Image2Life: A Model for 3D Mesh Reconstruction from a Single-Image

Lynda Ayachi[1] and Mohamed Rabia Benarbia[2]
[1]*Orange Innovation Tunisia, Sofrecom, Tunis, Tunisia*
[2]*National School of Computer Science, Manouba, Tunisia*

Keywords: 3D Reconstruction, Advanced Modeling Techniques, Large Reconstruction Model LRMs, Feature Aggregation.

Abstract: Reconstructing 3D models from a single 2D image is a complex yet fascinating challenge with applications in areas like computer vision, robotics, and augmented reality. In this work, we propose a novel approach to tackle this problem, focusing on creating accurate and detailed 3D representations from minimal input. Our model combines advanced deep learning techniques with geometry-aware methods to extract and translate meaningful features from 2D images into 3D shapes. By introducing a new framework for feature extraction and a carefully designed decoding architecture, our method captures intricate details and improves the overall reconstruction quality. We tested the model extensively on well-known datasets, and the results show significant improvements compared to existing methods in terms of accuracy and reliability.

1 INTRODUCTION

The art of creating a digital representation of any object, real or imagined in three dimensions, is the fundamental component of 3D modeling. Vertices , edges and faces must be placed precisely during this process. The end product is a flexible 3D model that can be animated, modified, and generated to suit a variety of purposes. 3D modeling has reshaped how industries create, visualize, and interact with digital and physical spaces. In the entertainment world, it is at the heart of storytelling, allowing studios such as Pixar and DreamWorks to craft realistic characters and detailed environments for animations and films (Whizzy Studios, 2023). Similarly, the video game industry depends on 3D models to design complex worlds and dynamic characters, enhancing the overall gaming experience (Cutting Edge R, 2023). In architecture and construction, 3D modeling provides architects and engineers with tools to visualize designs, conduct virtual walkthroughs, and efficiently plan projects. This approach helps to communicate ideas clearly and facilitates better decision-making during the design process (Adobe, 2023). The integration of 3D modeling into virtual reality (VR) and augmented reality (AR) has expanded its applications even further. VR uses 3D models to create fully immersive digital environments, while AR overlays virtual objects onto the physical world, enhancing areas such as education, training, and interactive experiences (Cutting Edge R, 2023). In the medical field, 3D modeling is used to generate highly detailed anatomical representations that support medical education, surgical planning, and prosthetic design. These models improve understanding of complex biological structures and contribute to better patient outcomes (Whizzy Studios, 2023). Also, the product design and manufacturing industries leverage 3D modeling to speed up prototyping, enable greater customization, and streamline production. By visualizing and refining designs before manufacturing, 3D modeling helps reduce costs and improve efficiency (Adobe, 2023). Our solution uses generative AI to create detailed 3D models from 2D images, primarily optimized for electronics and gadgets while maintaining effectiveness across other object types. This approach enhances reconstruction fidelity and realism while providing a more efficient and accessible alternative to traditional 3D modeling techniques.

2 RELATED WORK

This section will explore advanced techniques in 3D model generation, focusing on three main areas: Score Distillation for 3D Generation, 3D Generation with Sparse View Reconstruction, and Feed-forward 3D Generative Models.

Ayachi, L. and Benarbia, M. R.
Image2Life: A Model for 3D Mesh Reconstruction from a Single-Image.
DOI: 10.5220/0013349500003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1319-1326
ISBN: 978-989-758-737-5; ISSN: 2184-433X

- **Score Distillation for 3D Generation:**

Score distillation leverages large-scale image diffusion models to iteratively refine and produce high-quality 3D models without requiring extensive 3D datasets. DreamFusion (Poole et al., 2022) pioneered this approach with Score Distillation Sampling (SDS), which uses feedback from image diffusion models to align 3D models with target images or descriptions. This iterative refinement process enables highly detailed results without relying on pre-existing 3D data. ProlificDreamer (Wang et al., 2023) builds on this foundation by introducing Variational Score Distillation (VSD), addressing over-saturation issues and enhancing output diversity. However, these techniques face significant computational challenges - generating a single model on an NVIDIA RTX 4090 GPU can take several minutes with DreamFusion, while complex objects may require hours of processing time, limiting practical applications where speed is crucial.

