DOI: 10.1111/cgf.70325

EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

GeoFusionLRM: Geometry-Aware Self-Correction for Consistent 3D Reconstruction

Ahmet Burak Yildirim[1,*] Tuna Saygin[1,*] Duygu Ceylan[2] Aysegul Dundar[1]

[1]Bilkent University, Ankara, Turkey
[2]Adobe Research, London, United Kingdom

*These authors contributed equally.

Abstract

Single-image 3D reconstruction with large reconstruction models (LRMs) has advanced rapidly, yet reconstructions often exhibit geometric inconsistencies and misaligned details that limit fidelity. We introduce GeoFusionLRM, a geometry-aware self-correction framework that leverages the model's own normal and depth predictions to refine structural accuracy. Unlike prior approaches that rely solely on features extracted from the input image, GeoFusionLRM feeds back geometric cues through a dedicated transformer and fusion module, enabling the model to correct errors and enforce consistency with the conditioning image. This design improves the alignment between the reconstructed mesh and the input views without additional supervision or external signals. Extensive experiments demonstrate that GeoFusionLRM achieves sharper geometry, more consistent normals, and higher fidelity than state-of-the-art LRM baselines.

CCS Concepts

• ***Computing methodologies*** → ***Machine learning approaches; Image-based rendering;*** *Reconstruction; Probabilistic reasoning;*

1. Introduction

Recovering 3D geometry from images is fundamental to many applications in vision and graphics, such as content creation, AR/VR, and robotics. Reconstructing a full 3D model from a single image is particularly challenging due to the severe ambiguity of missing viewpoints. Recent Large Reconstruction Models (LRMs) [HZG*23, WZB*24, LTZ*23, WTB*23, JHP24, XBS*24], have made progress on this task by training transformer architectures on large collections of image–3D data pairs, enabling them to directly predict 3D assets from a single view. While these models succeed in generating 3D assets that capture the coarse 3D shape observed in the images, they often struggle to produce meshes that are consistent with the conditioning image in terms of geometric details. They suffer from limitations including inaccurate geometry, distorted surface normals, and misaligned details. These limitations highlight the need for approaches that improve consistency between the generated 3D assets and the input images.

Existing approaches [HZG*23, XCG*24, HBV*25, TCC*24] typically encode 2D image features and inject them into the reconstruction pipeline via attention-level conditioning. While these models provide a strong semantic prior, accurate 3D reconstruction from a single image remains fundamentally ill-posed due to missing volumetric information. In particular, single-image methods such as LRM [HZG*23] and SPAR3D [HBV*25] often struggle with occluded and back-facing regions, where geometry must be inferred from learned shape priors rather than direct visual evidence. On the other hand, recent LRM-based methods such as LGM [TCC*24] and InstantMesh [XCG*24] attempt to alleviate this limitation by leveraging pretrained image-to-multiview models [SWY*23, WS23, LWVH*23] to synthesize multiple views from a single input image, which are then jointly encoded to provide a stronger geometric prior. However, the synthesized multi-view images are not always geometrically consistent. Moreover, when operating directly on real multi-view images without view synthesis, the available views are sometimes insufficiently informative when the object geometry is difficult to infer from RGB observations alone. As a result, ambiguities in 3D understanding can persist, which occasionally leads to reconstruction failures where geometric errors in the mesh are visually masked in the rendered RGB outputs. (e.g., holes suggested by appearance but absent in the underlying geometry). To address this, our model predicts its own depth and normal maps from an initial reconstruction. These predicted geometric cues are then fused with image features in a second pass, allowing the network to refine depth, normals, and semantic information. This self-contained two-stage process improves geometric fidelity and enables more accurate and consistent mesh reconstruction without relying on external predictors. The improvements achieved by our approach on FLUX [Lab24]-generated synthesized images are illustrated in Fig. 1, demonstrating sharper

