

Real3D: Towards Scaling Large Reconstruction Models with Real Images

Hanwen Jiang Qixing Huang Georgios Pavlakos
 The University of Texas at Austin

Project & Code: <https://hwjiang1510.github.io/Real3D/>

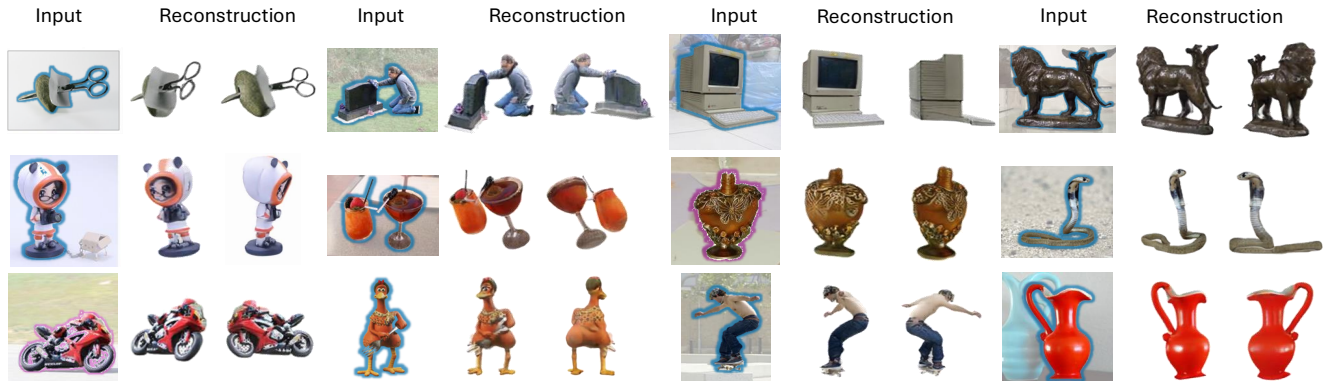


Figure 1. **Real3D reconstructions of in-the-wild instances.** We show the input view and two novel views of our reconstruction, including a novel side view and a novel back view. Real3D can reconstruct diverse shapes from in-the-wild images, including those in uncommon poses and shape categories.

Abstract

*Training single-view Large Reconstruction Models (LRMs) follows the fully supervised route, requiring multi-view supervision. However, the multi-view data typically comes from synthetic 3D assets, which are hard to scale further and are not representative of the distribution of real-world object shapes. To address these limitations, we introduce Real3D, the first LRM that uses **single-view real images** for training, benefiting from their scalability and capturing the real-world shape distribution. Real3D introduces a novel self-training framework, including unsupervised losses at the pixel- and semantic-level, enabling LRMs to learn from these single-view images without multi-view supervision. Simultaneously, to deal with the noise of real data, Real3D also presents an automatic data curation approach to gather high-quality examples that have positive impact on training. Our experiments show that Real3D consistently outperforms prior work in diverse evaluation settings that include real and synthetic data, as well as both in-domain and out-of-domain shapes.*

1. Introduction

The scaling law is the secret sauce of large foundation models [33]. By scaling both model parameters and training data, the foundation models demonstrate impressive emerging ca-

pabilities [2, 60]. Recently, the same recipe has been applied to build large reconstruction models (LRMs) [23], *i.e.*, foundation models for *single-view 3D reconstruction*. LRMs benefit from training on *multi-view images* from 3D/video data, where the increased dataset size [9, 10, 98] is the key of the improved model performance. However, the excessive reliance on *multi-view supervision* creates *two bottlenecks* for LRMs. First, expanding the scale of such data is hard – either creating synthetic 3D assets [9] or intentional video captures [98] is laborious. This leads to a *limited scale* of training data of LRMs, compared with LLMs [2, 76] and VLMs [36, 101], ceiling their scalability. Second, the multi-view data is *biased* towards shapes that are easy to model by artists or easy to capture in the round table setting [29, 62, 98], creating a domain gap with the real-life shapes we want to reconstruct in most applications.

To address these limitations, we propose **Real3D**. Our main objective is to expand the training data of LRMs to **single-view images**. This is motivated by the potential of using single-view images to solve the two aforementioned problems. First, single-view images are readily available in *existing large-scale datasets* [12, 66, 74] or *in-the-wild resources*, leading to a large-scale set of training examples. Second, by leveraging a large number of collected **real images** for training, we can capture the distribution of real objects more faithfully (Figure 1), closing the training-inference gap and

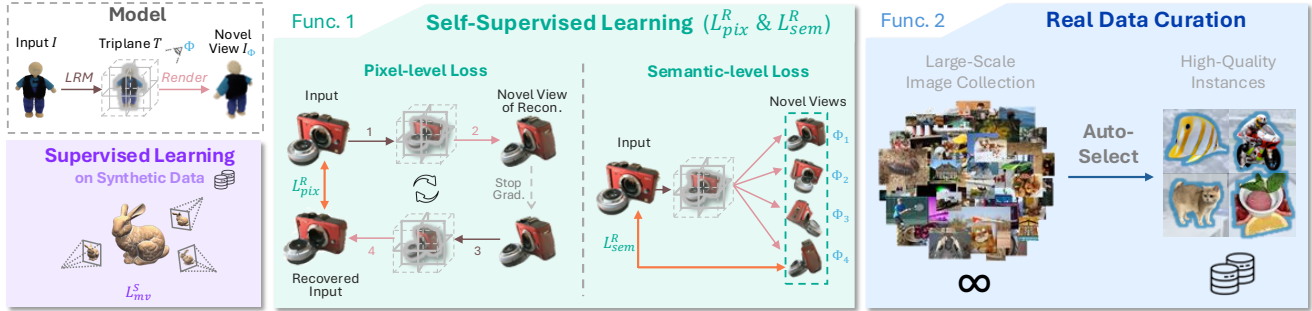


Figure 2. **Overview of Real3D.** (Left) Real3D adopts a standard LRM model architecture with Triplane 3D representation \mathbf{T} that can render a novel view I_Φ given a target camera pose/viewpoint Φ (Sec. 3). The LRM is first initialized with medium-scale multi-view synthetic data using supervised losses \mathcal{L}_{mv}^S . (Middle) To enable scaling, Real3D extends the training data to **single-view real images**, benefiting from their scalability and faithful capture of the real-world shape distribution. Real3D proposes a novel self-training method for improving LRM on this single-view data, including pixel-level, \mathcal{L}_{pix}^R , and semantic-level, \mathcal{L}_{sem}^R , self-supervision (Sec. 4.1). For the pixel-level supervision we leverage a cycle consistency rendering loss, while for the semantic-level self-supervision, we maximize the semantic similarity between rendered novel views of the reconstruction and the input image. (Right) To further benefit from self-training, Real3D automatically selects the high-quality instances from large-scale noisy images (Sec. 4.2), benefiting model accuracy and generalization capability.

improving generalization [38, 44, 94].

