.54	1.09	0.89	0.68	1.83	0.42	0.60	0.91	0.88
Ours	0.57	0.83	0.43	0.32	0.86	0.58	0.56	1.35	0.93	0.76	0.49	1.12	0.32	0.47	0.42	0.66

We propose a single tri-plane NeuS, labeled as 'ST', and progressively test the ablation studies of multiple techniques, including multi-resolution tri-plane 'MT', progressive tri-plane feature blending 'Pt', and progressive training set resolution growth 'Pr', and the consistency verification of multi-view normal priors 'MVN' based on both 'NCC' in NeuRIS and 'NAL' in our work, and finally 'Ours' denotes our full version.

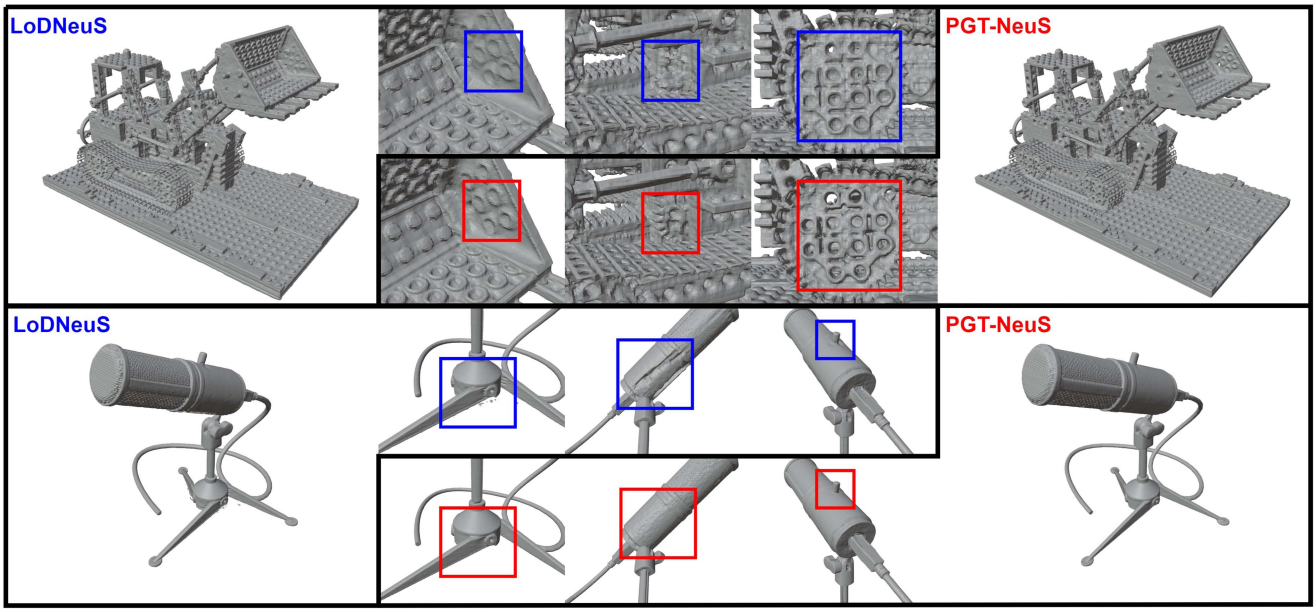


Fig. 9. Comparisons with LoD-NeuS. Our progressive-growing tri-planes capture more geometric details, such as the smooth surface of the microphone and the gear details within Lego.

are provided by its authors. Although LoD-NeuS employs a multi-scale tri-plane representation for high-fidelity geometry reconstruction, its simultaneous optimization of all scales introduces training instability, degrading final results. In contrast, our progressively growing tri-plane representation and compatible progressive feature blending strategy enhance training stability, enabling better geometric detail capture and superior reconstruction fidelity, as quantitatively validated in Table I.

Additional comparisons with NeuRIS on indoor datasets (Fig. 10)—where low-frequency textures favor NeuRIS's normal prior reliability—demonstrate our advantages. The upper three rows show NeuRIS reconstructions, while the lower three rows present our results. Our method outperforms NeuRIS in both fine structures and surface regularity, with quantitative improvements confirmed in Table V.