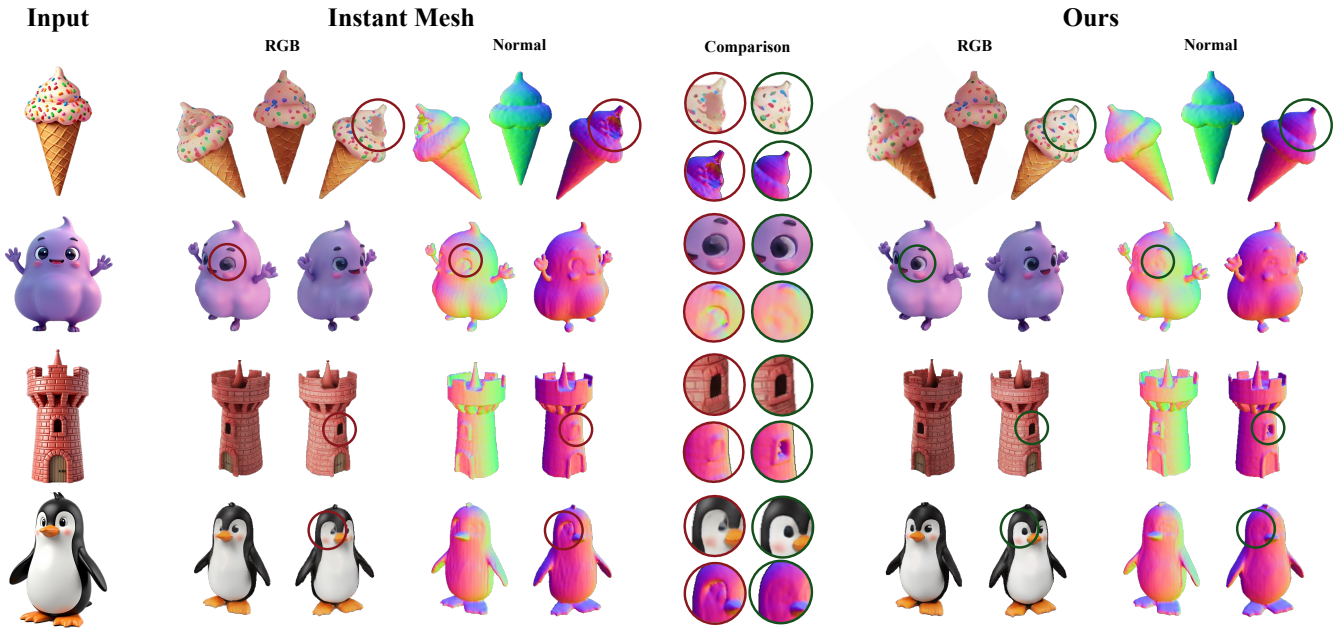

Figure 1: *Qualitative comparison using a synthesized input image generated by the FLUX image generator. The same synthesized image is provided as input to the InstantMesh baseline and our proposed GeoFusionLRM. The baseline struggles to preserve geometric fidelity, producing distorted normals and misaligned surface details. In contrast, our iterative geometric conditioning progressively corrects these errors, yielding reconstructions with sharper normals and RGB renderings that more closely match the GT view.*

geometry, more accurate normals, and better alignment with the input image compared to its baseline.

Building on this idea, we propose GeoFusionLRM, a geometry-aware self-corrective conditioning framework that integrates these predicted geometric cues for two-stage refinement. Once an initial mesh is first reconstructed from the input image, we introduce a geometry encoder that encodes features of the depth and normal maps of this initial reconstruction. The resulting features are fused with semantic features from the vision encoder through a fusion module. This two-pass process corrects residual geometric errors and produces more accurate and consistent meshes. Our method builds upon InstantMesh [XCG*24] as a baseline and introduces two key components: (i) a GeoFormer encoder, fine-tuned with geometric supervision to capture structural consistency from normals and depths, and (ii) the GeoFuser module, a lightweight token-wise network that fuses semantic features from the vision encoder with geometry-aware embeddings from GeoFormer.

Our contributions can be summarized as follows:

- **Self-predicted geometry-aware conditioning:** We introduce GeoFusionLRM, which refines meshes in a two-stage process by conditioning on depth and normal cues predicted from intermediate reconstructions.
- **GeoFormer encoder:** We propose a geometry-aware encoder, initialized from DINO [CTM*21] and fine-tuned with geometric supervision, to capture structural alignment with conditioning images.
- **GeoFuser module:** We design a lightweight token-level fusion network that merges semantic and geometric features to produce refined triplane conditioning.
- **Improved consistency and fidelity:** Extensive experiments show that GeoFusionLRM improves over InstantMesh [XCG*24] and other competing models, yielding sharper geometry, more accurate normals, and higher fidelity to input views.

2. Related Work

Single-image 3D reconstruction has advanced considerably with the advent of transformer-based architectures and large-scale training datasets. Earlier approaches predominantly employed category-specific encoder–decoder networks [BDL*21, CLG*19, GKM20, DGTC23a, DGTC23b], which restricted their ability to generalize beyond the categories seen during training. More recently, large reconstruction models (LRMs) have been introduced, trained on diverse collections of 3D assets [HZG*23], and shown to generate high-fidelity 3D geometry from sparse inputs such as a single image. These models shift the paradigm from category-specific learning toward general-purpose reconstruction, providing stronger robustness to variations in object shape and appearance. Architectures such as Instant3D [LTZ*23] and InstantMesh [XCG*24] exemplify this trend by integrating multi-view diffusion [SCZ*23] with transformer-based 3D decoders, producing triplane or volumetric representations that generalize across categories without requiring further fine-tuning. Recent extensions further improve efficiency and applicability, for example, through architectural simplifications and curated training data [TPL*24], or by jointly estimating camera

pose and geometry to enable reconstruction from unposed sparse inputs [WTB*23]. Collectively, these works highlight a transition from narrow, limited domain reconstructions toward more scalable and generalizable frameworks that form the basis for current research directions. However, despite their efficiency, existing LRMs remain sensitive to errors introduced during multiview synthesis and lack explicit mechanisms for geometry-aware correction, causing geometric inconsistencies and misalignment with the conditioning image to persist once reconstruction is completed.

Several feed-forward approaches [LXJ*23, ZPG*24, LLL*24, VYB*24] synthesize multiview observations using diffusion-based view generation conditioned on a single image [LWVH*23], followed by 3D reconstruction through task-specific geometric pipelines rather than a unified LRM-style representation. As an extension, methods such as Wonder3D and GeoWizard [LGL*24, FYH*24] jointly generate RGB images and surface normals, employing cross-domain attention mechanisms to ensure geometric consistency during view synthesis. Although these methods improve multiview consistency, reconstruction quality remains constrained by inaccuracies in the synthesized views. In contrast, optimization-based approaches leverage strong 2D diffusion priors through per-scene procedures such as Score Distillation Sampling (SDS) [TWZ*23], achieving high geometric fidelity at the cost of substantial computational overhead.

There have also been iterative approaches proposed for improving 3D reconstruction quality. Some methods refine reconstructions by progressively incorporating additional input views to update the underlying representation [LWC*25, KNS*25], which inherently assumes access to multi-view image sets. In contrast, our approach performs reconstruction from a single input image and applies refinement internally through predicted geometric cues rather than external view accumulation. Closely related, GTR [ZHW*24] introduces a lightweight per-instance refinement stage on top of Large Reconstruction Models, where NeRF color parameters and triplane features are fine-tuned at test time using differentiable mesh rendering. Its test-time optimization primarily targets appearance by refining RGB textures through color-based losses, without explicitly correcting the underlying geometry, in contrast to our geometry-aware refinement.