Given this abundance of single-view images, the key challenge is how to incorporate them in the training of LRMs. To achieve this, Real3D proposes a novel **self-training framework**. More specifically, we introduce **unsupervised training losses**, *i.e.*, both *pixel-level* and *semantic-level* self-supervision, mimicking how LRMs are trained with *multi-view supervision*. The pixel-level self-supervision leverages a cycle consistency rendering loss, enabling training with photometric losses on merely the single-view input image. The semantic-level self-supervision maximizes the semantic similarity between rendered novel views of the reconstruction and the input image (Figure 2, middle).

Although the large-scale image collections are beneficial, they tend to be highly noisy, leading to limited improvements for our system. To address this, Real3D introduces an automated **data curation** method for selecting high-quality examples, *i.e.*, unoccluded instances. Eventually, we train jointly on the curated in-the-wild data using self-training and on the synthetic data using supervised losses. This strategy enriches the model with knowledge from real single-view data while preventing divergence.

We evaluate Real3D on a diverse set of datasets, spanning both real and synthetic data and covering both in-domain and out-of-domain shapes. Experimental results highlight Real3D’s three key strengths: i) **Superior performance**. Real3D outperforms prior works in all evaluations, with an average relative improvement of 4.2% PSNR (15% MSE), 6.3% LPIPS, 13.5% FID, and 14.3% mesh accuracy. ii) **Effective use of real data**. Real3D demonstrates greater improvement using *single view real-data* than what the previous methods achieved using *multi-view real data*. iii) **Scalability**. Performance improves as more data is incorporated, demonstrating Real3D’s potential of scaling. We will make our code, models, and data available upon publication.

2. Related Work

Single-view 3D Reconstruction. Reconstructing 3D scenes and objects from a single image is a core task in 3D Computer Vision. One line of work focuses on developing better 3D representations to improve reconstruction quality. For example, different explicit representations, *e.g.*, voxels [8, 78], point clouds [14, 30], multiplane images [51, 77], meshes [16, 43], and 3D Gaussians [35, 71, 73], as well as implicit representations, *e.g.*, SDFs [49, 56, 69], and radiance fields [27, 52, 63, 97], have been explored. Recent methods explore incorporating various guidance to improve single-view 3D reconstruction. MCC [85] and ZeroShape [25] use depth guidance; however, accurate depth inputs are not always available. RealFusion [48], Makeit-3D [72] and Magic123 [58] harness diffusion priors as guidance [57]. However, they require slow per-shape optimization. Zero-1-to-3 [42] fine-tunes diffusion models for direct novel view generation. With multiple views, it is easier to get a 3D reconstruction. However, this route suffers from limited reconstruction quality, caused by inconsistency between the generated novel views [45, 68, 81]. Adversarial guidance is explored to distinguish reconstruction from input view [24, 90]. However, adversarial training is generally unstable and difficult to scale. Moreover, semantic guidance leverages the image-to-text inversion model and the similarity calculation model to supervise novel reconstruction views [11, 72]. This improves the semantic consistency of the reconstructions but harms the reconstruction details [15, 60]. In contrast, our method, Real3D, proposes two complementary unsupervised losses to improve both the semantic and spatial consistency of our reconstruction, enabling us to train using single-view images.

Large Reconstruction Models. Large reconstruction models are proposed for generalizable and fast feed-forward

3D reconstruction. They use scalable model architectures, e.g., Transformers [23, 26, 80] or Convolutional U-Nets [64, 73, 83] to encode diverse shape and texture priors and directly map 2D information to 3D representations. The models are trained with multi-view rendering losses, assuming access to 3D ground-truth. For example, LRM [23] uses triplane tokens to query information from 2D image features. Other methods improve reconstruction quality by exploiting better representations [84, 99, 102] and introducing generative priors to first generate multi-view images [18, 46, 73, 83, 91, 92]. However, one shortcoming of these models is that they require a normalized coordinate system and canonicalized input camera pose, which limits the scalability and effectiveness of training with multi-view real data. To solve this problem, we enable the model to perform self-training on real single images, without overly relying on multi-view supervision.

Unsupervised 3D Learning from Real Images. Learning to perform 3D reconstruction typically requires 3D ground-truth or multi-view supervision, which makes scaling up more challenging. To solve this problem, a promising avenue would be learning from massive unannotated data. Early works in this direction leverage category-level priors, where the reconstruction can benefit from category-specific templates and the definition of a canonical coordinate frame [3, 13, 21, 32, 34, 37, 39, 41, 53–55, 86]. Recent works extend this paradigm to general categories by adjusting adversarial losses [4, 40, 50, 96], multi-category distillation [1], synergy between multiple generative models [88], knowledge distillation [70], depth regularization [65], and multi-view compression [28]. However, these methods learn 3D reconstruction from scratch, without leveraging the available 3D annotations due to limitations of their learning frameworks. Thus, their 3D accuracy and viewpoint range are limited. In contrast, our model enables initialization with available 3D ground-truth from synthetic examples and is jointly trained with in-the-wild images using unsupervised losses, which improves the reconstruction quality.

Model Self-Training. Self-training helps improve performance when labeled data is limited or not available [61, 67, 89]. For example, contrastive learning methods use different image augmentations to learn visual representations [5, 6, 19]. This strategy has also been applied to 3D computer vision tasks, e.g., hand pose estimation [44], detection [93], segmentation [31] and depth estimation [94, 95]. In this paper, we propose novel losses to perform self-training on real images to improve 3D reconstruction.