- **3D Generation with Sparse View Reconstruction:**

Recent advancements in 3D reconstruction have introduced methods that generate multi-view consistent images from single-view inputs, which are then transformed into detailed 3D models. SyncDreamer (Liu et al., 2023c) employs a synchronized multi-view diffusion model to produce images with consistent geometry and color across various perspectives, facilitating accurate 3D reconstruction without extensive training data. Building upon this, Wonder3D (Long et al., 2023) utilizes a cross-domain diffusion approach to generate multi-view normal maps and corresponding color images, ensuring coherence across different views. This method effectively handles complex scenes and textures, resulting in high-fidelity textured meshes from single-view images. Zero123++ (Shi et al., 2023) further enhances multi-view image generation by leveraging refined diffusion techniques to manage dynamic lighting and shadows, producing realistic and cohesive images from diverse viewpoints. For the reconstruction phase, frameworks like NeuS and Neuralangelo have demonstrated significant capabilities. NeuS (Liu et al., 2023c) specializes in reconstruction of the neural surface using a signed distance function to implicitly represent geometry, which is particularly beneficial in handling sparse views and self-occlusions. Neuralangelo (Liu et al., 2023a) advances 3D surface reconstruction by employing multi-resolution 3D hash grids and neural surface rendering, achieving high fidelity through coarse-to-fine optimization strategies.

Despite these advancements, challenges persist, notably the computational intensity required for processing. For instance, NeuS requires approximately two hours to process a single object on an NVIDIA RTX 4090 GPU, while Neuralangelo takes around 40 minutes. Sparse view inputs can lead to incomplete reconstructions, and ensuring geometric consistency across different views remains critical.

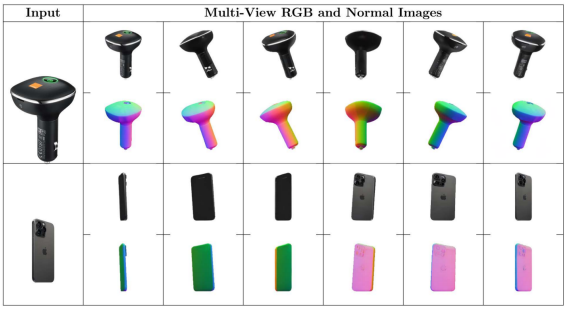

Figure 1: Wonder3D Multi-View Results.

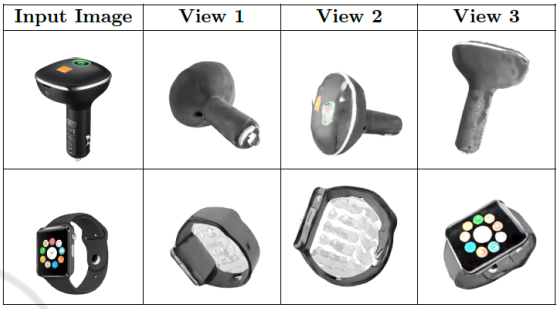

Figure 2: Examples of 3D Models Reconstructed by NeuS.

- **Feed-forward 3D Generative Models:**

Feed-forward 3D generative models offer a faster and more efficient alternative for generating 3D objects, bypassing the iterative optimization processes typically used in traditional methods. By employing a direct feed-forward approach, these models drastically reduce the time required for 3D object generation, making them particularly valuable for applications where rapid results are essential. Recent advancements in this domain have demonstrated significant progress, aided by extensive datasets like Objaverse (Deitke et al., 2023), which provides diverse and robust training data.

One notable example is **One-2-3-45** (Liu et al., 2023b), which transforms single images into detailed textured 3D meshes using Zero123 diffusion model for multi-view prediction and SparseNeuS for neural surface reconstruction. It generates results in 45 seconds without iterative optimization. The Large Reconstruction Model (LRM) (Hong et al., 2023) advances feed-forward 3D generation through transformer-based architecture with refined attention mechanisms, generating high-quality 3D objects in 10 seconds on an NVIDIA RTX 4090 GPU. While

the official model is private . Building on LRM, **Triposr** (Tochilkin et al., 2023) enhances its performance through improved data preprocessing, model optimization, and training strategies while maintaining the core LRM architecture. These refinements allow it to achieve superior reconstruction quality in approximately the same inference time.