In another line of recent work, surface normals have been particularly used to improve geometric fidelity of reconstruction models [PLS25,SZW*25]. These methods rely on an additional monocular depth or surface normal estimation network and provide normal maps as input to a reconstruction model directly. Our work differs in that we propose a two-stage self-corrective conditioning framework, which does not require additional networks or lengthy optimization steps.

3. Method

Our goal is to improve the geometric consistency of single-image 3D reconstruction with respect to the conditioning image. To this end, we introduce **GeoFusionLRM**, an iterative conditioning framework that augments Large Reconstruction Models (LRMs) with geometry-aware self-supervision. The overall architecture is illustrated in Fig. 2. We first review the baseline LRM pipeline before describing our geometry-aware extensions.

3.1. Preliminaries

Large Reconstruction Models (LRMs) such as InstantMesh [XCG*24] predict a 3D representation directly from a single conditioning image by employing a transformer-based architecture with 3D-aware cross-attention. In practice, InstantMesh first employs a multi-view diffusion model [SCZ*23] to synthesize six views of the object $\{I_k\}_{k=1}^{6}$ from the input image I, each associated with known camera parameters $\{C_k\}_{k=1}^{6}$. These views are then encoded by a transformer backbone (ViT Encoder) with DINO [CTM*21] initialization, where the camera parameters are injected into the AdaLN layers. The encoded semantic tokens are defined as

$$F_k^{\text{sem}} = E_{\text{sem}}(I_k, C_k), \tag{1}$$

where $F_k^{\text{sem}} \in \mathbb{R}^{N\times d}$ denotes the semantic token embeddings for view k. The triplane decoder transformer aggregates these tokens across views and generates the triplane tokens by applying cross-attention between the semantic tokens and triplane tokens. The resulting features form a triplane representation $\mathcal{T} \in \mathbb{R}^{3\times R\times R\times d}$ are subsequently decoded into a mesh via differentiable iso-surface extraction. Although this design provides stronger 3D awareness than earlier LRMs, the reliance on semantic embeddings still limits geometric fidelity. While camera parameters can inject 3D information into the 2D tokens, they often fail to capture structural details, leading to unwanted artifacts in the reconstructed geometry. This embedding bottleneck motivates our geometric refinement strategy, where the model receives feedback from geometry rendered under the input view to progressively refine the embeddings extracted from RGB images.

3.2. GeoFusionLRM Overview

GeoFusionLRM introduces geometry-aware conditioning into the LRM pipeline (see Fig. 2). Starting from an initial mesh $\mathcal{M}^{(0)}$ generated by the baseline LRM, we extract depth and normal maps

$$D^{(t)},\, N^{(t)} = \Pi(\mathcal{M}^{(t)}), \tag{2}$$

where $\Pi(\cdot)$ denotes differentiable rendering of the current mesh $\mathcal{M}^{(t)}$ into depth D and normals N as defined in Eq. 2. These maps are encoded by the geometry-aware **GeoFormer**, and fused with semantic tokens through the **GeoFuser** module to guide the next reconstruction pass.

3.3. GeoFormer: Geometry-Aware Encoder

GeoFormer is designed to capture structural consistency from normal and depth projections. It is initialized as a copy of the ViT encoder used in InstantMesh, including AdaLN-based camera parameter conditioning. To process normal and depth information, we extend the input layer from three channels (RGB) to four channels with zero-initialized weights, which enables the encoder to handle geometry maps instead of the color input. Formally, the geometry-aware tokens are defined as

$$F^{\text{geo}} = E_{\text{geo}}(D^{(t)}, N^{(t)}), \tag{3}$$

where E_{geo} denotes the GeoFormer encoder and $F^{\text{geo}} \in \mathbb{R}^{N\times d}$ are geometry-aware embeddings. GeoFormer is added as a parallel branch to the pipeline, which produces geometry-aware tokens

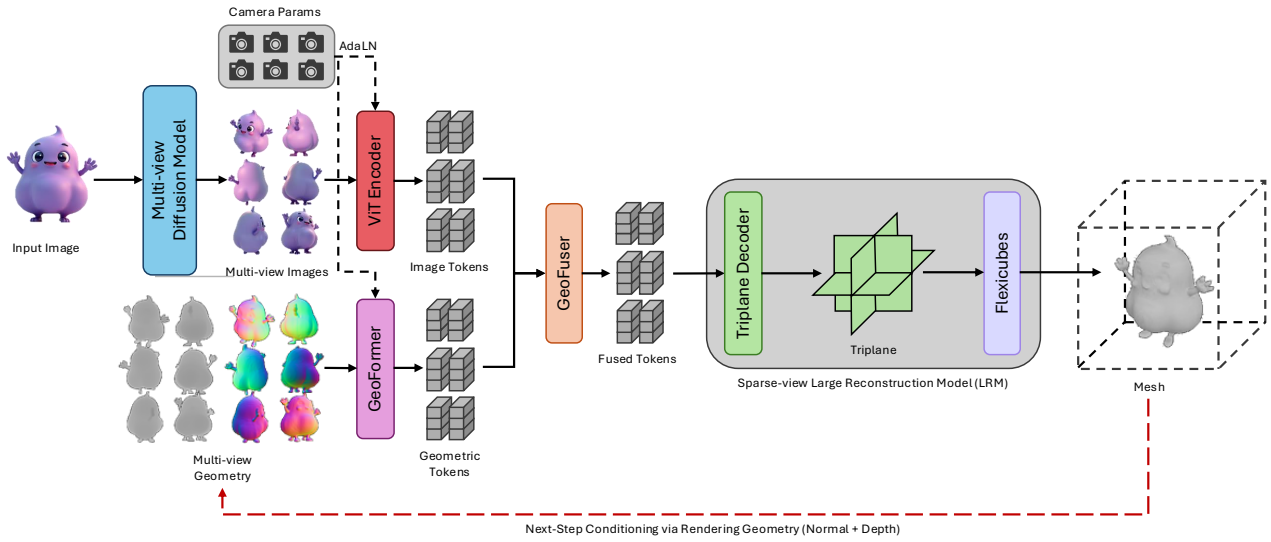

Figure 2: *Overview of the proposed GeoFusionLRM architecture. Given a conditioning image, semantic features are extracted with a pre-trained vision encoder, while geometric cues from normals and depths of the intermediate mesh are encoded by the geometry-aware **GeoFormer**. The **GeoFuser** module merges these two streams of embeddings at the token level to produce refined conditioning features, which guide the LRM in generating an updated 3D mesh. This process corrects residual geometric errors and improves the consistency of surface normals and RGB renderings with respect to the conditioning image.*