C. Ablation Study

To validate the effectiveness of each module in our proposed method, we conduct ablation experiments. The quantitative comparisons on the DTU dataset of all ablation experiments are shown in Table I. We propose a single tri-plane NeuS, labeled as 'ST', as a baseline, and progressively test the ablation studies of multiple techniques, including multi-resolution tri-plane 'MT', progressive tri-plane feature blending 'Pt', and progressive training set resolution growth 'Pr', and the consistency verification of multi-view normal priors 'MVN' based on both 'NCC' in NeuRIS and 'NAL' in our work, and finally our full version.

Multi-resolution Tri-plane Feature Encoding. We first verify the advantages of multi-resolution tri-plane encoding over single-resolution tri-plane encoding. For univariate comparison,

TABLE IV
QUANTITATIVE COMPARISONS OF CHAMFER DISTANCE ON NERF-SYNTHETIC DATASET

CD	chair	drums	ficus	hotdog	lego	materials	mic	ship	Mean
2DGS	0.110	0.037	0.061	0.099	0.118	0.019	0.023	0.057	0.066
NeuS	0.014	0.258	0.033	0.032	0.026	0.014	0.018	0.073	0.059
NeuS2	0.053	0.006	0.022	0.064	0.029	0.013	0.015	0.465	0.083
PET-NeuS	0.011	0.089	0.025	0.049	0.028	0.018	0.020	0.263	0.063
Ours	0.010	0.007	0.014	0.020	0.003	0.009	0.016	0.070	0.018

TABLE V
QUANTITATIVE COMPARISONS ON INDOOR DATASETS [13]

CD	case1	case2	case3
NeuRIS	1.53	3.26	$0.26\text{x}10^{-2}$
OurS	0.77	2.85	$0.18\text{x}10^{-2}$

