MeTTA: Single-View to 3D Textured Mesh Reconstruction with Test-Time Adaptation

Kim Yu-Ji[1]
ugkim@postech.ac.kr
Hyunwoo Ha[2]
hyunwooha@postech.ac.kr
Kim Youwang[2]
youwang.kim@postech.ac.kr
Jaeheung Surh[3]
jh.surh@bucketplace.net
Hyowon Ha[3,†]
hyowon.ha@bucketplace.net
Tae-Hyun Oh[1,2,4,†]
taehyun@postech.ac.kr

[1] Grad. School of AI
POSTECH, South Korea
[2] Dept. of Electrical Engineering
POSTECH, South Korea
[3] Bucketplace, Co., Ltd., South Korea
[4] Institute for Convergence
Research and Education
in Advanced Technology
Yonsei University, South Korea

Abstract

Reconstructing 3D from a single view image is a long-standing challenge. One of the popular approaches to tackle this problem is learning-based methods, but dealing with the test cases unfamiliar with training data (Out-of-distribution; OoD) introduces an additional challenge. To adapt for unseen samples in test time, we propose `MeTTA`, a test-time adaptation (TTA) exploiting generative prior. We design joint optimization of 3D geometry, appearance, and pose to handle OoD cases with only a single view image. However, the alignment between the reference image and the 3D shape via the estimated viewpoint could be erroneous, which leads to ambiguity. To address this ambiguity, we carefully design learnable virtual cameras and their self-calibration. In our experiments, we demonstrate that `MeTTA` effectively deals with OoD scenarios at failure cases of existing learning-based 3D reconstruction models and enables obtaining a realistic appearance with physically based rendering (PBR) textures.

1 Introduction

Understanding 3D scenes and objects from a single-view image is a long-standing fundamental challenge in computer vision [26]. It becomes particularly crucial in robotics for machine perception, extended reality systems for AR/VR, and virtual communication. They need the ability to comprehend and interact with the real 3D world. Moreover, representing real 3D scenes requires not only geometric accuracy but also realistic and physically-based properties, essential for creating lifelike and interactive virtual environments [3, 49].

† denotes corresponding authors.

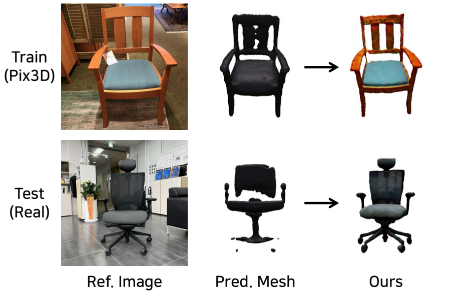


Figure 1: **Distribution gap between train and test.** "Train" refers to a sample on which the Image-to-3D is trained, and "Test" is an in-the-wild sample we captured.

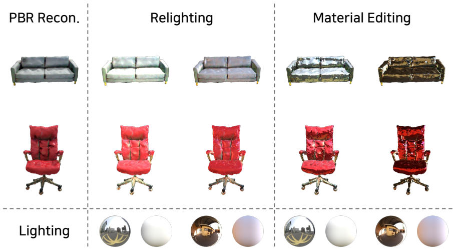


Figure 2: **Practical applications in graphics.** "PBR Recon." means reconstruction results with PBR textures by ours.

There have been growing efforts to understand holistic 3d scenes, *e.g.*, layout, object pose, and mesh, from a single-view image [4, 13, 23, 31, 50]. These methods operate effectively by utilizing a learning-based feed-forward approach with reasonable coarse geometry and viewpoint estimation when only given the single-view reference image. However, the feed-forward methods have the inherent limitation that they cannot perform well on real-world test images away from trained distribution. Those methods rely on training with {2D image, 3D shape}-paired datasets [11, 43], which have narrow data distribution compared to the tremendous diversity of real objects. It is infeasible to construct a large-scale dataset that covers such diversity, considering the difficulty and labor-intensive process of real 3D data acquisition. Thus, feed-forward methods trained on such a limited dataset can only learn the narrow expressivity of 3D shapes, as shown in "Pred. Mesh" of Fig. 1. It hints the vulnerability of such feed-forward models to out-of-distribution (OoD) cases.

To address this challenge, we propose `MeTTA`, a test-time adaptation (TTA) method for 3D reconstruction by utilizing only a single reference view image. To compensate for the limited information of single-view, we leverage a pre-trained multi-view generative model [25] as a prior. Given a single-view image, we obtain initial mesh and viewpoint predictions from the existing feed-forward model. We then design joint optimization of the mesh, texture, and camera viewpoint to deal with OoD cases. However, alignments between the reference image and the 3D mesh from the estimated viewpoint are not exactly matched, which may lead to erroneous results. To mitigate this, we propose carefully designed learnable virtual cameras with the self-calibrating method to align the 2D pixel information with the 3D shape by updating the initial guess of the viewpoint estimation.