3 Image2Life MODEL ARCHITECTURE

In this section, we present the detailed architecture of our model, illustrated in figure 3, highlighting its components and their functionalities.

3.1 Image Encoder

The DINO (Self-Distillation with No Labels) model (Touvron et al., 2023) is utilized as the image encoder within the architecture due to its ability to learn robust visual representations without requiring labeled data. Specifically, the integration of DINOv2 (Touvron et al., 2023), an enhanced version of DINO, addresses the demands of dense visual predictions and fine-grained feature extraction. Object views are processed through the encoder to extract feature vectors that encapsulate critical visual details, forming the foundation for subsequent stages in the 3D reconstruction pipeline. The DINOv2 model incorporates an optimized self-distillation process that significantly enhances its proficiency in capturing intricate details and high-resolution features. Input images are segmented into patches and passed through a multilayer transformer comprising 12 layers, each with a dimensionality of 768 (Touvron et al., 2023). This process generates detailed feature vectors that represent the visual properties of each patch while preserving a coherent understanding of the entire image. The maintenance of spatial coherence is ensured through the use of register tokens, which uphold spatial relationships within the image. This mechanism is critical for preserving structural integrity, allowing the reconstructed 3D model to accurately reflect the input image. The self-distillation approach employed by DINOv2 provides interpretable attention over image structures and textures, enabling detailed structural information to guide the reconstruction of geometry and color in 3D space.

3.2 Feature Aggregator

The Multi-Branch Attention Aggregator is a critical module that we designed to integrate features from multiple views into a unified representation, leveraging the complementary information provided by each view. This architecture is particularly effective in scenarios where different perspectives offer unique insights, enhancing the quality of the aggregated features for downstream tasks. Inspired by the patch-based processing of the DINO encoder (Touvron et al., 2023), the Feature Aggregator employs a multi-branch attention mechanism to efficiently combine information from various views. Each view is processed through a dedicated attention layer comprising sequential linear layers interspersed with ReLU activation functions, enabling each branch to independently highlight important features. Attention weights are dynamically calculated using a softmax layer, which ensures that the most relevant features across all views are emphasized. The aggregated features are subsequently refined through a final fully connected output layer, which adjusts the dimensionality to suit specific application requirements. This architecture offers enhanced feature representation by adaptively focusing on the most informative aspects of each view, resulting in a richer, more robust output. Additionally, the modular design of the aggregator allows for flexibility in the number of views and the dimensionality of the features, making it highly adaptable across various settings. This approach significantly improves the efficiency and effectiveness of multi-view feature integration, aligning well with modern 3D reconstruction pipelines.

3.3 Image-to-Triplane Decoder

The Image-to-Triplane Decoder plays a critical role in converting high-dimensional image features into structured 3D representations. By utilizing a transformer-based architecture, this decoder projects image features onto learnable spatial-positional embeddings, effectively mapping them to a triplane structure. This approach addresses the inherent challenges of single-image 3D reconstruction by providing a comprehensive representation of geometric and appearance information. The triplane representation consists of three planes aligned with the axes: T_{XY}, T_{YZ}, T_{XZ}, each with a spatial resolution of 64×64 and d_T feature channels. Features for any 3D point within the NeRF bounding box $[-1,1]^3$ are derived through bilinear interpolation, followed by processing with a multi-layer perceptron (MLP) to generate NeRF color and density values (Mildenhall et al., 2020). Learnable spatial-positional embeddings are integrated into the decoder, with dimensions $3 \times 32 \times 32 \times d_D$, to bridge the gap between image and 3D space. These embeddings query im-