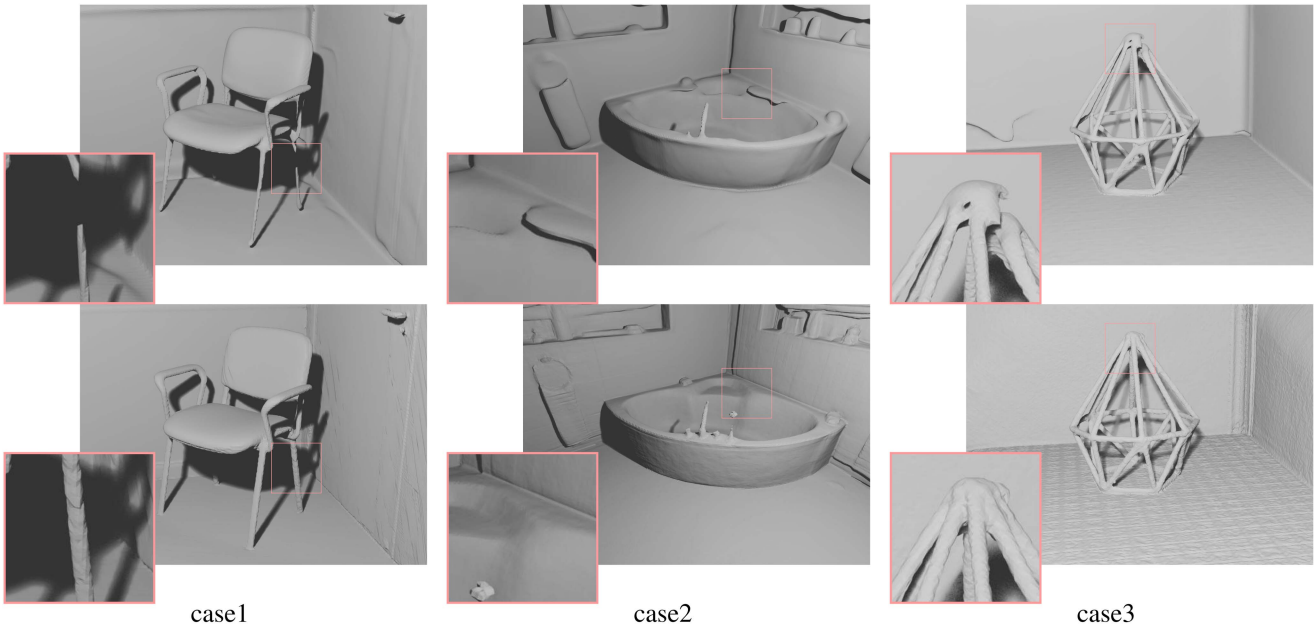


Fig. 10. Visual comparisons on the indoor dataset. The above and below show the NeuRIS and Ours reconstruction results, respectively. Our method has enhanced reconstruction quality for both intricate geometries and smooth surfaces.

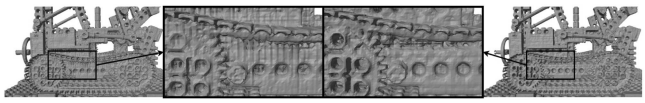


Fig. 11. Visual comparison of reconstruction between single-resolution and multi-resolution tri-planes. The left is single-resolution, and the right is multi-resolution. It can be seen that multi-resolution tri-planes can eliminate parallel patterns caused by the high discrete characteristic.

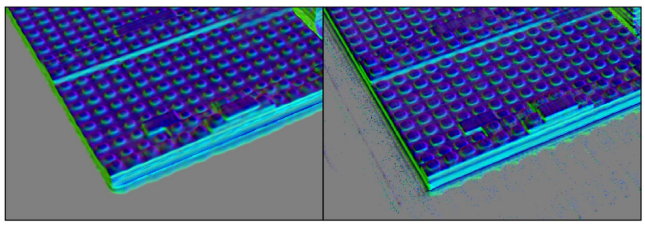


Fig. 12. Normal visualization at the iteration when the network first grows to the highest resolution of tri-plane. Our progressive-growing tri-plane representation (left) optimizes the space from coarse to fine, which can avoid the occurrence of high-frequency noise (right).

we do not adopt the progressive-growing strategy in multi-resolution tri-plane encoding. Fig. 11 shows that the multi-resolution tri-planes eliminate high-frequency artifacts caused by the discreteness of the high-resolution tri-plane, as the low-resolution tri-plane focuses on a larger local area, resulting in a smoother surface. The quantitative results presented in Table I are labeled as 'ST' and 'MT'. Compared to single tri-plane 'ST', our multi-resolution tri-plane 'MT' achieves a smaller CD value, further demonstrating its effectiveness.

Progressive-growing. Then we show the differences between our progressive growing training strategy and the ordinary network structure. Fig. 12 shows that our progressive-growing strategy can ensure our tri-plane effectively fits the scene from coarse to fine in the early training, without generating a large amount of noise, which is helpful for the overall training of the network. The quantitative results presented in Table I are labeled as 'MT+Pt' and 'MT+Pt+Pr'. The result of 'MT+Pt' shows a certain improvement compared to 'MT', indicating that our tri-plane feature progressive blending can also significantly enhance reconstruction on its own. The result of 'MT+Pt+Pr' shows further improvements, indicating that progressive increasing of training dataset resolution also has a positive effect on the reconstruction.

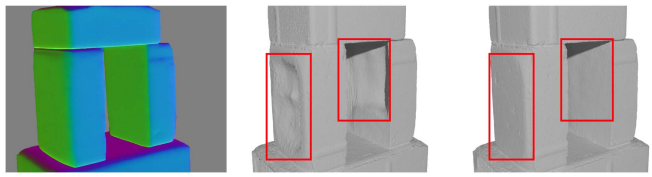


Fig. 13. The reliable normal prior based on multi-view consistency enhances reconstruction quality. From left to right, the images show the predicted normal prior, the reconstruction result without the normal prior, and the reconstruction result with the normal prior.