that are later fused with semantic tokens from the vision encoder through the GeoFuser module.

3.4. GeoFuser: Token-Level Feature Fusion

To integrate semantic and geometric features, we propose the **GeoFuser** module. Given semantic tokens F^{sem} from the input image and geometry-aware tokens F^{geo} extracted from rendered normal and depth maps of the same view (Eq. 2 and Eq. 3), GeoFuser produces corrective residuals that refine the semantic embeddings. Formally, the fused tokens are defined as

$$F^{\text{fused}} = F^{\text{sem}} + f_\theta(F^{\text{sem}}, F^{\text{geo}}), \tag{4}$$

where $f_\theta(\cdot)$ is a lightweight two-layer feed-forward network with a hidden SiLU activation. Its final linear layer is initialized with zero weights and bias, ensuring that the residual correction is disabled at the first iteration and enabled in the subsequent refinement. At the first iteration, the residual term is disabled, and the baseline LRM generates an initial mesh. In the following iteration, geometry maps rendered from the reconstructed mesh are encoded by GeoFormer and fused through Eq. 4 to refine the semantic embeddings based on inconsistencies between input image features and rendered geometry. The refined tokens F^{fused} are then injected into the cross-attention layers of the LRM, subsequently improving geometric fidelity.

During training, we unroll the refinement process for T = 3 steps. At each step (t = 0, 1, 2), the model renders the current reconstruction, encodes the resulting depth and normal maps, and receives supervision on the outputs for that step. This setup makes the optimization stable and keeps the overall training cost manageable. At inference time, however, we observe that an initial reconstruction stage followed by a single refinement stage is sufficient.

3.5. Training Strategy

During training, the baseline InstantMesh backbone is frozen, and only the parameters of *GeoFormer* and *GeoFuser* are optimized.

For supervision, we utilize the default InstantMesh losses on the rendered outputs and backpropagate immediately. In particular, this objective combines photometric MSE, perceptual LPIPS [ZIE*18], mask consistency, depth alignment, normal similarity, and FlexiCubes [SMH*23] regularization, with the same weights as InstantMesh:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{rgb}} + 2.0 \cdot \mathcal{L}_{\text{lpips}} + \mathcal{L}_{\text{mask}} \\ + 0.5 \cdot \mathcal{L}_{\text{depth}} + 0.2 \cdot \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{reg}}.\end{aligned} \tag{5}$$

4. Experiments

In this section, we present quantitative and qualitative results for GeoFusionLRM, detailing our training setup, evaluation datasets, baselines, and metrics.

Datasets. We train our model on Objaverse-1.0 [DSS*23], excluding assets with rendered alpha coverage of $\leq 10\%$ to remove highly transparent or very small objects. After filtering, the training set contains approximately 168k objects. For evaluation, we use the OmniObject3D [WZF*23] and Google Scanned Objects (GSO) [DFK*22] datasets. From OmniObject3D, we select five objects per category across 100 categories (500 in total), while we uniformly sample 500 objects from GSO. Each object is rendered on a fixed viewing grid, defined by elevations

$\{-20°, -10°, 0°, 10°, 20°\}$ crossed with six uniformly spaced azimuths in an orbital setting. In addition to these predefined views, we evaluate our method on the standard OmniObject3D benchmark test views, following the evaluation protocol used in prior work (e.g., InstantMesh). Specifically, we report results on 16 benchmark views per object provided by the OmniObject3D dataset.

Implementation Details. We fine-tune our model for 168k steps on 4 × A100 GPUs using the AdamW optimizer with an initial learning rate of 4×10^{-6}, $\beta_1 = 0.90$, $\beta_2 = 0.95$, and a weight decay of 0.01. We employ a cosine scheduler that anneals the learning rate to 0 over 100k steps. Training supervision is provided with 32 views per object, obtained by randomly sampling camera poses on a viewing sphere with radius $r \sim \mathcal{U}(1.5, 2.2)$ and uniformly distributed orientations.

Baselines. We compare GeoFusionLRM against four recent approaches: LRM [HZG*23], SPAR3D [HBV*25], LGM [TCC*24], and InstantMesh [XCG*24]. OpenLRM [HW23] weights are utilized for LRM comparison, which is an open-source implementation of the Large Reconstruction Model (LRM) [HZG*23]. It is a transformer-based framework trained to predict NeRF representations of the object from a single input image. SPAR3D follows a two-stage design: it first predicts a sparse 3D point cloud with a lightweight point-diffusion model, and then refines it into a detailed mesh conditioned on the input view. LGM reconstructs 3D objects as collections of Gaussian parameters for an efficient representation that can predict high-quality novel view synthesis. InstantMesh integrates an off-the-shelf multi-view diffusion model with an LRM-style sparse-view reconstructor, enabling fast feed-forward mesh generation from a single image.