3. Preliminaries

Large Reconstruction Model (LRM). Let $I \in \mathbb{R}^{H \times W \times 3}$ be an input image of the target object. The LRM outputs a 3D representation of the object, $\mathbf{T} = \text{LRM}(I)$, where $\mathbf{T} \in \mathbb{R}^{3 \times h \times w \times c}$ is the latent triplane. LRM performs vol-

ume rendering [4] to produce novel views. The rendering module is formulated as $\hat{I}_\Phi = \pi(\mathbf{T}, \Phi)$, where \hat{I}_Φ is the rendered image under a target camera pose $\Phi \in \text{SE}(3)$, and π represents the rendering process.

Training LRM on Synthetic Multi-view Images. The standard training of LRMs requires multi-view image supervision. For each training object, we have access to several views of it with the corresponding camera poses. We denote the views and poses as $\{(I_i^{gt}, \Phi_i) | i = 1, \dots, n\}$, where n images are collected. These multi-view images are usually obtained from synthetic 3D assets [9] using rendering tools. The loss function on these multiple views is:

$$\mathcal{L}_{\text{mv}}(I) = \frac{1}{n} \sum_{i=1}^n (\mathcal{L}_{\text{MSE}}(\hat{I}_{\Phi_i}, I_i^{gt}) + \lambda \cdot \mathcal{L}_{\text{LPIPS}}(\hat{I}_{\Phi_i}, I_i^{gt})), \quad (1)$$

where λ balances the two losses, \mathcal{L}_{MSE} and $\mathcal{L}_{\text{LPIPS}}$ [100] that provide pixel-level and semantic-level supervision, respectively.

Coordinate Frame System. LRM has a constant $\text{SE}(3)$ camera pose, ϕ , that renders the reconstruction triplane \mathbf{T} at the viewpoint of input images. This camera pose ϕ has an identity rotation and a fixed translation [23]. Since this constant camera pose is shared across any input image, we denote it as the canonical pose. This canonical pose is used in Sec. 4.1 for designing our self-training techniques.

4. Real3D

We propose a novel framework that enables training LRMs using **unlabeled real-world single-view images**, which are easier to collect/scale and can better capture the distribution of real object shapes. As shown in Fig. 2, we initialize an LRM on a synthetic dataset. Then we collect real-world object instances. We jointly train the model on synthetic multi-view data (Eq. 1) and perform self-training on real single-view data. The former prevents the model from diverging with the help of supervision from ground-truth novel views. The latter introduces new data in the model training, which improves the reconstruction quality and generalization capability (Sec. 4.1). We also propose an automatic data curation method to collect high-quality instances from real images to benefit self-training (Sec. 4.2).

4.1. Self-Training on Real Images

The effectiveness of the multi-view training loss (Eq. 1) originates from applying supervision for improving both *pixel-level* and *semantic-level* similarity between reconstruction and ground-truth novel views. Following this philosophy, we develop novel **unsupervised** pixel-level and semantic-level supervision when training the model on single-view images, where ground-truth multiple views are not available. For our base model, we use a fine-tuned TripoSR [75], an LRM without input pose and intrinsics conditioning [23].

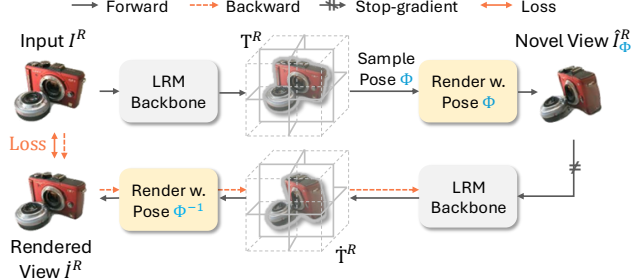


Figure 3. **Pixel-level supervision with cycle-consistency loss.** We use LRM to reconstruct the object (T^R) and render a novel view \hat{I}_Φ^R at viewpoint/pose Φ . We again apply LRM on this rendered novel view to reconstruct the object (\tilde{T}^R) and render it from the viewpoint of the original input, denoted as \tilde{I}^R . The difference between I^R and \tilde{I}^R provides the loss for training. We apply **stop gradient** on intermediate result \hat{I}_Φ^R to avoid model degradation.

4.1.1. Pixel-level Supervision

The pixel-level supervision has three components: 1) the **cycle consistency loss** for providing pixel-level supervision, 2) the **stop gradient regularizer**, which avoids minimizing the loss by learning trivial solutions, and 3) the **camera pose sampling curriculum** to improve stability.

Cycle-consistency Loss. The loss is illustrated in Fig. 3. It provides pixel-level supervision using the original input image. We formulate it as follows.

We input the model with an image I^R that contains a real-world shape instance and then reconstruct the triplane:

$$\mathbf{T}^R = \text{LRM}(I^R), \quad I^R \in \mathbb{R}^{H \times W \times 3}. \quad (2)$$

We *sample* a camera pose Φ to render a novel view of the reconstruction and create the second input \hat{I}_Φ^R in the cycle:

$$\hat{I}_\Phi^R = \pi(\mathbf{T}^R, \Phi), \quad \Phi = \phi \cdot \Delta\Phi, \quad (3)$$

where the sampled camera pose Φ is associated with the relative pose $\Delta\Phi$ between the sampled target view and the input view. Here, ϕ is the constant canonical pose of the input view (introduced in Sec. 3). We input this synthesized novel view back to the LRM for reconstruction, and render a image at the viewpoint of the original input I^R :

$$\tilde{\mathbf{T}}^R = \text{LRM}(\hat{I}_\Phi^R), \quad \text{and} \quad (4)$$

$$\tilde{I}^R = \pi(\tilde{\mathbf{T}}^R, \hat{\Phi}), \quad \text{where, } \hat{\Phi} = \phi \cdot (\Delta\Phi)^{-1}. \quad (5)$$

Here \tilde{I}^R is the image that closes the cycle and $\hat{\Phi}$ is the camera pose to render the image at the viewpoint of original input I^R . The pixel-level cycle consistency loss is defined as: $\mathcal{L}_{\text{pix}}^R = \mathcal{L}_{\text{MSE}}(\tilde{I}^R, I^R) + \lambda \cdot \mathcal{L}_{\text{LPIPS}}(\tilde{I}^R, I^R)$.