Consistency verification of multi-view normal priors. We employ consistency verification of multi-view normal priors to generate local reliability masks for filtering erroneous supervision. In some cases like scan40 of the DTU dataset, where supervisory camera viewpoints are predominantly frontal with limited side views, reconstructing lateral surfaces becomes challenging. Compounded by significant lighting and shadow variations across cameras, surface reconstruction accuracy is further compromised. As shown in Fig. 13, applying our multi-view normal consistency filtering (MVN) significantly improves reconstruction quality. Quantitative comparisons in Table I (rows

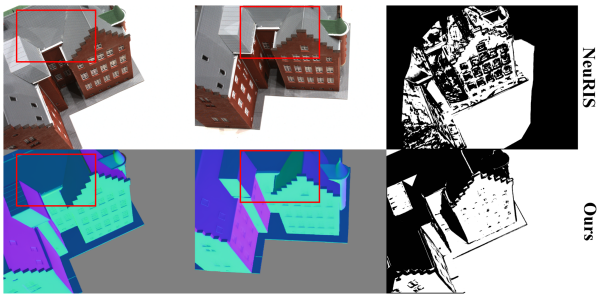


Fig. 14. Comparisons with NeuRIS on multi-view consistency assessment strategy. NeuRIS is susceptible to interference from inconsistent lighting and complex textures, leading to inaccurate reliability assessment of normal priors. Our method effectively overcomes these limitations and demonstrates superior performance in filtering out unreliable normal priors.

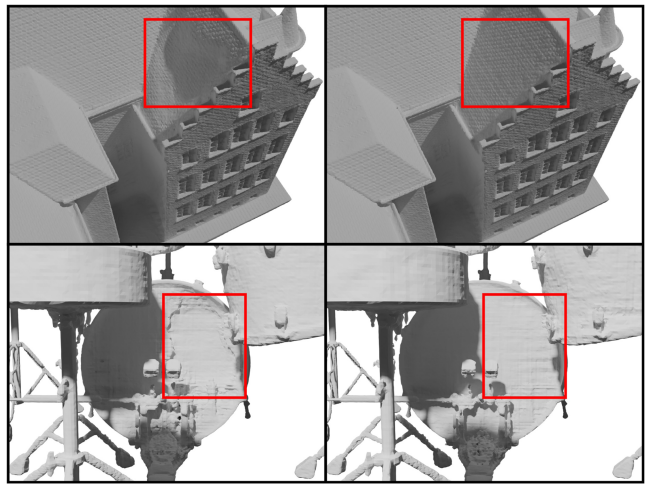
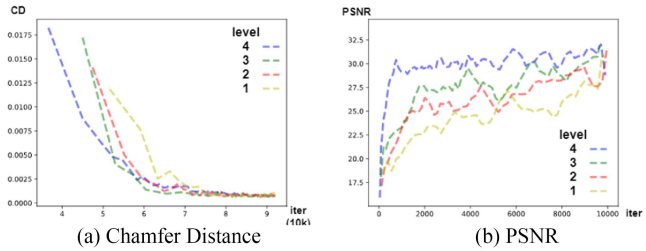


Fig. 15. The reconstruction quality is improved after applying perturbation and refinement to regions of unreliable normal priors.



(a) Chamfer Distance (b) PSNR

Fig. 16. Ablation of levels from 1 to 4 on the Lego dataset. The level of 4 accelerates convergence in the early training stage and has an advantage in the final training performance.

"MT+Pt+Pr" vs. "MT+Pt+Pr+MVN(NAL)") validate this enhancement, demonstrating our approach's robustness against viewpoint sparsity and illumination inconsistencies.

We compare the effectiveness of our method and NeuRIS in consistency verification of multi-view normal priors. Fig. 14 displays the local normal reliability masks (rightmost columns) generated by both approaches, where black regions indicate unreliable normals. In the DTU dataset, inconsistent lighting across views—such as shadow variations highlighted by red boxes in the first row—causes chaotic masking in NeuRIS. Our method's normal consistency verification, however, mitigates lighting variation interference, yielding more rational reliability assessments. Quantitative comparisons in Table I further validate this, with NeuRIS and our method labeled as "MVN(NCC)" and "MVN(NAL)", respectively. For fairness, we implement the "NCC" baseline using NeuRIS's recommended robust confidence threshold (0.66) while keeping other settings identical. Results confirm that our method more accurately evaluates normal prior reliability, thereby improving reconstruction quality.