Metrics. We evaluate both the visual quality and geometric accuracy of the generated 3D assets. For visual quality, we compare predicted and ground-truth pixel values in RGB space, reporting PSNR, SSIM, and LPIPS (higher is better for PSNR/SSIM, while lower is better for LPIPS). For geometry, we assess structural consistency by rendering normal maps from the generated meshes in Blender using the same camera grid, and then comparing predicted and ground-truth normals with the same set of metrics. All results are averaged over both views and objects.

Table 1: *Quantitative results on the **GSO dataset using uniform views** for RGB images and normal maps.*

Method	RGB			Normal		
	PSNR ↑	SSIM ↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS ↓
OpenLRM	17.2217	0.8885	0.1212	18.7437	0.8837	0.1243
Spar3d	16.1483	0.8790	0.1214	19.0522	0.8999	0.1160
LGM	19.0940	0.9016	0.0975	22.0817	0.9205	0.0886
InstantMesh	20.3072	0.9201	0.0832	25.8303	0.9468	0.0625
Ours	**20.3537**	**0.9212**	**0.0831**	**26.3866**	**0.9497**	**0.0592**

Table 2: *Quantitative results on **OmniObject3D using uniform views** for RGB images and normal maps.*

Method	RGB			Normal		
	PSNR ↑	SSIM ↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS ↓
OpenLRM	16.5965	0.8753	0.1189	19.2952	0.8696	0.1191
Spar3d	17.0840	0.8849	0.1048	19.6523	0.8942	0.1092
LGM	20.6049	0.9061	0.0819	23.3091	0.9054	0.0831
InstantMesh	21.9417	0.9150	0.0798	24.6765	0.9177	0.0781
Ours	**23.0465**	**0.9205**	**0.0722**	**26.1600**	**0.9265**	**0.0658**

Table 3: *Quantitative results on **OmniObject3D using benchmark views** for RGB images and normal maps.*

Method	RGB			Normal		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SPAR3D	16.15	0.881	0.1235	17.52	0.883	0.1248
LRM	16.85	0.874	0.1261	18.04	0.865	0.1245
LGM	20.98	0.905	0.0849	22.61	0.906	0.0826
InstantMesh	21.85	0.913	0.0805	24.24	0.918	0.0769
Ours	**22.75**	**0.916**	**0.0741**	**25.76**	**0.926**	**0.0648**

4.1. Quantitative Comparisons

Tables 1, 2, and 3 summarize quantitative results on the GSO and OmniObject3D datasets. Across both datasets, GeoFusionLRM consistently ranks first in all metrics, especially in the normal map scores.

In GSO, improvements are more pronounced in the normal metrics than in RGB, which reflects the geometry-focused conditioning of our pipeline. In OMNIOBJECT3D, the same pattern is observed for both uniform and benchmark views: RGB results show a modest improvement, while normal maps improve more clearly across all metrics, indicating better geometric accuracy, especially for objects with thin structures and curved surfaces.

The advantage over the other baselines is especially visible in geometry. Furthermore, we improve upon the strongest baseline, InstantMesh, both qualitatively and quantitatively.

4.2. Qualitative Comparisons

Figure 3 provides a qualitative comparison against state-of-the-art methods on the GSO dataset. These results reveal a critical weakness in several prior models: the tendency to prioritize plausible-looking textures at the expense of geometric fidelity. This is often achieved through aggressive surface smoothing and by baking complex shading cues directly into the albedo. While this can mask underlying shape defects in the RGB output, it leads to significantly distorted and uninformative surface normals.

This failure mode is particularly evident in the reconstructions from LGM and InstantMesh. For instance, with the turtle teapot, both methods flatten the complex pattern on the shell into a smooth, featureless surface. Similarly, for the bowl, the sharp rim is incorrectly rounded off. This geometric simplification is noticeably visible in their corresponding normal maps, which lack high-frequency detail. The most telling example is the decorative planter, where the crisp, cut-out patterns are degraded into shallow, indistinct indentations, demonstrating a failure to reconstruct complex topology.

In contrast, SPAR3D and our method, GeoFusionLRM, show a superior capability to preserve fine geometric details. Both methods successfully capture the sharp edges of the bowl's rim and the complex holes of the planter, resulting in crisp, well-defined normal maps that accurately reflect the object's true surface structure.

However, our method also improves on reconstructing dark, low-feature regions. This is best illustrated by the hat reconstruction in the final row. The dark band on the hat presents an ambiguous region for reconstruction. While most methods produce a simplified shape, SPAR3D hallucinates a large hole, failing to infer the continuous surface underneath the dark texture. GeoFusionLRM, on the