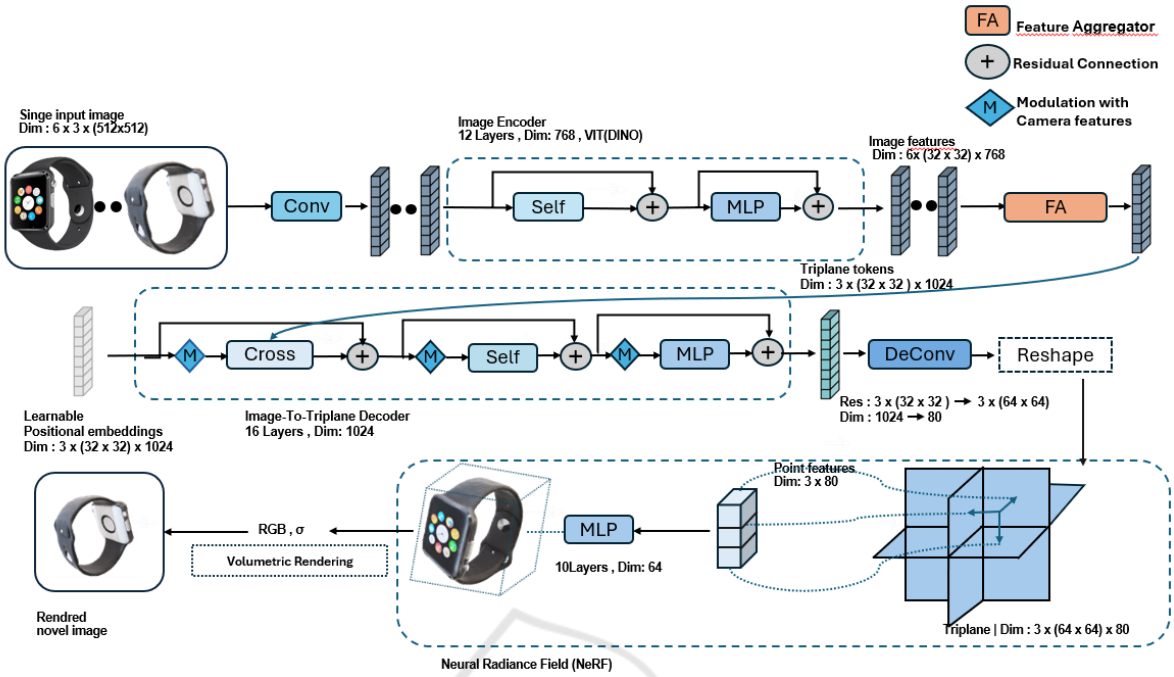

Figure 3: Image2Life Architecture.

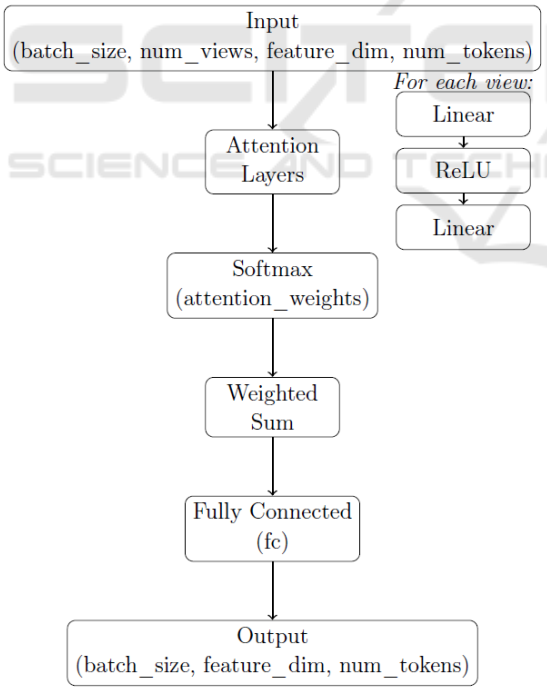

Figure 4: Architecture of The Feature Aggregator.

age features via cross-attention mechanisms, enabling the projection of image features to the triplane structure. Given the disparity in token counts between the initial embeddings $3 \times 32 \times 32$ and the final triplane representation $3 \times 64 \times 64$, the transformer output is upsampled to complete the triplane representation. The decoder employs a transformer architecture beginning with a token sequence of dimensions $(3 \times 32 \times 32) \times 1024$, where the token count $3 \times 32 \times 32$ corresponds to the number of spatial tokens and 1024 represents the hidden dimension of the transformer. This configuration includes 1025 image feature tokens for generating keys and values in the cross-attention layers. The transformer consists of 16 layers, each with 16 attention heads of dimension 64. In accordance with methods proposed by Touvron et al. (Touvron et al., 2023), bias terms are omitted, and a pre-normalization structure is employed, defined as: $x + f(\{LayerNorm\}(x))$. Each transformer layer consists of a cross-attention sub-layer to integrate image and triplane features, a self-attention sub-layer to capture internal triplane relationships, and an MLP sub-layer for non-linear feature transformations. The operations are defined as (Hong et al., 2023):

