

Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers

Zi-Xin Zou^{1*} Zhipeng Yu² Yuan-Chen Guo^{1*} Yangguang Li²

Ding Liang² Yan-Pei Cao^{2†} Song-Hai Zhang^{3,1†}

¹BNRist, Tsinghua University ²VAST ³Qinghai University

{zouzx19@mails., guoyc19@mails., shz@tsinghua.edu.cn, yuzhipeng21@mails.ucas.ac.cn}

liyangguang@vastai3d.com, liangding1990@163.com, caoyanpei@gmail.com

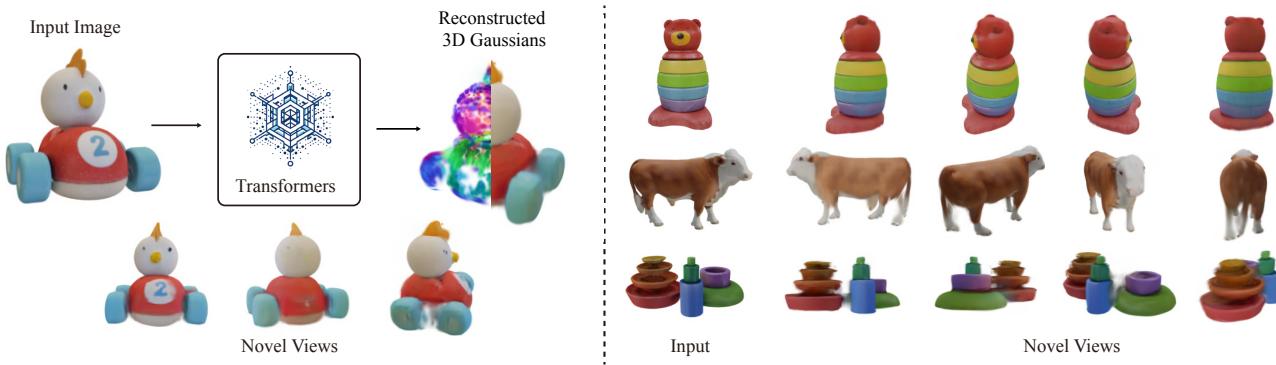


Figure 1. We propose a method that enables fast reconstruction from a single-view image. We build the 3D representation upon a hybrid Triplane-Gaussian representation by evaluating a transformer-based framework, from which 3D Gaussians would be decoded (Left). Based on this, we can perform novel view synthesis with high quality and fast rendering speed via Gaussian Splatting (Right).

Abstract

Recent advancements in 3D reconstruction from single images have been driven by the evolution of generative models. Prominent among these are methods based on Score Distillation Sampling (SDS) and the adaptation of diffusion models in the 3D domain. Despite their progress, these techniques often face limitations due to slow optimization or rendering processes, leading to extensive training and optimization times. In this paper, we introduce a novel approach for single-view reconstruction that efficiently generates a 3D model from a single image via feed-forward inference. Our method utilizes two transformer-based networks, namely a point decoder and a triplane decoder, to reconstruct 3D objects using a hybrid Triplane-Gaussian intermediate representation. This hybrid representation strikes a balance, achieving a faster rendering speed compared to implicit representations while simultaneously delivering superior rendering quality than explicit representations. The point decoder is designed for generating point clouds from single images, offering an explicit representation which is then utilized by the triplane decoder to query Gaussian features for each point. This

design choice addresses the challenges associated with directly regressing explicit 3D Gaussian attributes characterized by their non-structural nature. Subsequently, the 3D Gaussians are decoded by an MLP to enable rapid rendering through splatting. Both decoders are built upon a scalable, transformer-based architecture and have been efficiently trained on large-scale 3D datasets. The evaluations conducted on both synthetic datasets and real-world images demonstrate that our method not only achieves higher quality but also ensures a faster runtime in comparison to previous state-of-the-art techniques. Please see our project page at <https://zouzx.github.io/TriplaneGaussian/>

1. Introduction

Digitizing 3D objects from single 2D images represents a crucial and longstanding challenge in both computer vision and graphics, with wide applications in augmented reality (AR) and virtual reality (VR). However, the inherent ambiguity and lack of information in single images pose a substantial challenge in accurately recovering the complete,

* Intern at VAST

† Corresponding authors

high-quality shape and texture of an object from such a constrained perspective.