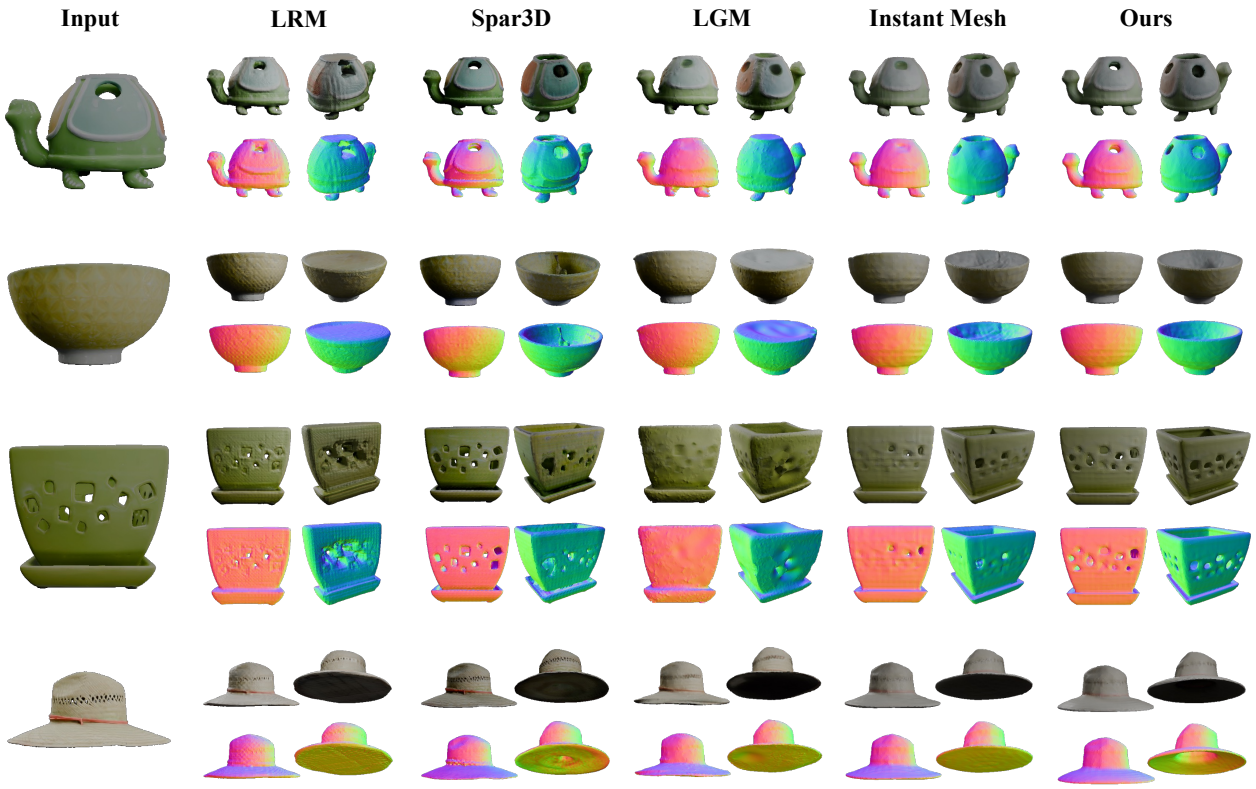

Figure 3: *Qualitative results on GSO. Columns show the conditioning input image (left), followed by LRM, SPAR3D, LGM, InstantMesh, and our GeoFusionLRM. For each method, we display results rendered from the same camera viewpoints, showing RGB outputs (top) and surface normals (bottom).*

other hand, correctly reconstructs the complete, coherent geometry of the hat, demonstrating a superior ability to reason about shape even in areas with ambiguous visual information. This highlights our model's improved capacity for producing not only detailed but also geometrically complete reconstructions.

4.3. Ablation Studies

We conduct a comprehensive ablation analysis to evaluate the contribution of each design choice in our method. Specifically, we study the effects of input conditioning, architectural components, and the initialization strategy of the geometry-aware encoder. All ablations are evaluated using SSIM and LPIPS.

Table 4 reports ablation results evaluated on both RGB and normal map reconstruction quality. Using only depth or only normal conditioning consistently degrades performance compared to using both conditions jointly, which confirms that depth and normal information provide complementary geometric signals. Replacing the proposed geometry-aware fusion with simple token concatenation or removing the GeoFormer initialization leads to a noticeable drop in reconstruction quality. While these variants still outperform the InstantMesh baseline, they remain below the proposed method across all metrics. The overall trends are consistent across RGB and normal evaluations, with the proposed method achieving the best performance in both RGB and normal reconstructions.

Table 4: *Ablation results evaluated on RGB and normal map reconstruction quality.*

	RGB		Normal	
Method	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS ↓
InstantMesh	0.915	0.0798	0.918	0.0781
Random Init	0.914	0.0755	0.924	0.0680
Token Concat	0.916	0.0739	0.926	0.0663
Normal Only	0.916	0.0738	0.926	0.0662
Depth Only	0.916	0.0738	0.926	0.0661
Ours (Proposed)	**0.920**	**0.0722**	**0.927**	**0.0658**

Finally, we analyze the effect of the initialization strategy. Initializing GeoFormer from random weights leads to consistent drops in SSIM and LPIPS in both RGB and normal evaluations. In contrast, initializing from a pretrained ViT encoder provides a stronger and more stable starting point for learning geometry-aware features. Since the DINO encoder used in InstantMesh is trained on a large

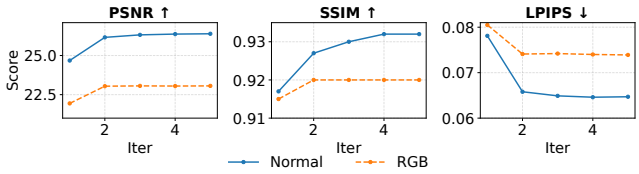

Figure 4: *Performance across refinement iterations on the OmniObject3D dataset under uniform views.*

dataset with geometry-aware adaptive normalization through camera parameters, reusing this encoder for the conditioning branch smooths the optimization landscape and allows the refinement stage to focus on correcting geometric inconsistencies rather than learning geometry-aware 2D-to-3D feature mappings from scratch.

4.4. Iteration Analysis

GeoFusionLRM's ability to condition on its own rendered depth and normal maps enables iterative refinement of the generated geometry. To assess the effectiveness of this strategy, we evaluate performance across varying numbers of refinement steps, where Iteration 1 corresponds to a single forward pass without geometric conditioning. As shown in Figure 4, both geometric and appearance metrics (PSNR, SSIM, LPIPS) improve substantially after the first refinement pass, while additional iterations yield diminishing returns and quickly plateau. Considering the computational cost of each pass (Sec. 5), we adopt a two iteration setting, one initial reconstruction followed by one geometry aware refinement, as the optimal trade-off between quality and efficiency.

4.5. Computational Cost Analysis

As refinement requires additional passes, we evaluate the inference-time computational cost of GeoFusionLRM in comparison to InstantMesh. Table 5 reports TFLOPs and single-object inference time measured on an NVIDIA RTX 3090, where the same object is reconstructed by both methods. While GeoFusionLRM improves reconstruction quality through geometry-aware refinement, this improvement comes with a higher computational cost. Specifically, the refinement process introduces extra forward passes over the base InstantMesh encoder, resulting in increased TFLOPs and longer inference time. This highlights the trade-off between reconstruction quality and computational efficiency for refinement-based methods.