We compare NAL using soft weights vs. hard thresholds, with results for soft weights shown in Table III ('weighted-unr'). Under identical settings, its reconstruction quality is notably worse than 'Ours'. Our method classifies NAL scores via a threshold to identify unreliable normals, guiding retraining on these pixels. This threshold-based approach efficiently targets only a subset of low-reliability regions. In contrast, incorporating NAL as weights forces sampling from all pixels (with varying weights), reducing sampling probability in highly unreliable areas. It also disturbs originally reliable regions (despite smaller weights), hurting training efficiency and reconstruction quality within the same epochs.

Perturbing and refining strategy on regions of unreliable normal priors. For regions with unreliable normal priors, we propose a perturbing and refining strategy that generates large gradients through controlled perturbations to focus optimization on potentially misreconstructed areas. As shown in Fig. 15, surfaces in these regions (highlighted by red boxes in the left two panels) suffer from poor reconstruction quality due to their reliance solely on 2D image loss (normal supervision being unavailable). After applying our perturbation and re-optimization, these surfaces exhibit significant improvement (right two panels, red boxes), which is quantitatively validated in Table I by comparing "MT+Pt+Pr+MVN(NAL)" with "Ours (full)".

Levels of tri-planes. We use 4 levels of multi-resolution tri-plane in all experiments. We also test the levels of 1, 2, and 3 and compare the CD and PSNR in Fig. 16. The colors blue, green, red, and yellow represent the levels of 4 to 1 respectively. The variations of values indicate that the level of 4 accelerates convergence in the early training stage and has an advantage in the final training performance.

V. Conclusion

A. Technical Summary

In this paper, we present a neural-based reconstruction method for generating detailed meshes from multi-view images. We introduce a multi-resolution tri-plane feature encoding, where high-resolution features capture fine geometric details, while low-resolution features suppress high-frequency artifacts caused by insufficient optimization and tri-plane discretization. A progressive training strategy, compatible with image supervision, is developed to hierarchically integrate scene details from coarse to fine granularity, ensuring training stability and convergence.

To address challenges from sparse viewpoints and inconsistent lighting in image datasets—common causes of surface

reconstruction failures—we leverage monocular normal priors as supervision. However, their effectiveness depends critically on prediction accuracy. We propose a multi-view normal prior consistency verification method to assess reliability, preventing erroneous normals from degrading reconstruction.

For regions with unreliable normals, we design a perturbing and refining strategy: problematic areas are perturbed into learnable convex states, followed by re-optimization to globally enhance reconstruction quality. This approach ensures geometric correctness even for initially under-reconstructed concave surfaces, while preserving accuracy in reliable regions through adaptive sampling.

B. Limitations and Future Work

While our method achieves fine geometric detail reconstruction, it has two main limitations.

First, reconstructing a single model requires approximately 16–20 hours, as the pursuit of high-quality outputs sacrifices computational efficiency. While integrating accelerated NeRF techniques [56] could mitigate this, balancing speed and fidelity remains an open challenge.

Second, even with relatively reliable normal supervision, our method struggles with specular surfaces (e.g., mirrors) and transparent objects (e.g., glass). Future work will focus on reconstructing surfaces with reflective, refractive, or translucent properties by incorporating advanced material-aware constraints.

REFERENCES

[1] B. Mildenhall, P.P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421.

[2] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, 1987.

[3] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NEUS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 27171–27183.

[4] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH Comput. Graph.*, vol. 18, no. 3, pp. 165–174, 1984.

[5] Y. Wang, I. Skorokhodov, and P. Wonka, "HF-NeuS: Improved surface reconstruction using high-frequency details," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1966–1978.

[6] E. R. Chan et al., "Efficient geometry-aware 3D generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16123–16133.

[7] Y. Wang, I. Skorokhodov, and P. Wonka, "Pet-neus: Positional encoding Tri-planes for neural surfaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12598–12607.