The recent surge in image generation using diffusion models [54, 56] enabled approaches to “imagine” images from novel viewpoints by incorporating pose transformation conditions [36] or novel view rendering feature maps [5, 82, 83]. Despite these advancements, achieving consistent novel view synthesis remains challenging due to the lack of 3D structural constraints. Even with the integration of multi-view attention or 3D-aware feature attention for simultaneous multi-view generation [22, 37, 59], this challenge persists. Furthermore, these methods require a time-intensive, object-specific optimization process to distill diffusion priors [7, 33, 40, 50, 60], combined with monocular cues (depth or normals [14, 38]), into neural 3D representations [43, 58]. This optimization procedure is often too slow for practical applications requiring rapid 3D creation.

Instead of optimizing a 3D representation, alternative approaches utilize category-specific shape priors to regress 3D representations (e.g., point clouds [15, 70], voxels [19, 69], meshes [63, 67], and NeRF [77]) from single images. The triplane representation stands out for its compactness and efficient expressiveness, leveraging its implicit nature for learning 3D structures via volume rendering [43]. Nevertheless, volume rendering’s inherent complexity incurs significant runtime and memory costs, hindering training efficiency and real-time rendering. Recently, Gaussian Splatting [28, 71] has advanced by employing anisotropic 3D Gaussians for high-quality radiance field representation, improving both novel view synthesis quality and rendering speed through adaptive density control and efficient, differentiable rendering. Yet, the direct learning of 3D Gaussian representations from images presents a substantial challenge due to their discrete, non-structural, and higher-dimensional nature compared to implicit representations that only require decoding RGB and density. Moreover, the dimension of Gaussian is higher than that of the point cloud, and there are few correlations between different parameters (e.g., position and SH coefficients). This complexity poses a challenge for the model to learn the intricate relationships between each parameter in the same latent space. Consequently, developing an effective method to unleash the potential of Gaussian Splatting for generalizable 3D learning remains a significant area for exploration.

In this work, we introduce *TGS*, a novel 3D reconstruction approach from single-view images, combining the strengths of **Triplane** and **Gaussian Splatting**. Our method employs a hybrid explicit-and-implicit 3D representation, facilitating fast and high-quality 3D reconstruction and novel view synthesis. We diverge from traditional 3D representation methods by proposing a hybrid Triplane-Gaussian model, which marries an explicit point cloud with

an implicit triplane field. The explicit point cloud outlines the object’s rough geometry, while the implicit triplane field refines the geometry and encodes 3D Gaussian properties such as opacity and spherical harmonics. This separation of position (explicit) and Gaussian properties (implicit) significantly enhances the quality of 3D reconstruction, while preserving the efficiency of Gaussian Splatting. Our approach consists of two networks for reconstructing the point cloud and triplane from the input image, employing a fully transformer-based architecture for both. This design enables interaction between latent features and input image features through cross-attention, ensuring scalability and supporting large-scale, category-agnostic training for enhanced real-world object generalizability. Moreover, we augment each 3D Gaussian with local image features by projecting explicit geometry onto the input image plane.

To the best of our knowledge, this is the first study to achieve generalizable 3D reconstruction from single-view images using Gaussian Splatting, opening new possibilities in fast 3D content creation and rendering. Through extensive experiments on the Google Scanned Object (GSO) dataset [13], our method exhibits superior performance over existing baselines in both geometry reconstruction and novel-view synthesis. Owing to its feed-forward model and the hybrid representation, our approach achieves object reconstruction in less than a second, significantly faster than prior methods.