In addition, we parameterize the texture map with physically based rendering (PBR) parameters, including diffuse, specularity, and normal. This enables us to utilize our results in off-the-shelf graphics tools, *e.g.*, Blender [8]; thereby ours can be facilitated to editing for relighting and material control as shown in Fig. 2. This is an underexplored feature in previous holistic 3D scene understanding researches [4, 13, 23, 31, 50] that predominantly focus on shapes and poses of objects, where we extend to output material property, texture, and mesh complying with input reference image.

Our key contributions are summarized as follows:

- We propose `MeTTA`, which closes the domain gap between training and test time by jointly updating mesh, texture, and viewpoint with the aid of the generative model prior.

- We design viewpoint self-calibration and textured mesh reconstruction using only a single view reference image.
- We achieve high-fidelity geometry along with a realistic appearance with physically based rendering (PBR) textures, which can be compatible with real graphics engines.

2 Related Work

Our task is related to the feed-forward reconstruction methods at single-view and the iterative test-time adaptation aided by a generative prior. We briefly review these lines of work.

Feed-forward reconstruction methods. This task aims to reconstruct 3D mesh from a single-view image captured in a real-world environment [48, 52]. A line of work [13, 23, 31, 47, 50] have proposed learning-based models that reconstruct image-aligned 3D meshes and poses of objects from a single 2D image. While they could reconstruct the geometry of objects of given single-view image in an feed-forward manners, they are vulnerable to out-of-distribution (OoD) scenarios beyond the training dataset. The out-of-distribution cases for this task are common since the intricacy and the diversity of object shapes in a real-world environment are too complicated to be learned from the limited scale and diversities of existing {2D image, 3D shape}-aligned and -paired datasets [7, 11, 21, 43]. Moreover, these methods could not represent the texture. A recent work [4] has explored the reconstruction of 3D mesh and texture from a single image. However, their feed-forward estimation of shape and texture also could not generalize to real-world cases. Also, the model only estimates the RGB color and does not model the physically based rendering (PBR) characteristics, which may limit the realism of the reconstructed texture.

Iterative reconstruction methods using generative priors. Recent advances in the field of 2D generative models [1, 2, 10, 32, 35, 36, 38, 39] have shown remarkable capabilities as the prior for 2D inverse problems [5, 6, 15, 41]. For our task of single-view 3D textured mesh reconstruction, prior knowledge about 3D object geometry and textures is mandatory to embody a test-time adaptability for OoD cases. However, directly constructing a 3D object geometry or appearance prior is challenging, considering its unmeasured diversity.

A seminal work, DreamFusion [33] unlocked the capabilities of a pre-trained text-to-image diffusion model and proposed the Score-Distillation Sampling (SDS), which acts as a 2D generative prior for the 3D generation task [3, 14, 22, 45]. We exploit the idea of using a pre-trained generative model as a prior for 3D tasks. Specifically, we propose to use a multi-view diffusion model [25] as a generative prior to mitigate the test-time distribution shift of the 3D shape, texture and poses. Additionally, recently proposed feed-forward reconstruction methods with generative priors [24, 46] also cannot model the realistic PBR properties.

3 Method

We first provide the overall `MeTTA` pipeline in Sec. 3.1. Following that, we explain how we obtain the coarse object geometry in Sec. 3.2 and align the virtual camera to match with the 2D single-view image in Sec. 3.3. We describe our test-time adaptation (TTA) process for 3D reconstruction in Sec. 3.4 and explain the details of texture representation in Sec. 3.5.