Stop-gradient Regularizer. We observe that simply training with the cycle consistency loss leads to degraded performance. In detail, $\mathcal{L}_{\text{pix}}^R$, which is defined on 2D images,

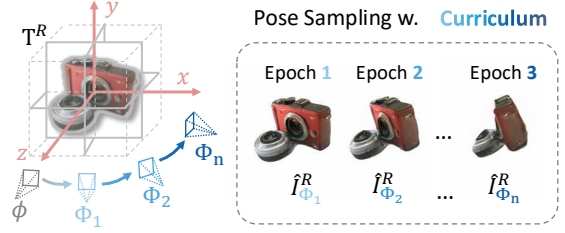


Figure 4. **Pose sampling curriculum.** By guiding the sampling process of pose Φ , we control the training difficulty. At the beginning of training, we ensure a small variation between sampled pose Φ and canonical pose ϕ , achieving steady training. We gradually enlarge the upper bound of this variation ($\Phi_n > \dots > \Phi_2 > \Phi_1$) during training to increase difficulty and improve learning.

can be minimized by adopting trivial solutions that do not improve 3D geometry quality. Please see Appendix for visualization. To solve the problem, we apply a stop-gradient regularization on the intermediate rendering result \hat{I}_Φ^R , as shown in the right side of Fig. 3. The stop-gradient operation is widely used in self-supervised learning to improve training effectiveness by avoiding trivial solutions [6, 17, 79].

This stop-gradient operator divides the cycle into two parts. The second half of the cycle (from \hat{I}_Φ^R to \tilde{I}^R with $\mathcal{L}_{\text{pix}}^R$) is the normal method for training LRM with novel view rendering losses. The first half of the cycle (from I^R to \hat{I}_Φ^R) can be interpreted as the process of collecting the input image \hat{I}_Φ^R of the second half.

Learning with Curriculum. When using the stop-gradient regularizer, the training performance is dependent on the quality of the intermediate rendered view \hat{I}_Φ^R . However, since the model is imperfect, the created novel view \hat{I}_Φ^R can be inaccurate, limiting the effectiveness of self-training.

To address this problem, we introduce a **curriculum learning** approach. This curriculum is based on our observation that the rendered novel view tends to be more accurate when the camera pose has a smaller variation with the input image. Thus, we can control the learning difficulty and uncertainty by varying the pose change between the sampled pose Φ , which is used for rendering, and the canonical pose ϕ . As shown in Fig. 4, initially, the model learns on simpler cases with a small pose variation, where the rendered view \hat{I}_Φ^R tends to be more accurate. We formulate the camera sampling method under the curriculum as:

$$\Phi = \phi \cdot \Delta\Phi, \quad \Delta\Phi \sim \text{Uniform}(-\Delta\Phi_{\text{max}}^j, \Delta\Phi_{\text{max}}^j), \quad (6)$$

where we denote *Uniform* as the uniform sampling in the $\text{SE}(3)$ pose space, achieved by parametrizing poses with the polar coordinate system. $\Delta\Phi_{\text{max}}^j$ is the maximum sampling range of relative camera pose *at the training iteration* j . Note $\Delta\Phi_{\text{max}}^j$ is parameterized by the relative azimuth θ_{max}^j

and elevation φ_{\max}^j , which are

$$\theta_{\max}^j = j/j_{\max} \cdot (\theta_{\max} - \theta_{\min}) + \theta_{\min}, \text{ and} \quad (7)$$

$$\varphi_{\max}^j = j/j_{\max} \cdot (\varphi_{\max} - \varphi_{\min}) + \varphi_{\min}, \quad (8)$$

where j_{\max} is the number of total training iterations. We start the curriculum with $\theta_{\min} = \varphi_{\min} = 15^\circ$, and finalize with $\theta_{\max} = \varphi_{\max} = 90^\circ$. The camera pose sampling range keeps increasing in the curriculum during the training process.

4.1.2. Semantic-level Supervision

The semantic supervision is performed between the novel view of the reconstruction and the input view. We leverage CLIP to compute the semantic similarity loss, as CLIP is trained on image-text pairs and can capture high-level image semantics. The image semantic similarity loss is $\mathcal{L}_{\text{CLIP}}(\hat{I}, I) = -\langle f(\hat{I}), f(I) \rangle$, where f represents the CLIP visual encoder for predicting normalized image features.

A simple strategy for applying the loss is rendering multiple novel views of a reconstruction and calculating the loss on all of them. However, this leads to the multi-head problem, a trivial solution to minimize the loss (see visualization in Appendix). The reason is that CLIP is not fully invariant to the camera viewpoint.

To address this issue, we apply the loss to a single rendered novel view, specifically the one least similar to the input view among multiple candidates, leveraging hard negative mining. Besides, we avoid rendering novel views that are too far from the input viewpoint, as the back of the shape may exhibit semantic differences from the input view.

We formulate the semantic-level supervision as follows. For the latent triplane \mathbf{T}^R of real image I^R , we render m novel views using m sampled rendering camera poses as:

$$\hat{I}_{\Phi_i}^R = \pi(\mathbf{T}^R, \Phi_i), \text{ for } i \in \{1, \dots, m\}. \quad (9)$$

We sample the camera poses using a similar strategy in Eq. 6:

$$\Phi_i = \phi \cdot \Delta\Phi_i, \quad \Delta\Phi_i \sim \text{Uniform}(-\Delta\Phi_{\max}, \Delta\Phi_{\max}), \quad (10)$$

where $\Delta\Phi_{\max}$ is parameterized by $\theta'_{\max} = 120^\circ$ and $\varphi'_{\max} = 45^\circ$. Unlike the sampling of pixel-level supervision (Eq. 6), this sampling does not depend on the training iteration j .

Finally, we calculate the semantic-level supervision as:

$$\mathcal{L}_{\text{sem}}^R = \mathcal{L}_{\text{CLIP}}(\hat{I}_{\Phi_k}^R, I^R), \text{ where} \quad (11)$$

$$k = \arg \max_{i \in \{1, \dots, m\}} \mathcal{L}_{\text{CLIP}}(\hat{I}_{\Phi_i}^R, I^R). \quad (12)$$

Training Target. The losses applied in self-training on real data can be defined as:

$$\mathcal{L}_{\text{self}}^R = \lambda_{\text{in}}^R \cdot \mathcal{L}_{\text{in}}^R + \lambda_{\text{pix}}^R \cdot \mathcal{L}_{\text{pix}}^R + \lambda_{\text{sem}}^R \cdot \mathcal{L}_{\text{sem}}^R, \quad (13)$$

where $\mathcal{L}_{\text{in}}^R$ is the rendering loss on the input view, and \hat{I}_{ϕ}^R is rendered using the canonical pose of input as $\hat{I}_{\phi}^R = \pi(\mathbf{T}^R, \phi)$. The final losses on both synthetic and real-world data can be defined as $\mathcal{L} = \mathcal{L}_{\text{mv}}^S + \mathcal{L}_{\text{self}}^R$.