5. Conclusion

We have presented GeoFusionLRM, which aims to improve geometric fidelity of large reconstruction models through a refinement strategy. In particular, we introduce a geometric feature encoder and fusion modules into a baseline LRM method (InstantMesh in our experiments). Given an initial 3D reconstruction, these modules extract geometric features from the depth and normal maps of the reconstruction that act as residual features and improve the reconstruction. Our experiments show that this approach improves upon the base LRM method, especially in terms of geometric fidelity. The gains in geometric fidelity are most evident on surfaces exhibiting coherent shape deviations, such as flattened bumps, softened edges, and merged or missing holes, where geometric errors persist consistently across a region.

Table 5: *Computational cost and inference time comparison between InstantMesh and GeoFusionLRM.*

Model	TFLOPs ↓	Inference Time ↓
InstantMesh	3.878	0.989
Ours	8.687	3.854

Limitations. As discussed in Section 5, the proposed refinement strategy increases inference-time computation due to additional forward passes. In terms of reconstruction quality, thin structures such as small branches and fine root segments remain a challenging case for LRM-based models. Geometry-conditioned self-refinement improves branching structures for the plant object shown in Figure 5 compared to InstantMesh. The refinement effectively fixes gaps and discontinuities observed in the InstantMesh reconstruction along coarse root branches, resulting in improved structural coherence and surface continuity. However, very fine root branches remain missing in both the baseline and our result. This limitation stems from the low-resolution triplane representation used by the InstantMesh backbone, which restricts the recovery of extremely fine-scale geometric details. Overall, while the refinement step consistently improves branch coherence and corrects large-scale structural errors, details below the effective resolution of the underlying representation remain challenging for both methods.

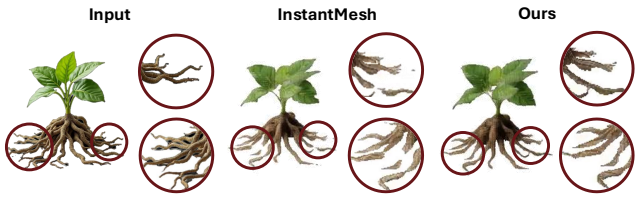

Figure 5: *Limitations on thin structure reconstruction. Our refinement improves coarse branches by closing gaps (see zooms), but very thin root segments remain missing due to the limited resolution of the InstantMesh triplane backbone.*

Future work. Future research could extend our approach beyond local geometric cues like depth and normals. One promising direction is to incorporate global geometric priors, including symmetry constraints or semantic structural consistency, directly into the fusion module to further regularize the reconstruction.

6. Acknowledgments

This work was supported by the BAGEP Award of the Science Academy. We acknowledge the EuroHPC Joint Undertaking for awarding the project ID EHPC-BEN-2025B05-045 access to the MareNostrum5 ACC at Barcelona, Spain. The author also acknowledges support from the Scientific and Technological Research Council of Turkey (TUBITAK) through the 2211 Graduate Scholarship Program.

References

[BDL*21] BHATTAD A., DUNDAR A., LIU G., TAO A., CATANZARO B.: View generalization for single image textured 3d models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 6081–6090. 2

[CLG*19] CHEN W., LING H., GAO J., SMITH E., LEHTINEN J., JACOBSON A., FIDLER S.: Learning to predict 3d objects with an interpolation-based differentiable renderer. In *Advances in Neural Information Processing Systems* (2019), pp. 9609–9619. 2

[CTM*21] CARON M., TOUVRON H., MISRA I., JÉGOU H., MAIRAL J., BOJANOWSKI P., JOULIN A.: Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2021). 2, 3

[DFK*22] DOWNS L., FRANCIS A., KOENIG N., KINMAN B., HICKMAN R., REYMANN K., MCHUGH T. B., VANHOUCKE V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)* (2022), IEEE, pp. 2553–2560. 4

[DGTC23a] DUNDAR A., GAO J., TAO A., CATANZARO B.: Fine detailed texture learning for 3d meshes with generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*, 12 (2023), 14563–14574. 2

[DGTC23b] DUNDAR A., GAO J., TAO A., CATANZARO B.: Progressive learning of 3d reconstruction network from 2d gan data. *IEEE Transactions on Pattern Analysis and Machine Intelligence 46*, 2 (2023), 793–804. 2

[DSS*23] DEITKE M., SCHWENK D., SALVADOR J., WEIHS L., MICHEL O., VANDERBILT E., SCHMIDT L., EHSANI K., KEMBHAVI A., FARHADI A.: Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023), pp. 13142–13153. 4

[FYH*24] FU X., YIN W., HU M., WANG K., MA Y., TAN P., SHEN S., LIN D., LONG X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision* (2024), Springer, pp. 241–258. 3

[GKM20] GOEL S., KANAZAWA A., MALIK J.: Shape and viewpoint without keypoints. *arXiv preprint arXiv:2007.10982* (2020). 2

[HBV*25] HUANG Z., BOSS M., VASISHTA A., REHG J. M., JAMPANI V.: Spar3d: Stable point-aware reconstruction of 3d objects from single images. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (2025), pp. 16860–16870. 1, 5

[HW23] HE Z., WANG T.: Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. Accessed: 2025-09-26. 5

[HZG*23] HONG Y., ZHANG K., GU J., BI S., ZHOU Y., LIU D., LIU F., SUNKAVALLI K., BUI T., TAN H.: LRM: Large reconstruction model for single image to 3D. *arXiv preprint arXiv:2311.04400* (2023). URL: https://arxiv.org/abs/2311.04400, arXiv:2311.04400. 1, 2, 5