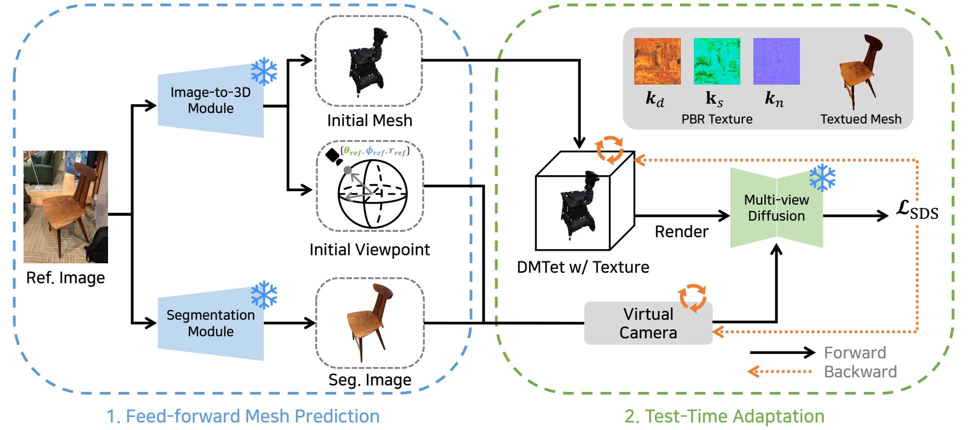
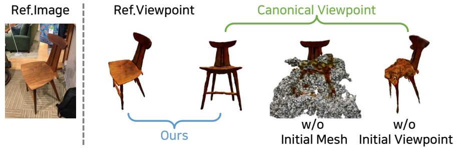


Figure 3: **Overview of MeTTA.** We propose a test-time adaptation pipeline to reconstruct a 3D mesh with PBR texture from a single-view image. "Ref. Image" refers to the reference input image. "Seg. Image" refers to the object-segmented image from "Ref. Image".



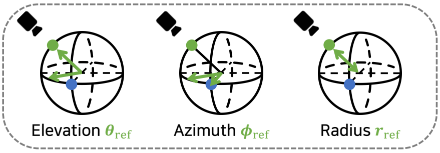


Figure 4: **Ablation studies.** To validate our pipeline design, we perform ablation studies where the initial mesh or viewpoint prediction is absent. In the case of a missing initial mesh, we initialize our 3D space with ellipsoid. Canonical viewpoint means that the azimuth and elevation angles are 0°.

Figure 5: **Learnable virtual camera.** The reference image is taken with viewpoint $(\theta_{\text{ref}}, \phi_{\text{ref}}, r_{\text{ref}})$, which we estimate and optimize. **Green dot** means predicted viewpoint given single-view image. **Blue dot** means canonical viewpoint with both elevation and azimuth angles are 0°.

3.1 Overall Pipeline

When provided with a single-view reference image during test time, we employ a feed-forward reconstruction method to obtain initial coarse shape and viewpoint predictions in the first stage (blue box) of Fig. 3. We update coarse geometry to fine-grained shape with realistic textures and viewpoints aligned with a 2D image in the second stage (green box) of Fig. 3. We utilize a multi-view diffusion model [25] to guide the adaptation process through Score-Distillation Sampling (SDS) loss [33]. We leverage the segmentation module [16, 17, 37] to obtain a white-background object image. The initial estimated viewpoint has an ambiguity between the 3D object and the reference image. To mitigate the vagueness, we assume a learnable virtual camera space with its self-calibration which aids in finding well-aligned 2D pixel to 3D space mapping, facilitating seamless adaptation. We demonstrate the effectiveness of our design, composed of both the initial feed-forward mesh and viewpoint prediction stage and the subsequent test-time adaptation stage, as illustrated in Fig. 4.

3.2 Feed-forward Initial Prediction

Given a single view input image, we first predict a coarse mesh and its viewpoint by the base Image-to-3D model. We can adapt a pre-trained 2D detector (e.g., Faster R-CNN [12]) into our system, ensuring that it encompasses the specific class we intend to reconstruct. We then integrate the separate 3D detection and mesh prediction networks that have the 2D detections as input and output SDF representation for mesh and its viewpoint for each object in the input scene, respectively. We train the 3D networks on the Pix3D [43] and SUN RGB-D [42] datasets. We refer to the whole pipeline as the base model [31, 50].

3.3 Learnable Virtual Camera

Recall that we obtain predictions for the initial mesh and camera viewpoint (e.g., radius, elevation and azimuth angles) using the feed-forward model. At test time, the camera parameters of camera focal length and pose parameters are unknown, leading to the ambiguity between 2D pixel information and 3D shape mapping. To address this ambiguity, we define a learnable virtual camera, where we set pre-defined camera intrinsics and adapt the extrinsic pose of the virtual camera. We need refinement to align the mapping because the viewpoint estimation from the previous step is just an initial guess and may be erroneous.