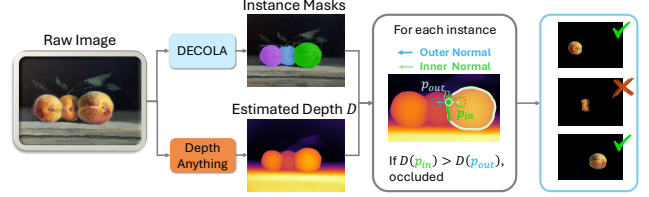


Figure 5. **Pipeline for automatic selection of clean instances.** We use an instance segmentation model (DECOLA [7]) to detect masks of object instances and a monocular depth model (DepthAnything [94]) to estimate depth maps D . For each instance, we sample points on its boundary and select pixels along the direction of the inner and outer boundary normal (p_{in} and p_{out} respectively). If $D(p_{in}) < D(p_{out})$, we infer that the instance “owns” this boundary. If this inequality holds for all boundary points, we conclude that the object is unoccluded and is considered a clean instance.

4.2. Automatic Data Curation

Our data curation aims to identify unoccluded instances from real images. This is important because occluded instances have incomplete side-view information and lead to limited model improvement (verified in Table 8). We use models for open-vocabulary instance segmentation (DECOLA [7]) and monocular depth estimation (Depth Anything [94]) to predict instance masks and depth maps (Fig. 5). From the instance mask, we detect contacting boundaries between different object instances. We sample points on the boundary, and query the predicted depth map along the direction of the inner and outer normal. We infer occlusion if the depth of the inner point exceeds the depth of the outer point. We also incorporate robust estimation methods by dense point sampling and voting. Please see Appendix for more details.

5. Experiments

We introduce our evaluation results on diverse datasets, including real images in controlled and in-the-wild settings, for comparison with previous work.

Implementation Details. On the synthetic data, we supervise the model on $n = 4$ renderings. On the real data, we render $m = 4$ novel views to apply the semantic supervision. We use 1 view to apply the pixel-level supervision. All evaluations are performed with a rendering resolution of 224. We set the learning rate as $4e - 5$ using the AdamW optimizer [47]. We train the model with $j = 40,000$ iterations with a cosine learning rate scheduler. We use a batch size of 80, where we have half synthetic samples and half real-data samples. We set $\lambda, \lambda^R, \lambda_{\text{pix}}^R, \lambda_{\text{sem}}^R$ as 1.0, 0.3, 5.0 and 1.0. Please see more details in the Appendix.

Datasets. We train Real3D on our collected WildImages real single-view data and the synthetic multi-view renderings of Objaverse [9] jointly. We evaluate the models on test splits of WildImages, MVImageNet, CO3D, and OmniObject3D.

• **WildImages** contains 300K single-view segmented

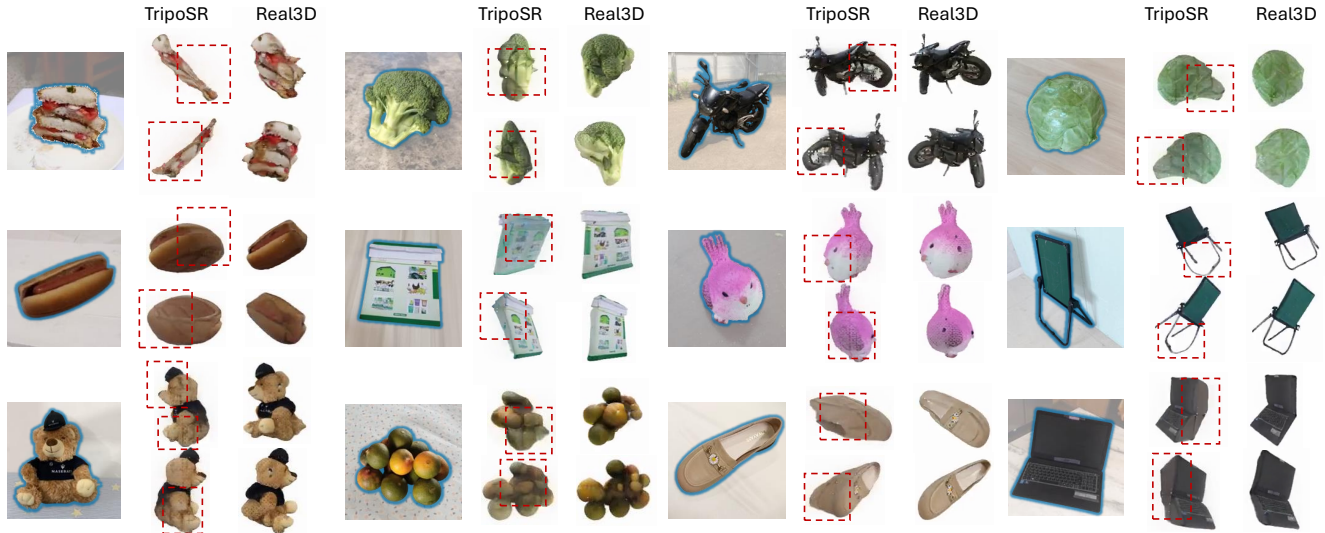


Figure 6. **Comparison with TripoSR.** We compare Real3D with the base model TripoSR. Real3D uses real images for training and performs better on real images. Red boxes highlight failure cases of TripoSR due to the limited data coverage of its synthetic training data.

objects, collected from datasets in diverse domains [12, 38, 98], which is filtered from more than 3M real object instances. We keep aside a test split with 1000 images.

- **Objaverse** [9] is a large-scale synthetic dataset. We use a filtered subset consisting of 260K high-quality shapes for training. We use renderings from [42, 59]. We note this is a widely used subset [59, 73], as Objaverse contains many low-quality instances, which harm learning.

- **MVImgNet** [98], **CO3D** [62] and **OmniObject3D** [87] are used for evaluation, containing real data and out-of-domain synthetic data.