$$f^j_{\text{cross}} = \text{CrossAttn}\big(\text{ModLN}_{\text{cam}}(\text{LayerNorm}(f^j_{\text{in}})); \{h_i\}_{i=1}^n\big) + f^j_{\text{in}} \tag{1}$$

$$f^j_{\text{self}} = \text{SelfAttn}\big(\text{ModLN}_{\text{cam}}(\text{LayerNorm}(f^j_{\text{cross}}))\big) + f^j_{\text{cross}} \tag{2}$$

$$f^j_{\text{out}} = \text{MLP}_{\text{tfm}}(\text{LayerNorm}(f^j_{\text{self}})) + f^j_{\text{self}} \tag{3}$$

This combination of components ensures a robust and detailed 3D representation derived from high-dimensional image features.

Then, camera features are constructed by flattening the 4×4 extrinsic matrix E and concatenating it with the focal length foc and principal point pp, creating a 20-dimensional vector $\mathbf{c}$. This vector is normalized and then transformed into a high-dimensional embedding $\tilde{\mathbf{c}}$ via an MLP. These camera features are used to modulate the triplane tokens through adaptive layer normalization. The true camera parameters $\mathbf{c}$ are incorporated into the Image-to-Triplane decoder only during the training phase. In the inference phase, these features are substituted with an encoding of the standard, fixed camera parameters.

$$\gamma, \beta = \text{MLP}_{\text{mod}}(\tilde{\mathbf{c}}) \tag{4}$$

$$\text{ModLN}_{\mathbf{c}}(f_j) = \text{LayerNorm}(f_j) \cdot (1+\gamma) + \beta \tag{5}$$

After processing through all transformer layers, the final triplane features f_{out} are upsampled using a learnable de-convolution layer to form the final triplane representation T. This layer transforms the transformer output from $(3 \times 32 \times 32) \times 1024$ to $3 \times (64 \times 64) \times 80$. The de-convolution layer has a kernel size of 2, a stride of 2, and no padding.

The Triplane-NeRF (Neural Radiance Fields) component is designed to estimate RGB values and density σ from point features derived from the triplane representation T. This approach is based on the triplane framework described by (Chan et al., 2022), utilizing a multi-layer perceptron (MLP) to process these point features effectively. The architecture of the MLP consists of three main parts. First, the input comprises point features extracted from the triplane representation T. These features are then passed through 10 hidden layers, each employing linear transformations interleaved with ReLU activation functions and having a dimensionality of 64. Finally, the output layer generates a 4-dimensional vector, where the first three dimensions represent the RGB color values, and the fourth corresponds to the density σ. This structured design enables the MLP to process the input features and produce precise estimations of the color and density values.

In terms of the detailed process, querying point features is a critical step. For each point within the 3D bounding box, the point is mapped onto the triplane planes (T_{XY}, T_{YZ}, T_{XZ}). Bilinear interpolation is then applied to extract feature values from these planes, ensuring accurate and efficient computation for 3D representation tasks.

The features obtained from the three planes are aggregated into a single feature vector that represents the 3D point. This combined feature vector is then processed using the MLP_{NeRF}, which consists of a series of 10 linear layers with ReLU activation functions. The output from the MLP_{NeRF} is a 4-dimensional vector, where the first three dimensions correspond to the RGB color values, and the fourth dimension represents the density σ. This structured pipeline ensures the efficient transformation of the triplane features into meaningful 3D representations suitable for rendering.

The design of this framework offers several advantages. First, the triplane representation provides a compact yet expressive encoding of 3D structures, enabling efficient querying of point features. Second, the use of MLP_{NeRF} facilitates accurate predictions of color and density, resulting in high-quality 3D reconstructions and renderings. Finally, the modular nature of the triplane and MLP components ensures scalability to complex scenes and objects, making it a versatile solution for diverse 3D tasks.

4 EXPERIMENTATIONS

4.1 Training

The training process for the proposed model is conducted in two stages.