Getting aligned 3D mesh to 2D image observation is essential to utilize multi-view diffusion priors. In the pre-optimization stage, we set the initial viewpoint from these predictions and first update the radius of our virtual camera by optimizing the initial mesh rendering to be aligned with the reference image with mask loss. In the main optimization stage, we propose to self-calibrate the virtual camera pose by simultaneously optimizing our 3D mesh with PBR texture to achieve a more accurate alignment between the 2D image and the 3D space. We estimate and update the reference viewpoint $(\theta_{ref}, \phi_{ref}, r_{ref})$ to align between 2D reference image and the 3D shape, as shown in Fig. 5. This approach refines the mapping between a 2D image and 3D space and obtains consistent 3D results, which is vital for holistic scene reconstruction. Based on the reference viewpoint, we sample the relative viewpoint $(\Delta\theta, \Delta\phi, \Delta r)$ as a condition to the multi-view diffusion model [25].

3.4 Test-Time Adaptation for 3D Reconstruction.

We employ DMTet [40] as our 3D representation, which is characterized by two essential features; a deformable tetrahedral grid used to represent 3D shapes and a differentiable marching tetrahedral (MT) layer designed to extract explicit triangular meshes. DMTet has V_T vertices in the tetrahedral grid T, which can be expressed as (V_T, T).

DMTet initialization from coarse geometry. To model the geometry and texture of a 3D object, for each vertex $v_i \in V_T$, we learn the signed distance function (SDF) $s(v_i)$, vertex deformation offset Δv_i and per-vertex physically based rendering (PBR) material properties $\mathbf{k}_{\text{PBR}}$, with hash-grid positional encoding [29] function τ as follows:

$$[s(v_i), \Delta v_i, \mathbf{k}_{\text{PBR}}] = \Theta(\tau(v_i); \theta), \tag{1}$$

where MLP network Θ has the parameters θ. Before optimizing the target object from the reference image, we initialize DMTet with the initial shape obtained from the base model. From this initial mesh, we randomly sample a set of points $\{p_i \in \mathbb{R}^3\}$ where p_i represents a point in P which is the mesh vertices. We initialize the DMTet grid and its neural parameters to fit the initial mesh prediction by solving a SDF optimization problem as follows:

$$\theta^* = \arg\min_{\theta} \sum_{p_i \in P} \|s(\tau(p_i); \theta) - \text{SDF}(p_i)\|_2^2. \tag{2}$$

Using the pre-optimized network Θ and a differentiable renderer R, *e.g.*, Nvdiffrast [18], we obtain the RGB rendering image $\mathbf{x}$ as $\mathbf{x} = R(\theta, c)$, where c represents the sampled camera viewpoint. We randomly sample camera viewpoints within the range of [-45°, 45°] for the elevation angle and [0°, 360°] for the azimuth angle.

Jointly optimizing shape, texture & camera. Given the initialized DMTet and its corresponding MLP Θ, we proceed to adapt the shape, texture and the virtual camera pose jointly. To update Θ parameterized by θ, we utilize Score-Distillation Sampling (SDS) loss, which calculates per-pixel gradients by computing the difference between predicted noise and added noise as follows:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\psi, \mathbf{x}) = \mathbb{E}\left[w(t)(\varepsilon_\psi(\mathbf{z}_t; \mathbf{y}, t) - \varepsilon) \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \theta} \right], \tag{3}$$

where ψ parameterizes multi-view aware image diffusion model, $\mathbf{x}$ represents the RGB rendering output, $w(t)$ signifies a weight function for different noise levels, $\mathbf{z}_t$ denotes the latent encoding of $\mathbf{x}$ with the addition of noise ε, and ε_ψ is the predicted noise with reference image $\mathbf{y}$ and noise level t.

We leverage several additional loss terms to aid in the optimization. To promote the photometric consistency between the reference image and rendered textures of the 3D reconstruction, we introduce the photometric loss $\mathcal{L}_{\text{photo}} = ||I_{\text{ref}} - \mathbf{x}_{\text{ref}}||_1$ between the reference image I_{ref} and the rendering from the reference viewpoint $\mathbf{x}_{\text{ref}}$. Similar to the photometric loss, we also leverage the mask loss $\mathcal{L}_{\text{mask}} = ||M(I_{\text{ref}}) - M(\mathbf{x}_{\text{ref}})||_1$, which M is the masking function used for binary separation between the object and the background. It compares the mask of the reference image with the mask of the rendering to promote shape consistency.