Metrics. We use two sets of metrics to evaluate the models on data with and without multi-view annotations:

- **Novel View Synthesis (NVS) Metrics.** Following prior work, we evaluate reconstructions on data with multi-view information, using standard metrics, including PSNR, SSIM [82], and LPIPS [100].

- **Semantic and Self-Consistency Metrics.** To evaluate reconstruction quality on single-view data without ground-truth multi-view images, we introduce novel metrics. First, we render novel views of a reconstruction and measure the semantic similarity with the input view. In detail, we render 7 views where the azimuths are uniformly sampled in range $[0, 360]$ and no elevations. We use semantic metrics, including LPIPS [100], CLIP similarity [60], and FID score [22]. Second, we evaluate the self-consistency of the reconstruction. We render a designated novel view of the reconstruction, use it as input for the LRM, and render the second reconstruction from the viewpoint of the original input. We evaluate the consistency using NVS metrics.

- **Mesh Quality Metrics.** We report the mesh quality using the Chamfer-L1 Distance (CD) with mesh scale 2.0. To get mesh from the triplane representation, we use marching cubes with resolution 256 for TripoSR and Real3D.

Baselines. We compare Real3D with OpenLRM [20], TripoSR [75], LGM [73], CRM [83] and InstantMesh [91]. We use OpenLRM [20], an open-sourced LRM for comparisons, as LRM is close-sourced. All models are trained on Objaverse [9] unless noted otherwise. Specifically, InstantMesh uses the larger synthetic dataset Objaverse-XL [10]. Moreover, we finetune TripoSR, as it predicts reconstruction with random scales on different inputs with non-clean backgrounds, which leads to inferior evaluation results (see Appendix) and failure of self-training. TripoSR is a base model directly comparable to ours¹. For all baselines, we use the official checkpoints. Furthermore, we follow the specific settings of each model to normalize target camera poses.

5.1. Experimental Results

Qualitative Results. We show Real3D reconstruction examples in Fig. 6, showing that Real3D can recover the geometry with high fidelity. Compared with the baseline method TripoSR, Real3D demonstrates less reconstruction error on objects from *uncommon categories* or under *uncommon camera poses* that lead to ambiguity. Specifically, Real3D shows better accuracy for reconstructing *side-view* and *back* of objects. This performance gain shows that Real3D leverages the *semantic knowledge* learned from large-scale real data to resolve shape ambiguity, verifying the effectiveness of our self-training and data curation methods. Please also see Appendix for comparisons with baselines on evaluation sets.

Quantitative Results. Real3D consistently outperforms prior works on all four test sets of our evaluation. As shown in Table 1 to Table 4, Real3D showcases a 0.74 dB PSNR

¹We highlight that our TripoSR is also fine-tuned with the synthetic dataset we use. Thus, the performance gain comes from self-training. We provide results of TripoSR official weights in Appendix.

Method	Eval. on Input View						Eval. on GT Novel Views		
	Semantic Similarity			Self-Consistency			Novel View Synthesis Quality		
	CLIP \uparrow	LPIPS \downarrow	FID \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
METHODS W. GENERATIVE PRIORS									
LGM [73]	0.820	0.204	158.4	14.20	0.833	0.227	15.95	0.813	0.181
CRM [83]	0.823	0.179	172.5	15.72	0.873	0.168	17.54	0.853	0.142
InstantMesh [91]	0.873	0.188	153.5	14.81	0.861	0.171	14.70	0.806	0.197
METHODS W/O GENERATIVE PRIORS (DETERMINISTIC)									
OpenLRM [20]	0.868	0.160	147.6	17.69	0.873	0.140	19.75	0.864	0.112
TripoSR [75]	0.860	0.157	129.7	17.72	0.874	0.143	19.81	0.864	0.116
Real3D (ours)	0.892	0.147	116.9	19.80	0.893	0.125	20.53	0.871	0.107

Table 1. Evaluation results on the real-world MVImageNet dataset. We evaluate the performance both on input view and novel views. We bold and highlight the best results.

Method	Eval. on GT Novel Views		
	Novel View Synthesis Quality		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
METHODS W. GENERATIVE PRIORS			
LGM [73]	15.14	0.802	0.187
CRM [83]	16.38	0.840	0.153
InstantMesh [91]	13.99	0.789	0.199
METHODS W/O GENERATIVE PRIORS			
OpenLRM [20]	18.31	0.849	0.126
TripoSR [75]	18.44	0.848	0.127
Real3D (ours)	19.18	0.855	0.119

Table 2. Evaluation of novel view synthesis quality on real-world CO3D data.

Method	Eval. on Input View					
	Semantic Similarity			Self-Consistency		
	CLIP \uparrow	LPIPS \downarrow	FID \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
METHODS W. GENERATIVE PRIORS						
LGM [73]	0.843	0.188	146.0	14.13	0.825	0.210
CRM [83]	0.807	0.174	162.4	15.67	0.862	0.160
InstantMesh [91]	0.841	0.182	152.5	14.68	0.854	0.163
METHODS W/O GENERATIVE PRIORS						
OpenLRM [20]	0.847	0.149	144.9	18.09	0.872	0.129
TripoSR [75]	0.877	0.148	128.5	18.18	0.874	0.125
Real3D (ours)	0.892	0.144	106.5	19.00	0.882	0.117

Table 3. Evaluation results on WildImages test set. Due to the absence of ground-truth novel views, we perform evaluation on the input view.

Method	Eval. on GT Novel Views		
	Novel View Synthesis Quality		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
METHODS W. GENERATIVE PRIORS			
LGM [73]	15.83	0.791	0.197
CRM [83]	16.75	0.823	0.182
InstantMesh [91]	15.83	0.791	0.197
METHODS W/O GENERATIVE PRIORS			
OpenLRM [20]	18.20	0.831	0.144
TripoSR [75]	19.43	0.847	0.128
Real3D (ours)	20.17	0.855	0.119

Table 4. Evaluation of novel view synthesis quality on synthetic out-of-domain OmniObject3D data.

(4% relatively, equivalent to 16% better MSE) improvement, a 6.3% relative LPIPS improvement, and a 13.5% relative FID improvement *on average* over the directly comparable TripoSR model, demonstrating the effectiveness of our self-training method using real data. Results in Table 3 also highlight the advantage of our self-training method by using a broader data distribution to handle *in-the-wild* test data.