[8] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu, "NeuS2: Fast learning of neural implicit surfaces for multi-view reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3295–3306.

[9] Z. Li et al., "Neuralangelo: High-fidelity neural surface reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8456–8465.

[10] Q. Fu, Q. Xu, Y.-S. Ong, and W. Tao, "Geo-Neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction," 2022, *arXiv:2205.15848*.

[11] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, "Neural RGB-D surface reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6290–6301.

[12] E. Insafutdinov, D. Campbell, J. F. Henriques, and A. Vedaldi, "SNeS: Learning probably symmetric neural surfaces from incomplete data," 2022, *arXiv:2206.06340*.

[13] J. Wang et al., "NeuRIS: Neural reconstruction of indoor scenes using normal priors," 2022, *arXiv:2206.13597*.

[14] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction," 2022, *arXiv:2206.00665*.

[15] A. Tewari et al., "Advances in neural rendering," in *Comput. Graph. Forum*, vol. 41. New York, NY, USA: Wiley Online Library, 2022, pp. 703–735.

[16] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-NeRF: Neural radiance fields for dynamic scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10318–10327.

[17] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhofer, C. Lassner, and C. Theobalt, "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12959–12970.

[18] K. Park et al., "HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–12, 2021.

[19] J. Fang et al., "Fast dynamic radiance fields with time-aware neural voxels," in *Proc. SIGGRAPH Asia 2022 Conf. Papers*, 2022, pp. 1–9.

[20] R. Shao, Z. Zheng, H. Tu, B. Liu, H. Zhang, and Y. Liu, "Tensor4D: Efficient neural 4D decomposition for high-fidelity dynamic reconstruction and rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16632–16642.

[21] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt, "Neural actor: Neural free-view synthesis of human actors with pose control," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–16, 2021.

[22] S. Peng et al., "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9054–9063.

[23] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan, "NeuMan: Neural human radiance field from a single video," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 402–418.

[24] Z. Zheng, X. Zhao, H. Zhang, B. Liu, and Y. Liu, "AvatarReX: Real-time expressive full-body avatars," 2023, *arXiv:2305.04789*.

[25] K. Rematas et al., "Urban radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12932–12942.

[26] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-NERF: Scalable construction of large-scale NeRFs for virtual fly-throughs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12922–12931.

[27] M. Tancik et al., "Block-NeRF: Scalable large scene neural view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8248–8258.

[28] Y. Xiangli et al., "BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 106–122.

[29] Y.-J. Yuan, Y.-T. Sun, Y.-K. Lai, Y. Ma, R. Jia, and L. Gao, "NeRF-editing: Geometry editing of neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18353–18364.

[30] Y.-H. Huang, Y. He, Y.-J. Yuan, Y.-K. Lai, and L. Gao, "StylizedNeRF: Consistent 3D scene stylization as stylized NeRF via 2D-3D mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18342–18352.

[31] K. Jiang, S.-Y. Chen, F.-L. Liu, H. Fu, and L. Gao, "NeRFFaceEditing: Disentangled face editing in neural radiance fields," in *Proc. SIGGRAPH Asia 2022 Conf. Papers*, 2022, pp. 1–9.

[32] T. Wu, J.-M. Sun, Y.-K. Lai, and L. Gao, "DE-NeRF: Decoupled neural radiance fields for view-consistent appearance editing and high-frequency environmental relighting," in *Proc. ACM SIGGRAPH 2023 Conf. Proc.*, 2023, pp. 1–11.

[33] G. Lin et al., "SketchFaceNeRF: Sketch-based facial generation and editing in neural radiance fields," *ACM Trans. Graph.*, vol. 42, pp. 1–17, 2023.

[34] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, "FastNeRF: High-fidelity neural rendering at 200FPS," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14346–14355.

[35] P. Hedman, P.P. Srinivasan, B. Mildenhall, J. T. Barron, and P. Debevec, "Baking neural radiance fields for real-time view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5875–5884.

[36] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "Plenoctrees for real-time rendering of neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5752–5761.