To impose regularization on the mesh surface, parameterized by SDF representations, we employ SDF regularization methods akin to those proposed by Liao et al. [20] and [19]. Utilizing the binary cross entropy (BCE), the sigmoid function σ, and the sign function, we can express the SDF regularizer $\mathcal{L}_{\text{reg}} = \sum_{(i,j)\in\mathbb{S}} \Big(BCE(\sigma(s_i), \text{sign}(s_j)) + BCE(\sigma(s_j), \text{sign}(s_i)) \Big)$, where s_i is the SDF value at the vertex v_i and $\mathbb{S}$ is set of unique edges. To further encourage the smoothness of the reconstructed surface, we regularize the mean curvature of SDF, which can be computed from discrete mesh Laplacian. The Laplacian loss is defined as $\mathcal{L}_{\text{lap}} = \frac{1}{N} \sum_{i=1}^{N} |\nabla^2 s_i|$. The overall loss can be defined as the combination of $\mathcal{L}_{\text{SDS}}, \mathcal{L}_{\text{photo}}, \mathcal{L}_{\text{mask}}, \mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{lap}}$. We backpropagate the losses to jointly update the 3D shape, PBR texture, and poses of the learnable virtual camera.

3.5 Neural PBR Texture Optimization

As aforementioned in Eq. 1, we employ DMTet in conjunction with a physically based rendering (PBR) material model [27], similar to [30]. This choice allows us to incorporate spatially-varying Bidirectional Reflectance Distribution Function (BRDF) modeling for textures, yielding a more realistic appearance. The PBR material properties, $\mathbf{k}_{\text{PBR}}$ is composed of three key components: diffuse lobe parameters $\mathbf{k}_d \in \mathbb{R}^3$, the roughness and metalness term $\mathbf{k}_{rm} \in \mathbb{R}^2$, and the normal variation term $\mathbf{k}_n \in \mathbb{R}^3$. The specular highlight color, denoted as $\mathbf{k}_s \in \mathbb{R}^3$, can be determined with the renowned Cook-Torrance microfacet BRDF model [9]. Given diffuse value $\mathbf{k}_d$ and the metalness factor m, we compute $\mathbf{k}_s$ as: $\mathbf{k}_s = (1 - m) \cdot 0.04 + m \cdot \mathbf{k}_d$. It enables us to achieve photorealistic surface rendering and enhances the potential of diffusion models for improved realism. More details are in the supplementary material.

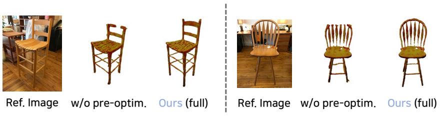


Figure 6: **Necessity of pre-optimization for radius.** The "w/o pre-optim." cases exhibit geometry cut-off for exceeding the camera space boundary and degradation of details.

Metric	Ours (w/o self-calibration)	Ours (full)
Chamfer Distance ↓	0.0593	**0.0580**
F-Score (%) ↑	50.35	**51.15**

Table 1: **Effectiveness of self-calibration for angles.** Ours (full) shows better consistency, depicting the self-calibration effectiveness. We average over all fifteen samples.

4 Experiments

In this section, we first explain the experimental setup in Sec. 4.1. Following that, we show the verification of our system design choices (*e.g.*, virtual camera and test-time adaptation) in Sec. 4.2. We demonstrate our high-fidelity textured mesh reconstruction results in respect of quality and quantity in Sec. 4.3 and Sec. 4.4, respectively.

4.1 Experimental Setup

To evaluate the cross-domain robustness of MeTTA's 3D reconstruction performance, we conduct experiments on the 3D-Front dataset [11], which has not been used in previous single-view to 3D reconstruction methods [31, 50], and we select fifteen samples for evaluation. To demonstrate that our pipeline is working in real-world, out-of-domain scenarios, we manually acquire images from the real scene and the web. For in-domain evaluations, we extract a subset from the Pix3D dataset [43]. Due to time complexity considerations at the optimization, we had to limit the number of dataset selections to a few dozen.

4.2 Verification of System

In this section, we show the experiments to verify the effectiveness of our system design choices, especially for the learnable virtual camera and the test-time adaptation stage.

Effectiveness of learnable virtual camera. We show the ablation studies of camera pre-optimization and self-calibration. The pre-optimization stage is crucial to find the proper radius scale for detailed structures, as shown in Fig. 6. We also present an ablation study of the camera self-calibration in Table 1. We add angle perturbations of [-15, -10, -5, 5, 10, 15] degrees to initial viewpoint estimations. Then, we measure the average scores of the results with respect to the 3D mesh obtained with no perturbation. The self-calibration stage is essential to refine the mapping between a 2D image and 3D space and obtain physically accurate and consistent 3D results, which is vital for total scene reconstruction.

Effectiveness of test-time adaptation. We show the intermediate iteration results during the second stage to present the necessity of the test-time adaptation (TTA) in Fig. 7. While bad initials occur quite often in the Image-to-3D module due to an out-of-distribution gap between training and test, the intermediate results clearly show the strength and necessity of our second TTA stage.