Additionally, we observe that methods with generative priors do not perform well on out-of-distribution data. These methods generate novel views and use those views to perform sparse-view reconstruction. We conjecture that the reason is the compounding error of the novel view synthesis and reconstruction stages. This is another argument in favor of single-stage methods, like ours.

Mesh Quality. We report that the CD for InstantMesh (the best baseline with generative prior) and TripoSR (the best deterministic baseline) are 0.395 and 0.321 respectively. In contrast to that, our method Real3D achieves a CD of 0.275, an improvement by 14.3%. We include the visualization of the mesh reconstructions in Appendix. We observe that both baselines perform worse than Real3D, particularly in cases with non-common object shapes, while InstantMesh specifically struggles to faithfully reconstruct thin structures.

5.2. Ablation Studies

With our ablation studies, we analyze the proposed losses and examine key design choices for semantic- and pixel-level supervision. Additionally, we evaluate the effectiveness of

our data curation procedure.

Ablation of proposed losses. As shown in Table 5, when we use only the \mathcal{L}_{in}^R loss, we observe slight improvements of PSNR, but SSIM and LPIPS have limited improvements. This pattern implies that \mathcal{L}_{in}^R can only help the model render more realistic pixels by learning the real-image pixel distribution, but it can not improve 3D reconstruction quality. Further adding the proposed semantic- and pixel-level unsupervised losses helps improve the performance.

Semantic-level Supervision Designs. As shown in Table 6, naively applying the CLIP-based semantic loss on all rendered novel views degrades the performance. We present visualization results in Appendix, observing the multi-head problem of reconstructions. We conjecture that the reason is that copying the input view to all other views is a trivial solution to minimize the loss. This motivates us to incorporate regularization as we do with our semantic supervision, which improve the performance across all metrics.

Pixel-level Supervision Designs. As shown in Table 7, the pixel-level supervision is only useful when stopping the gradient of intermediate rendering and applying a training curriculum. The result demonstrates the importance of each proposed component. We present more visualization for the ablation in Appendix.

Effectiveness of Data Curation. As shown in Table 8, simply training with all data leads to smaller performance gains, compared to using clean data filtered from real images

\mathcal{L}_{in}^R	\mathcal{L}_{sem}^R	\mathcal{L}_{pix}^R	Eval. on GT Novel Views		
			Novel View Synthesis Quality		
			PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
✗	✗	✗	18.44	0.848	0.127
✓	✗	✗	18.60	0.850	0.127
✓	✗	✗	18.81	0.853	0.125
✓	✓	✓	19.18	0.855	0.119

Table 5. Ablation study of proposed losses, including the input view rendering loss (\mathcal{L}_{in}^R), semantic-level supervision (\mathcal{L}_{sem}^R) and pixel-level supervision (\mathcal{L}_{pix}^R).

	Eval. on GT Novel Views		
	Novel View Synthesis Quality		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
No sem. loss	18.60	0.850	0.127
Naive sem. loss	17.89	0.830	0.151
\mathcal{L}_{sem}^R	18.81	0.853	0.125

Table 6. Ablation study of semantic-level supervision. ‘‘Naive’’ means simply applying CLIP-semantic loss on all novel views. We do not apply the pixel-level supervision (\mathcal{L}_{pix}^R) for these experiments.

Stop Grad.	Curriculum	Eval. on GT Novel Views		
		Novel View Synthesis Quality		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
✗	✓	17.78	0.821	0.140
✓	✗	18.63	0.848	0.125
✓	✓	19.18	0.855	0.119

Table 7. Ablation study of pixel-level supervision with cycle-consistency. ‘‘Stop Grad.’’ means stopping the gradient on intermediate rendering results. \mathcal{L}_{sem}^R is incorporated for these experiments.

	Eval. on GT Novel Views		
	Novel View Synthesis Quality		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
All data	18.79	0.852	0.123
Clean data	19.18	0.855	0.119

Table 8. Ablation study of data curation for getting clean data. All losses are incorporated for these experiments

Method	Eval. on GT Novel View									Eval. on Input View		
	MVIImgNet			CO3D			OmniObject3D			WildImages		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Δ multi-view	0.410	0.003	0.007	0.510	0.003	0.007	0.330	0.006	0.006	0.260	0.000	0.000
Δ ours	0.720	0.007	0.009	0.740	0.007	0.008	0.740	0.008	0.009	0.720	0.008	0.008

Table 9. Performance gain comparison between: i) using multi-view real data with multi-view rendering losses (Δ multi-view) and ii) using our real images with self-training losses (Δ ours). Δ multi-view uses 260K MVIImgNet videos with about 8 million images. In contrast, our method uses 300K high-quality images selected from about 3 million real images for self-training.

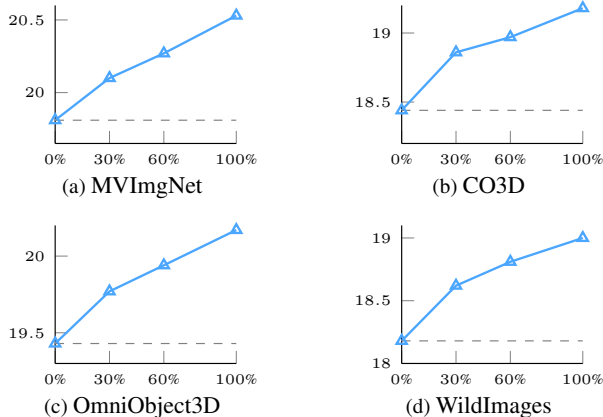


Figure 7. Real3D performance (PSNR) using different amounts of real data for training. The PSNR is evaluated on novel views for (a)-(c), and it is evaluated on (d) with self-consistency.

with our method. This highlights both the importance of data curation and the effectiveness of our method.

5.3. Scalability and Effectiveness of Self-Training

We provide experimental results to further understand the scalability and effectiveness of our self-training method.

Scalability - Data Amount Analysis. Here, we evaluate the effect of scaling up the training data. As shown in Fig. 7, Real3D achieves consistent improvements when more real images are incorporated for training. This performance gain curve demonstrates the potential benefits of further scaling up the real data for our self-training method.