[37] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530127
[38] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," 2022, *arXiv:2209.14988*.
[39] C.-H. Lin et al., "Magic3D: High-resolution text-to-3D content creation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 300–309.
[40] J. Xu et al., "Dream3D: Zero-shot text-to-3D synthesis using 3D shape prior and text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20908–20918.
[41] Z. Wang et al., "ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation," 2023, *arXiv:2305.16213*.
[42] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5855–5864.
[43] J. T. Barron, B. Mildenhall, D. Verbin, P.P. Srinivasan, and P. Hedman, "Mip-NeRF 360: Unbounded anti-aliased neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5470–5479.
[44] J. T. Barron, B. Mildenhall, D. Verbin, P.P. Srinivasan, and P. Hedman, "Zip-NERF: Anti-aliased grid-based neural radiance fields," 2023, *arXiv:2304.06706*.
[45] W. Hu et al., "Tri-MIPRF: Tri-mip representation for efficient anti-aliasing neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19774–19783.
[46] F. Darmon, B. Bascle, J.-C. Devaux, P. Monasse, and M. Aubry, "Improving neural implicit surfaces geometry with patch warping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6260–6269.
[47] A. Dogaru, A.-T. Ardelean, S. Ignatyev, E. Zakharov, and E. Burnaev, "Sphere-guided training of neural implicit surfaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20844–20853.
[48] H. Guo et al., "Neural 3D scene reconstruction with the manhattan-world assumption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5511–5520.
[49] Y. Zhuang et al., "Anti-aliased neural implicit surfaces with encoding level of detail," 2023, *arXiv:2309.10336*.
[50] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "TensoRF: Tensorial radiance fields," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 333–350.
[51] Q. Gao, Q. Xu, H. Su, U. Neumann, and Z. Xu, "Strivec: Sparse tri-vector radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17569–17579.
[52] C. Ye et al., "StableNormal: Reducing diffusion variance for stable and sharp normal," *ACM Trans. Graph.*, vol. 43, no. 6, pp. 1–18, 2024.
[53] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, pp. 153–168, 2016.
[54] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, "2D Gaussian splatting for geometrically accurate radiance fields," in *Proc. ACM SIGGRAPH 2024 Conf. Papers*, 2024, pp. 1–11.
[55] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–141, 2023.
[56] R. Li, M. Tancik, and A. Kanazawa, "NerfAcc: A general NeRF accleration toolbox," 2022, *arXiv:2210.04847*.

Xue-Kun Xiang is a student with the Institute of Computing Technology, Chinese Academy of Sciences, supervised by Prof. Lin Gao. At present, his main research work focuses on rendering and geometric reconstruction of 3D scenes, mainly using NeRF and 3D-GS methods based on deep learning.

Yu-Jie Yuan received the bachelor's degree in mathematics from Xi'an Jiaotong University, in 2018. He is currently the PhD degree in the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics and neural rendering.

Wen-Bo Hu received the BSc degree in computer science and technology from the Dalian University of Technology, China, in 2018, and the PhD degree in computer science from the Chinese University of Hong Kong. His research interests lie in the interface of Computer Graphics, Computer Vision, and AI, with a focus on 3D learning, including Reconstruction, Rendering, Understanding, and Generation. He is especially excited to explore how new-generation AI technologies benefit fundamental problems in computer vision and graphics.

Yu-Tao Liu is a student with the Institute of Computing Technology, Chinese Academy of Sciences, supervised by Prof. Lin Gao. His current research interests include computer graphics, computer vision, and deep learning.

Yue-Wen Ma received the PhD degree from Nanyang Technological University, Singapore, in 2013. He has been engaged in the research and product of computer graphics and 3D vision for a long time. He is currently the leader of 3D reconstruction with ByteDance Pico.

Lin Gao (Member, IEEE) received the PhD degree from Tsinghua University. He is currently a professor with the Institute of Computing Technology, Chinese Academy of Sciences and the University of Chinese Academy of Sciences. He has been awarded the Newton Advanced Fellowship from Royal Society and Asia Graphics Association Young Researcher Award. His research interests include computer graphics and geometric processing.