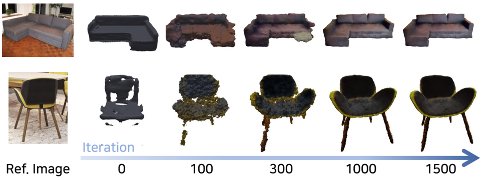


Figure 7: **Intermediate results.** Our method robustly refines meshes and textures iteratively, even with poor initialization.

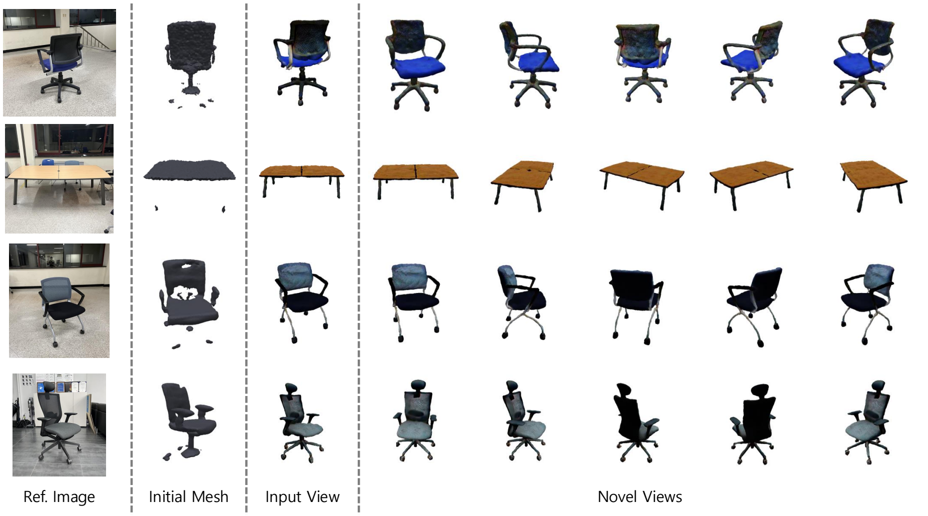


Figure 8: **Unseen real-world experiments about manually acquired data.** We showcase the effectiveness of our test-time adaptation for real scenarios.

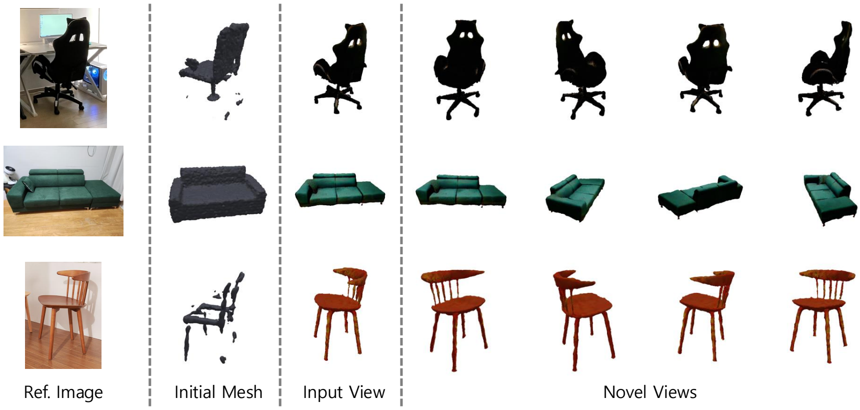


Figure 9: **Unseen real-world experiments about in-the-wild web images.** We showcase the effectiveness of our test-time adaptation for real scenarios.

4.3 Qualitative Analysis

We evaluate and compare the 3D mesh reconstruction quality of `MeTTA` with the competing methods. For more qualitative results, please refer to the supplementary material.

Textured 3D mesh reconstruction. We assess the quality of reconstructed 3D textured meshes in terms of geometric and appearance attributes. In Fig. 8, our results show notable achievement, where we can reconstruct a realistically textured novel-view 3D mesh only from a partial observation of the 3D object in the previously unseen scenarios. In Fig. 9, we conduct another real-world experiment about web images and show fine-grained detailed 3D textured mesh reconstruction results. In Fig. 10, feed-forward methods [31, 50] predict the coarse geometry corresponding to the reference image to some extent. However, for detailed