2. Related Works

Single View 3D Reconstruction. 3D reconstruction from single view image is an ill-posed problem that garnered substantial attention from many researchers, due to the lack of geometry cues from single view. Most studies learn strong 3D priors from 3D synthetic models [6] or real scans [9]. Previous works have constructed priors based on collections of 3D primitives using an encoder-decoder approach [42, 75]. For example, they utilize an image encoder to encode visual information into compact latent features (e.g., vectors and volumes), and the 3D model can be decoded from the features for specific representation. However, these methods often are limited to specific categories and have poor performance on texture. Recently, transformer architecture [61] has shown its capacity and scalability in the field of computer vision [4, 12, 23, 52, 55] and other areas [3, 11]. GINA-3D [57] and LRM [21] propose a full transformer-based pipeline to decode a NeRF representation from triplane features. Some other works based on LRM architecture are proposed for multi-view image generation [76], sparse-view reconstruction [31], and pose-free sparse-view reconstruction [64]. MCC [68] also trains a transformer decoder to predict occupancy and color from RGB-D inputs. In this work, we adopt the scalable transformer-based architecture to learn generalizable 3D representation from a single image.

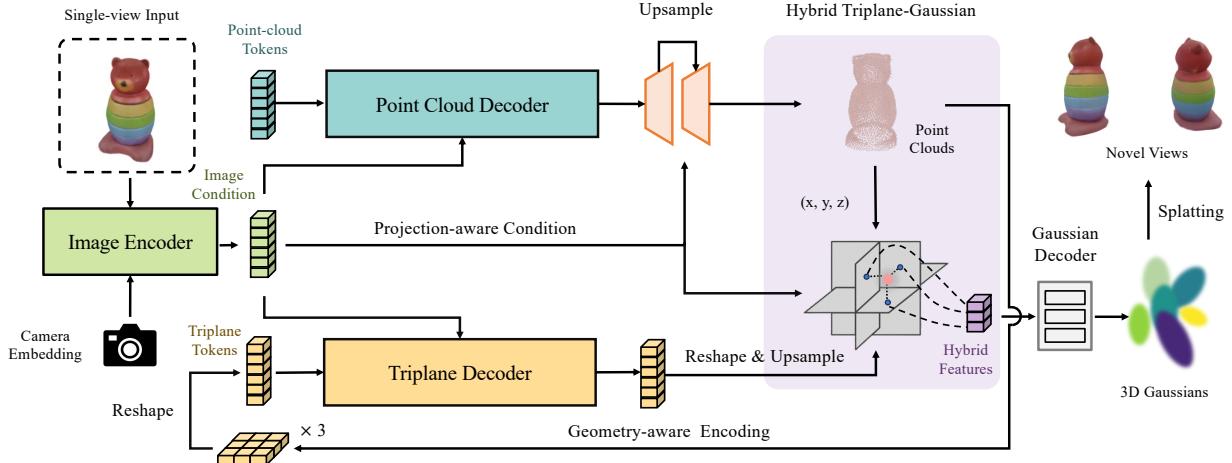


Figure 2. The overview of our framework. Given an image with its camera parameters, we first encode them into a set of latent feature tokens by leveraging a pre-trained ViT model. Our two transformer-based networks, point cloud decoder, and triplane decoder, take initial positional embedding as input and project image tokens onto latent feature tokens of respective 3D representation via cross-attention. Subsequently, a point cloud and a triplane can be de-tokenized and upsampled from the output of decoders, respectively. Additionally, we employ condition-aware projection to enhance both point cloud upsampling and Gaussian features. We also utilize geometry-aware encoding to project point cloud features into the initial positional embedding of triplane latent, aiming to achieve better alignment between two representations. Finally, 3D Gaussians are decoded by the point cloud, the triplane features and image features for novel view rendering.