Stage 1: The first stage focuses on training the feature aggregator, a critical component of the model's architecture. This stage involves fine-tuning key training parameters to enhance the model's efficiency and ensure robust feature extraction and integration. The training process incorporates techniques and parameters derived from the LRM framework, complemented by improvements inspired by recent advancements in Triposr.

To optimize the balance between computational efficiency and reconstruction quality, instead of rendering full-resolution images at a 512×512 resolution, the model processes smaller 128×128 patches, prioritizing patches more likely to cover foreground regions. This approach ensures that the model focuses on critical areas, enabling detailed surface reconstructions while maintaining computational efficiency. The rendering process was configured with 128 samples per ray, a radius of 0.87, an exponential density activation function, and a density bias of -1.0. The model was trained over 27 epochs .

Stage 2: In the second stage, we fine-tune the last five layers of the Image-to-Triplane Decoder and the Triplane-NeRF while freezing the rest of the model to

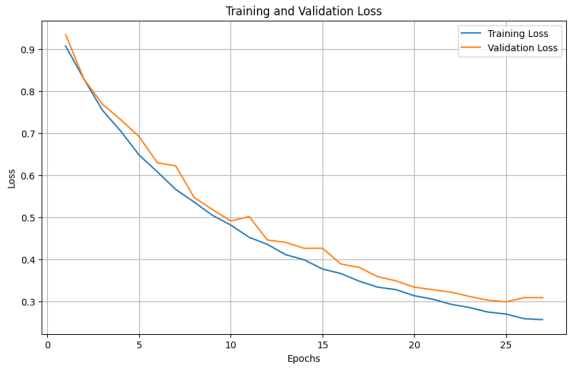

Figure 5: Training and Validation Loss During Stage 1.

adapt it to specific tasks. This stage uses the same hyperparameters and renderer settings as the first stage, ensuring consistency and stability during training.

4.2 Evaluation

4.2.1 Qualitative Results

We conducted qualitative comparisons to evaluate our model's performance against established benchmarks.

Figure 6 compares OpenLRM and our model with only the feature aggregator trained, using images sourced from OpenLRM and the MyTek website. The results demonstrate that our model, leveraging Zero123++ (Shi et al., 2023), generates more accurate and superior views.

Figures 7 compare our fine-tuned model with OpenLRM on electronic items from retailers like Amazon and Orange Store. Our model consistently delivers more realistic geometry and appearance, significantly improving results on electronic objects while generalizing well across various categories. In Figures 8 comparisons with baselines such as CRM (Wang et al., 2024) and One-2-3-45 highlight our model's advantages.

TripoSR struggles with imaginative capabilities and often produces degraded textures, while Image2Life use of 80 channels in the Triplane-NeRF architecture ensures detailed geometry and textures.

CRM, while recent, fails to generate smooth surfaces, whereas our model, aided by mask loss during training, excels in smooth and realistic reconstructions.

4.2.2 Quantitative Results

To evaluate the quantitative performance of our model, we conducted an extensive analysis using 50 unseen objects from our curated dataset and an additional 50 objects from the High-Quality Alignment Subset of Objaverse-XL. For each object, 15 reference views were processed, and five of these views

Figure 6: Comparison between OpenLRM and our model with only the feature aggregator trained.

were individually used as input for our model to reconstruct the corresponding 3D object. The rendered images were then evaluated against all 15 reference views to assess the quality and accuracy of the reconstruction. This evaluation methodology, inspired by the approach utilized in the LRM framework, provides a robust assessment of our model's reconstruction precision and robustness. Notably, our model demonstrates significant computational efficiency, achieving high-quality object generation within just 10 seconds. The rendered images were captured from the following angles:

As shown in Table 2, Image2Life surpasses baseline models in 2D novel view synthesis metrics, achieving higher SSIM and LPIPS scores, which highlight its superior perceptual quality. It also leads in FID and CLIP-Similarity, demonstrating its ability to produce visually appealing and contextually accurate images. While its PSNR is slightly lower than the best-performing baseline, this is attributed to the

Figure 7: Comparison of fine-tuned model and OpenLRM on electronic objects.

Table 1: Rendered View Angles (Az: Azimuth, El: Elevation).