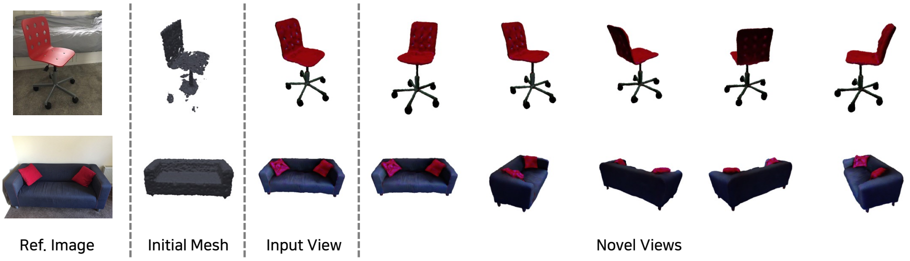


Figure 10: **In-domain experiments.** We showcase the effectiveness of our test-time adaptation of in-domain datasets in which the Image-to-3D module is trained.

geometry and realistic texture, it is essential to apply our test-time adaptation process, even for the in-domain settings.

Comparison with feed-forward methods. We compare ours to previous feed-forward reconstruction methods [31, 50] for visual quality. Thanks to the test-time adaptation with multi-view generative prior, we can get accurate 3D shapes with realistic PBR textures, as shown in Fig. 11.

Comparison with iterative methods using generative priors. We compare our single image to 3D reconstruction results to existing generative priors methods [25, 28, 44]. Because previous methods do not deal with viewpoint information as our learnable virtual cameras, their 3D reconstruction results are not aligned with the reference image and show distorted results, as shown in Fig. 12.

4.4 Quantitative Analysis

We also conduct quantitative comparisons to assess the quality of textured mesh reconstruction and the effectiveness of geometric properties.

Comparison with feed-forward methods. We compare ours to feed-forward reconstruction methods [31, 50] which are also the base models to evaluate whether they have a valid and accurate 3D structure. We evaluate the Chamfer Distance of sampled points between the ground-truth mesh and output mesh of each method. In Table 2, MeTTA outperforms geometry reconstruction than competing methods. Note that our optimization process does not access the ground-truth 3D information, *e.g.*, point clouds, voxels, and meshes, while previous methods are trained to minimize Chamfer Distance with ground-truth 3D shapes as direct supervision. Note that MeTTA also reconstruct fine-grained geometries with utilizing only 2D reference image, compared to others which are trained with 3D shape dataset [43].

Comparison with iterative methods using generative priors. We compare the texture reconstruction quality of MeTTA with the competing methods: RealFusion [28], Zero-1-to-3 [25] and Make-It-3D [44]. In Table 3, we measure the similarity between the reference image and the rendered image at the reference view and novel views, respectively. We use three metrics: PSNR, LPIPS [51], and CLIP score [34]. The CLIP score evaluates the semantic similarity. To see the appearance consistency between novel views, we also report the minimum value of the CLIP score. MeTTA mostly outperforms the competing methods in both reference view and novel view rendering qualities. The results highlight the MeTTA's capability of preserving the semantics of 3D objects, even for the occluded novel views, while achieving high-fidelity 3D reconstruction.

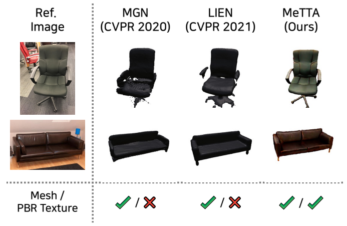


Figure 11: **Comparison with feed-forward methods.**

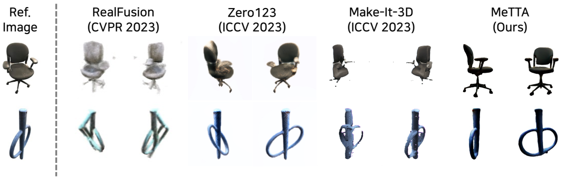


Figure 12: **Comparison with iterative methods using generative priors.** Ours show photo-realistic texture details with physically accurate geometry.

Metric	MGN [31]	LIEN [50]	MeTTA (Ours)
Chamfer Distance ↓	0.1089	0.0975	**0.0943**

Table 2: **Cross-domain evaluation of the single-view to mesh methods.** We evaluate on unseen test dataset [11].