3D Generation conditioned on the image. Contrary to learning 3D geometry prior and reconstructing an object from the image in a feed-forward manner, some recent works [32] treat this task as a conditional generation. They aim to exploit techniques such as GAN [16], diffusion model [8, 27, 41, 78, 81] to sample a high 3D content from pre-trained distribution. Point-E [45] and Shap-E [26] are trained on several million 3D models to generate point clouds and parameters of implicit function (e.g. neural radiance field). However, these 3D native generative models are limited due to the lack of a substantial amount of high-quality 3D model datasets.

Inspired from success of 2D generative models (e.g. DALL-E [54] and Stable Diffusion [56]) with impressive and high-quality image generation, a large number of studies [7, 33, 40, 50] incorporate 2D diffusion and CLIP priors into per-shape optimization via various score distillation sampling strategy [50, 62, 66], or with other monocular prior losses. Some works explore view-conditioned diffusion to learn control mechanisms that manipulate the camera viewpoint in large-scale diffusion models [36]. However, these methods necessitate a time-consuming per-shape optimization for different 3D representations (i.e., NeRF, mesh, 3D Gaussian, SMPL human model) via differentiable rendering. Some other works improve the multi-view consistency across different novel views [37, 59], which directly enables NeRF optimization with pixel-wise loss with distillation, achieving impressive performance in few minutes. One-2-3-45 [35] proposes to combine the 2D generative models and multi-view 3D reconstruction, leading to superiority in both quality and efficiency.

Neural 3D Representation and Rendering. With the recent advancement of neural networks, neural implicit representations [25, 42, 47], especially neural radiance field (NeRF) [43] with its variants [2, 17, 34, 72, 80], have shown its promising performance in novel-view synthesis and 3D reconstruction, only using dense images via differentiable volume rendering. Despite efforts to enhance the rendering speed of neural implicit representations [18, 20, 44], their efficiency still lags behind point-based rendering [29, 30]. Certain works augment points with neural features and utilize UNet [1, 53] for fast or even real-time rendering, leveraging its advantages of fast GPU/CUDA-based rasterization. These methods, relying on 2D CNN and lacking 3D awareness, often suffer from over- or under-reconstruction in challenging scenarios (e.g. featureless areas or thin structures). The 3DGS [28], as a special point-based rendering, stands out for its distinctive utilization of explicit representation and differential point-based splatting techniques, thereby facilitating real-time rendering at novel views. However, directly regressing 3D Gaussians from the image for high-quality novel view synthesis remains challenging. In this paper, we leverage a hybrid 3D representation, facilitating fast and generalizable reconstruction and rendering.

3. Method

In the subsequent sections, we present our approach for 3D object reconstruction from single-view images. We introduce a new hybrid 3D representation that combines explicit point cloud geometry and implicit triplane features, allowing for efficient rendering without compromising on qual-

ity (Section 3.1). In order to deduce the hybrid representation from a single-view input, we first employ a transformer-based point cloud decoder to predict coarse points from image features and upsample the coarse points to a dense point cloud. Subsequently, a triplane decoder takes these points along with the image features and outputs the triplane features. Finally, the hybrid Triplane-Gaussian will encompass point clouds, triplane features, and projected image features to generate representative hybrid features for 3D Gaussian attributes decoding (Section 3.2). Benefiting from differentiable splatting, we can train the full framework end-to-end on large-scale category-agnostic datasets efficiently (Section 3.3).

3.1. Hybrid Triplane-Gaussian

Gaussian Splatting, while offering advantages such as high-quality rendering and speed, remains unexplored in its application to generalizable reconstruction settings. A preliminary approach might consider treating 3D Gaussians as an explicit point cloud enriched with specific attributes. One possibility involves adapting existing point cloud reconstruction or generation models [41, 45] to include additional 3D Gaussian attributes besides point positions. Pursuing this concept, we initially developed a 3D Gaussian reconstruction model based on a transformer architecture aimed at directly predicting 3D Gaussians. However, as shown in the experiment in Section 4.6, it fails to generate satisfied 3D objects from a single image. This shortfall might be attributed to the discrete, non-structural, and high-dimensional nature of 3D Gaussians, which complicates the learning process. In response, we introduce *Triplane-Gaussian*, a new hybrid 3D representation that merges the benefits of both triplane and point cloud approaches for 3D Gaussian representation. This hybrid model offers fast rendering, high quality, and generalizability, as depicted in Figure 5.