[JHP24] JIANG H., HUANG Q., PAVLAKOS G.: Real3d: Scaling up large reconstruction models with real-world images. 1

[KNS*25] KANG G., NAM S., SUN X., KHAMIS S., MOHAMED A., PARK E.: ilrm: An iterative large 3d reconstruction model. *arXiv preprint arXiv:2507.23277* (2025). 3

[Lab24] LABS B. F.: Flux. https://github.com/black-forest-labs/flux, 2024. 1

[LGL*24] LONG X., GUO Y.-C., LIN C., LIU Y., DOU Z., LIU L., MA Y., ZHANG S.-H., HABERMANN M., THEOBALT C., ET AL.: Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2024), pp. 9970–9980. 3

[LLL*24] LI P., LIU Y., LONG X., ZHANG F., LIN C., LI M., QI X., ZHANG S., XUE W., LUO W., ET AL.: Era3d: High-resolution multiview diffusion using efficient row-wise attention. *Advances in Neural Information Processing Systems 37* (2024), 55975–56000. 3

[LTZ*23] LI J., TAN H., ZHANG K., XU Z., LUAN F., XU Y., HONG Y., SUNKAVALLI K., SHAKHNAROVICH G., BI S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214* (2023). 1, 2

[LWC*25] LI Z., WANG D., CHEN K., LV Z., NGUYEN-PHUOC T., LEE M., HUANG J.-B., XIAO L., ZHU Y., MARSHALL C. S., ET AL.: Lirm: Large inverse rendering model for progressive reconstruction of shape, materials and view-dependent radiance fields. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (2025), pp. 505–517. 3

[LWVH*23] LIU R., WU R., VAN HOORICK B., TOKMAKOV P., ZAKHAROV S., VONDRICK C.: Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision* (2023), pp. 9298–9309. 1, 3

[LXJ*23] LIU M., XU C., JIN H., CHEN L., VARMA T M., XU Z., SU H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems 36* (2023), 22226–22246. 3

[PLS25] PATEL A., LAGA H., SHARMA O.: Normal-guided detail-preserving neural implicit function for high-fidelity 3d surface reconstruction. *Proceedings of the ACM on computer graphics and interactive techniques 8*, 1 (2025), 1–24. 3

[SCZ*23] SHI R., CHEN H., ZHANG Z., LIU M., XU C., WEI X., CHEN L., ZENG C., SU H.: Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110* (2023). 2, 3

[SMH*23] SHEN T., MUNKBERG J., HASSELGREN J., YIN K., WANG Z., CHEN W., GOJCIC Z., FIDLER S., SHARP N., GAO J.: Flexible isosurface extraction for gradient-based mesh optimization. *ACM Trans. Graph. 42*, 4 (jul 2023). URL: https://doi.org/10.1145/3592430, doi:10.1145/3592430. 4

[SWY*23] SHI Y., WANG P., YE J., LONG M., LI K., YANG X.: Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512* (2023). 1

[SZW*25] SHEN Y., ZHOU K., WANG H., YANG Y., SHAO T.: High-fidelity 3d object generation from single image with rgbn-volume gaussian reconstruction model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2025), pp. 21558–21569. 3

[TCC*24] TANG J., CHEN Z., CHEN X., WANG T., ZENG G., LIU Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision* (2024), Springer, pp. 1–18. 1, 5

[TPL*24] TOCHILKIN D., PANKRATZ D., LIU Z., HUANG Z., LETTS A., LI Y., LIANG D., LAFORTE C., JAMPANI V., CAO Y.-P.: Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151* (2024). 2

[TWZ*23] TANG J., WANG T., ZHANG B., ZHANG T., YI R., MA L., CHEN D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF international conference on computer vision* (2023), pp. 22819–22829. 3

[VYB*24] VOLETI V., YAO C.-H., BOSS M., LETTS A., PANKRATZ D., TOCHILKIN D., LAFORTE C., ROMBACH R., JAMPANI V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision* (2024), Springer, pp. 439–457. 3

[WS23] WANG P., SHI Y.: Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201* (2023). 1

[WTB*23] WANG P., TAN H., BI S., XU Y., LUAN F., SUNKAVALLI K., WANG W., XU Z., ZHANG K.: Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024* (2023). 1, 3

[WZB*24] WEI X., ZHANG K., BI S., TAN H., LUAN F., DESCHAINTRE V., SUNKAVALLI K., SU H., XU Z.: Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385* (2024). 1

[WZF*23] WU T., ZHANG J., FU X., WANG Y., REN J., PAN L., WU W., YANG L., WANG J., QIAN C., ET AL.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 803–814. 4

[XBS*24] XIE D., BI S., SHU Z., ZHANG K., XU Z., ZHOU Y., PIRK S., KAUFMAN A., SUN X., TAN H.: Lrm-zero: Training large reconstruction models with synthesized data. *Advances in Neural Information Processing Systems 37* (2024), 53285–53316. 1

[XCG*24] XU J., CHENG W., GAO Y., WANG X., GAO S., SHAN Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191* (2024). 1, 2, 3, 5

[ZHW*24] ZHUANG P., HAN S., WANG C., SIAROHIN A., ZOU J., VASILKOVSKY M., SHAKHRAI V., KOROLEV S., TULYAKOV S., LEE H.-Y.: Gtr: Improving large 3d reconstruction models through geometry and texture refinement. *arXiv preprint arXiv:2406.05649* (2024). 3

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (2018). 4

[ZPG*24] ZHENG X.-Y., PAN H., GUO Y.-X., TONG X., LIU Y.: Mvdˆ 2: Efficient multiview 3d reconstruction for multiview diffusion. In *ACM SIGGRAPH 2024 conference papers* (2024), pp. 1–11. 3