Effectiveness - Performance Gain Analysis. We intend to further validate the effectiveness of using single-view images for self-training. For this purpose, we compare the

performance gain of using self-training with another method to scale up the training data, i.e., *using captured multi-view real images*. We report the performance gain of training with multi-view images by comparing LRM variants, i.e., an LRM when trained on synthetic data only and when trained with both synthetic and real-world MVIImgNet multi-view capture [20]. We report this gain as Δ multi-view. We also report the gain of our self-training method as Δ ours. We report the results on the four datasets. As shown in Table 9, our method achieves a larger performance gain while using much fewer images for training, showing its effectiveness. The limited improvement of Δ multi-view also indicates the limitation of leveraging multi-view real data for training. We include a detailed discussion and report the detailed numbers for computing Δ multi-view in Appendix.

More interestingly, Δ multi-view uses only MVIImgNet data for training and has limited, nearly zero, improvement on in-the-wild images. In contrast, our method achieves larger improvements by leveraging more in-the-wild data, demonstrating its success at improving generalization.

6. Conclusion

We present Real3D, the first LRM that can leverage single-view real images for training. This has the major advantage of enabling training on a seemingly endless data source, which is representative of the general object shape distribution. We propose a self-training framework using unsupervised losses to leverage these images. Additionally, we develop an automatic data curation method to deal with the noisy real-world data. Compared to previous work, Real3D shows consistent improvements in diverse evaluation sets and highlights the potential to improve LRMs by training in large-scale image collections.

Acknowledgment. We acknowledge the support of NSF-2047677, NSF-2413161, NSF-2504906, NSF-2515626, and Gifts from Adobe and Google. We would like to thank UT GenAI Center for generous support of GPU hours.

References

- [1] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for super-sizing 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3782, 2022.
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [3] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.
- [4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022.
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.
- [7] Jang Hyun Cho and Philipp Krähenbühl. Language-conditioned detection transformer. *arXiv preprint arXiv:2311.17902*, 2023.
- [8] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016.
- [9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.
- [10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.
- [11] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchun Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023.
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [13] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1536–1546, 2022.
- [14] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.
- [15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [16] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019.
- [17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- [18] Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models. In *European Conference on Computer Vision*, pages 333–350. Springer, 2025.
- [19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [20] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. <https://github.com/3DTopia/OpenLRM>, 2023.
- [21] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7498–7507, 2020.
- [22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [23] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.
- [24] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. Self-supervised 3d mesh reconstruction from single images.

- In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6002–6011, 2021.
- [25] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Zeroshape: Regression-based zero-shot shape reconstruction. *arXiv preprint arXiv:2312.14198*, 2023.
- [26] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. Leap: Liberate sparse-view 3d modeling from camera poses. *arXiv preprint arXiv:2310.01410*, 2023.
- [27] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *2024 International Conference on 3D Vision (3DV)*, pages 31–41. IEEE, 2024.
- [28] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. *arXiv preprint arXiv:2505.00702*, 2025.
- [29] Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin Shu, Kai Zhang, Sai Bi, Xin Sun, et al. Megasynt: Scaling up 3d scene reconstruction with synthesized data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16441–16452, 2025.
- [30] Li Jiang, Shaoshuai Shi, Xiaojuan Qi, and Jiaya Jia. Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In *Proceedings of the European conference on computer vision (ECCV)*, pages 802–816, 2018.
- [31] Ying Jin, Jiaqi Wang, and Dahua Lin. Semi-supervised semantic segmentation via gentle teaching assistant. *Advances in Neural Information Processing Systems*, 35:2803–2816, 2022.
- [32] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [33] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [34] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1966–1974, 2015.
- [35] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.
- [36] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.
- [37] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 452–461, 2020.
- [38] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.
- [39] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 677–693. Springer, 2020.
- [40] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *European Conference on Computer Vision*, pages 515–534. Springer, 2022.
- [41] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdfsrn: Learning signed distance 3d object reconstruction from static images. *Advances in Neural Information Processing Systems*, 33:11453–11464, 2020.
- [42] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023.
- [43] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7708–7717, 2019.
- [44] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021.
- [45] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.
- [46] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024.
- [47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [48] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8446–8455, 2023.
- [49] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

- [50] Lu Mi, Abhijit Kundu, David Ross, Frank Dellaert, Noah Snavely, and Alireza Fathi. im2nerf: Image to neural radiance field in the wild. *arXiv preprint arXiv:2209.04061*, 2022.
- [51] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.
- [52] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [53] Tom Monnier, Matthew Fisher, Alexei A Efros, and Mathieu Aubry. Share with thy neighbors: Single-view reconstruction by cross-instance consistency. In *European Conference on Computer Vision*, pages 285–303. Springer, 2022.
- [54] KL Navaneet, Ansu Mathew, Shashank Kashyap, Wei-Chih Hung, Varun Jampani, and R Venkatesh Babu. From image collections to point clouds with self-supervised shape and pose networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1132–1140, 2020.
- [55] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019.
- [56] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.
- [57] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [58] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.
- [59] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. *arXiv preprint arXiv:2311.16918*, 2023.
- [60] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [61] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omniscient supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018.
- [62] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021.
- [63] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021.
- [64] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- [65] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. Vq3d: Learning a 3d-aware generative model on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4240–4250, 2023.
- [66] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- [67] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.
- [68] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.
- [69] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.
- [70] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. *arXiv preprint arXiv:2303.01416*, 2023.
- [71] Stanislaw Szymanowicz, Christian Ruppert, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. *arXiv preprint arXiv:2312.13150*, 2023.
- [72] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22819–22829, 2023.
- [73] Jiayang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024.
- [74] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [75] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian

- Laforte, Varun Jampani, and Yan-Pei Cao. Tripopr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.
- [76] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [77] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020.
- [78] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017.
- [79] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [81] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.
- [82] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [83] 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.
- [84] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrn: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024.
- [85] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9065–9075, 2023.
- [86] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1–10, 2020.
- [87] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, 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*, pages 803–814, 2023.
- [88] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2383–2393, 2023.
- [89] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.
- [90] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022.
- [91] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.
- [92] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024.
- [93] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10368–10378, 2021.
- [94] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.
- [95] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.
- [96] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8843–8852, 2021.
- [97] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.
- [98] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023.
- [99] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-irm: Large reconstruction model for 3d gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024.
- [100] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

- [101] Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schroff, Hartwig Adam, Ting Liu, Boqing Gong, et al. Distilling vision-language models on millions of videos. *arXiv preprint arXiv:2401.06129*, 2024.
- [102] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. *arXiv preprint arXiv:2312.09147*, 2023.