The hybrid representation includes a point cloud $P \in R^{N \times 3}$ which provides explicit geometry, and a triplane $T \in R^{3 \times C \times H \times W}$ which encodes an implicit feature field, where 3D Gaussian attributes can be decoded. The triplane T comprises three axis-aligned orthogonal feature planes $\{T_{xy}, T_{xz}, T_{yz}\}$. For any position \mathbf{x} , we can query the corresponding feature vector from triplane by projecting it onto axis-aligned feature planes, and concatenate three trilinear interpolated features as the final feature $f_t = \text{interp}(T_{xy}, p_{xy}) \oplus \text{interp}(T_{xz}, p_{xz}) \oplus \text{interp}(T_{yz}, p_{yz})$, where interp and \oplus denote the trilinear interpolation and concatenation operation, and p denotes the projected position on each plane.

3D Gaussian Decoder. For a given position $\mathbf{x} \in R^3$ from point cloud P , we query feature f from the triplane and adopt an MLP ϕ_g to decode the attributes of the 3D Gaussians derived from the point cloud. The 3D Gaussians’

attributes include opacity α , anisotropic covariance (represented by a scale s and rotation q) and spherical harmonics (SH) coefficients sh [28]:

$$(\Delta\mathbf{x}', \alpha, s, q, sh) = \phi_g(\mathbf{x}, f) \quad (1)$$

Since surface points are not always the best choice for the 3D Gaussian representation, we additionally predict a small offset $\Delta\mathbf{x}$ for the position, and the final position is computed by $\mathbf{x} = \mathbf{x} + \Delta\mathbf{x}$.

Regarding the single-view 3D reconstruction task we focus on in this paper, we additionally augmented the queried triplane features with projected image features as in PC² [41] to fully utilize the local features from the input image. Specifically, we concatenate the triplane feature f_t with projected local features f_l from explicit geometry as f in Equation 1. Given an input camera pose π and a point cloud P , the local projection feature can be calculated by the projection function \mathcal{P} , where $f_l = \mathcal{P}(\pi, P)$. Following [41], we consider the self-occlusion problem of the point cloud and implement the projection by point rasterization. The local features include the RGB color, DINOV2 [46] feature, mask, and 2-dimensional distance transform corresponding to the mask region. In this way, the projection-aware condition can improve the texture quality under the input viewing angle.

Rendering. Based on such representation, we can perform image rendering by utilizing the efficient Gaussian Splatting rendering approach [28] at any novel viewpoint. It’s a differentiable tile-based rasterization that allows fast α -blending of anisotropic splats and fast backward pass by tracking accumulated α values, ensuring that our pipeline can be trained end-to-end more efficiently with higher resolution images and less GPU memory cost.

3.2. Reconstruction from Single-View Images

In this section, we detail the process of learning the Triplane-Gaussian representation from a single image. Drawing inspiration from recent breakthroughs in 3D reconstruction using transformers [21], our approach employs two transformer-based decoders, each tailored for reconstructing a point cloud P and a triplane F from input images. The adoption of a scalable transformer architecture substantially enhances the model’s capability and generalization potential. Moreover, we harness local features projected from the input image to enhance both the point cloud up-sampling process and the 3D Gaussian decoder. This strategy significantly improves initial point cloud reconstruction and facilitates more accurate novel-view synthesis.

Image Encoder. We first leverage a pre-trained ViT-based model [12] (e.g., DINOV2 [46]) to obtain patch-wise feature tokens from the input image. In order to better utilize camera information in our pipeline, we implement an adaptive layer norm (adaLN) to modulate the DINOV2 features

with the camera features, which is similar to [21, 48]. Finally, we implement an MLP to predict the scale and shift based on the camera features after each normalization layer. These camera modulation layers are trained to make image features aware of the observation viewpoint, thereby better guiding both the point cloud network and the triplane network.