View	Az	El	View	Az	El
Back	180°	0°	Right	270°	0°
Back Left	135°	0°	Top	0°	90°
Back Right	225°	0°	Top Back	180°	45°
Front	0°	0°	Top Front	0°	45°
Front Left	45°	0°	Top Left	90°	45°
Front Right	315°	0°	Top Right	270°	45°
Left	90°	0°			

multi-view diffusion model (Zero123++) generating less pixel-perfect but perceptually richer views. We argue that perceptual quality outweighs pixel-level accuracy for novel view synthesis, given the inherent variability of such views.

Figure 8: Comparison of Our Model with One-2-3-45 and CRM.

For 3D geometric metrics, Image2Life markedly outperforms baselines in both Chamfer Distance (CD) and F-Score (FS), demonstrating superior shape fidelity. The addition of mask loss during training effectively minimizes 'floater' artifacts, resulting in a significant improvement in CD.

5 CONCLUSION

In conclusion, our model consistently demonstrates exceptional performance in generating realistic geometry and visually appealing appearances, particularly for electronic objects sourced from various online retailers such as Amazon and Orange Store. Its robust generalization capabilities extend beyond electronic items, achieving superior results across a diverse range of object categories. Qualitative comparisons have shown that our fine-tuned model outperforms baseline approaches by producing smoother and more accurate reconstructions.

Comprehensive quantitative evaluations further validate the effectiveness of our approach. Our model excels in key perceptual quality metrics such as SSIM, LPIPS, FID, and CLIP-Similarity, while also achieving state-of-the-art performance in geometric fidelity metrics like Chamfer Distance (CD) and F-Score (FS).

Despite these achievements, there is room for improvement. Addressing challenges such as textural in-

Table 2: Quantitative comparison of our method with baseline models across multiple metrics. ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	FS ↑	FID ↓	CLIP-Similarity ↑
TripoSR	23.681	0.872	0.204	0.246	0.879	25.459	0.812
DreamGaussian	19.204	0.789	0.277	0.382	0.635	57.257	0.815
CRM	22.790	0.891	0.137	0.201	0.802	23.846	0.880
One-2-3-45	18.558	0.726	0.296	0.421	0.633	98.261	0.720
Ours	23.297	0.901	0.124	0.179	0.892	22.547	0.930

Table 3: Comparison of our model with only the feature aggregator trained against OpenLRM using CD and LPIPS metrics.

Method	CD ↓	LPIPS ↓
OpenLRM	0.271	0.209
Our Model (Only Feature Aggregator Trained)	0.194	0.153

distinctness in occluded regions and reducing the dependency on the quality of multi-view images generated by Zero123++ are crucial next steps. Enhancing the initial stages of multi-view image generation or developing alternative strategies could further boost the performance and reliability of the 3D reconstruction process. These advancements will pave the way for even more robust and precise 3D reconstruction capabilities in future work.

REFERENCES

Adobe (2023). What is 3d modelling & what is it used for?

Chan, E. R. et al. (2022). Efficient triplane-nerf for 3d object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cutting Edge R (2023). 10 exciting applications of 3d modeling in various industries.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. (2023). Objaverse: A universe of annotated 3d objects. pages 13142–13153.

Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., and Tan, H. (2023). Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Liu, J., Zhang, Z., Wang, X., Li, S., Zhang, Z. Y., Yang, M.-Y., Kautz, J., Hilliges, O., and Tulsiani, S. (2023a). Neuralangelo: High-fidelity neural surface reconstruction. *arXiv preprint arXiv:2306.03092*.

Liu, M., Xu, C., Jin, H., Chen, L., Varma, M. T., Xu, Z., and Su, H. (2023b). One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*.

Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., and Wang, W. (2023c). Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Long, X., Liu, Y., Lin, C., Zeng, Z., Liu, L., Komura, T., and Wang, W. (2023). Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, pages 405–421.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., and Su, H. (2023). Zero123++: A single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*.

Tochilkin, A. et al. (2023). Triposr: High-efficiency 3d reconstruction from minimal data inputs. *arXiv preprint arXiv:2308.12045*.

Touvron, H., Bojanowski, P., Caron, M., Misra, I., Mairal, J., and Joulin, A. (2023). Dinov2: Learning robust visual features without labels. *arXiv preprint arXiv:2304.07193*.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. (2023). Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., Li, C., Su, H., and Zhu, J. (2024). Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*.

Whizzy Studios (2023). Applications of 3d modeling.