Method	Reference View			Novel Views	
	LPIPS ↓	PSNR [dB] ↑	CLIP Score ↑	CLIP Score ↑	min. CLIP Score ↑
RealFusion [28]	0.1809	21.56	0.8494	0.7538	0.7030
Zero-1-to-3 [25]	0.1079	**23.53**	0.9170	0.7661	0.6670
Make-It-3D [44]	0.0867	22.45	0.9386	0.8937	0.8046
MeTTA (ours)	**0.0777**	22.89	**0.9465**	**0.8942**	**0.8286**

Table 3: **Comparisons of texture reconstruction and perceptual quality.**

5 Discussion, Limitation, and Conclusion

In this work, we present MeTTA, a monocular 3D textured mesh reconstruction with generative test-time adaptation. Our approach addresses several challenges in reconstructing a 3D textured mesh from a single image. First, we highlight the limitations of single-view to 3D mesh prediction methods based on feed-forward manners, which often struggle to ensure high-quality mesh estimation results due to limited 3D shape representation learned from the existing closed training set. Second, we emphasize the necessity of self-calibrating the learnable virtual camera to connect different coordinate spaces between Image-to-3D shape models and the multi-view image generative prior model. Tackling the challenges enables us to achieve quality geometry and photo-realistic texture appearance, complying with input. Finally, We discuss our limitations and conclude with future directions.

Optimization-based system. Ours is much faster than fair competitors, optimization-based approaches [28, 44]. Specifically, our test-time adaptation stage takes 30 minutes per object, compared to 193 minutes of RealFusion [28] and 91 minutes of Make-It-3D [44]. However, we acknowledge that there is still work to achieve practicality, especially in real-time.

Category generalization. Our definition of "cross-domain" implies training and testing on different datasets within the same intra-category, *e.g.*, furniture to furniture. Trained on a small-scale 3D dataset [43], our Image-to-3D module's prediction is category-specific. Despite this, testing in an inter-category scenario in Fig. 13 shows our method is reasonably effective, albeit not designed for such cases.

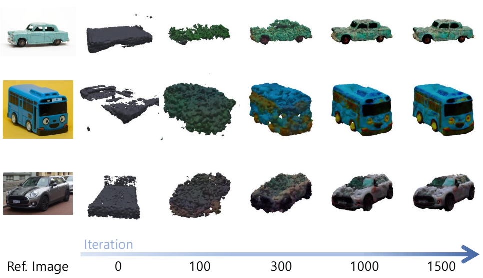


Figure 13: **Possibility of category extension.** Because the Image-to-3D module is trained with 9 indoor object classes [43], it predicts the image as a "bed" rather than a "car".

Future direction. Our two-stage optimization method could be integrated into an end-to-end approach for improved speed and performance. Enhancing the Image-to-3D stage with more data may improve category generalization. We aim to investigate this in future work.

Acknowledgment. This project was supported by Bucketplace and also supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220290, Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense; No.RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities; and No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH)).

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2 (3):8, 2023.

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023.

[4] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. *arXiv preprint arXiv:2311.00457*, 2023.

[5] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *NeurIPS*, 2022.

[6] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023.

[7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, 2022.

[8] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL http://www.blender.org.

[9] R. L. Cook and K. E. Torrance. A reflectance model for computer graphics. *ACM TOG*, 1(1), jan 1982. doi: 10.1145/357290.357293.

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

[11] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *ICCV*, 2021.

[12] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[13] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *ICCV*, 2019.

[14] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *ICCV*, 2023.

[15] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022.

[16] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[18] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM TOG*, 39(6):1–14, 2020.

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

[20] Yiyi Liao, Simon Donne, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2916–2925, 2018.

[21] Joseph J. Lim, Hamed Pirsiavash, and Antonio Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV*, 2013.

[22] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023.

[23] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *ECCV*. Springer, 2022.

[24] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023.

[25] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023.

[26] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[27] Stephen McAuley, Stephen Hill, Naty Hoffman, Yoshiharu Gotanda, Brian Smits, Brent Burley, and Adam Martinez. Practical physically-based shading in film and game production. In *ACM SIGGRAPH 2012 Courses*, pages 1–7. 2012.

[28] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *CVPR*, 2023.

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

[30] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *CVPR*, 2022.

[31] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *CVPR*, 2020.

[32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[33] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*. PMLR, 2021.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[37] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[40] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *NeurIPS*, 2021.

[41] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*, 2023.

[42] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[43] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018.

[44] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023.

[45] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023.

[46] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024.

[47] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. In *NeurIPS*, 2017.

[48] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T. Freeman, and Joshua B. Tenenbaum. Learning Shape Priors for Single-View 3D Completion and Reconstruction. In *ECCV*, 2018.

[49] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *CVPR*, 2024.

[50] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *CVPR*, 2021.

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[52] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Joshua B. Tenenbaum, William T. Freeman, and Jiajun Wu. Learning to Reconstruct Shapes from Unseen Classes. In *NeurIPS*, 2018.