Transformer Backbone. In our framework, we use a set of feature tokens $\{f_s\}_p$ and $\{f_i\}_t$ for the latent features of two different 3D representations, i.e., points and triplane, respectively. The latent tokens are initialized from learnable positional embeddings, and fed to the transformer blocks. Similar to other transformer architecture designs [24], each transformer block comprises a self-attention layer, a cross-attention layer, and a feed-forward layer. The viewpoint-augmented image tokens guide the two decoders via cross-attention, respectively.

Point Cloud Decoder. The point cloud decoder provides the coarse geometry of the object, where 3D Gaussians can be produced based on the coordinates of the points. We employ a 6-layer transformer backbone to decode a point cloud using image conditions from a fixed number of learnable positional embeddings, i.e., point cloud tokens. We treat each latent token as a point, which is similar to Point-E [45]. However, our approach operates in a feed-forward manner, where the point cloud can be predicted directly from the final token, as opposed to the diffusion process employed in Point-E [45]. Due to limited computation and memory resources, we only decode a coarse point cloud with 2048 points in this step.

Point Upsampling with Projection-Aware Conditioning. Since the quality of novel view synthesis from Gaussian Splatting is highly influenced by the number of Gaussians, it's not enough to generate 3D Gaussians from such a low-resolution point cloud. Therefore, we adopt a lifting module with two-step Snowflake point deconvolution (SPD) [73, 74] to densify the point clouds from 2048 points to 16384 points. Snowflake utilizes a global shape code extracted from the input point cloud along with point features, to predict the point displacements for the upsampled point cloud. In order to reconstruct the detailed geometry guided by the input image, we integrate local image features into the shape code of SPD from the given camera pose by Projection-Aware Conditioning [41]. Such projection enables the generation of high-resolution point clouds that are well-aligned with the input image.

Triplane Decoder with Geometry-Aware Encoding. The triplane decoder outputs the implicit feature field based on the image and initial point cloud, from which 3D Gaussian properties will be decoded by position queries. It shares a similar transformer architecture to decode the features from learnable positional embedding by image tokens.

We also encode the point cloud into the initial learnable positional embeddings, resulting in better geometry awareness. More specifically, the points are also augmented with projection features from the input image as in point cloud upsampling. The feature-rich points are then fed into a shallow PointNet [51] with local pooling [49], and then we perform an orthographic projection onto the three axis-aligned planes. The features projected to the same token are average-pooled and added with the learnable positional embedding.

3.3. Training

We train the full pipeline by using 2D rendering loss along with 3D point cloud supervision:

$$\begin{aligned} \mathcal{L} = & \lambda_c \mathcal{L}_{\text{CD}} + \lambda_e \mathcal{L}_{\text{EMD}} + \\ & \frac{1}{N} \sum_{i=1}^N (\mathcal{L}_{\text{MSE}} + \lambda_m \mathcal{L}_{\text{MASK}} + \lambda_s \mathcal{L}_{\text{SSIM}} + \lambda_l \mathcal{L}_{\text{LPIPS}}) \end{aligned} \quad (2)$$

where \mathcal{L}_{CD} is the Chamfer Distance (CD) and \mathcal{L}_{EMD} is the Earth Mover's Distance (EMD). For training the triplane decoder and 3D Gaussian decoder, we apply the rendering losses, including a pixel-wise MSE loss $\mathcal{L}_{\text{MSE}} = \|I - \hat{I}\|_2^2$, a mask loss $\mathcal{L}_{\text{MASK}} = \|M - \hat{M}\|_2^2$, a SSIM loss $\mathcal{L}_{\text{SSIM}}$ [65] and a perceptual loss $\mathcal{L}_{\text{LPIPS}}$ [79].

4. Experiments

4.1. Implementation Details

We utilize a pre-trained DINOv2 [46] as our image encoder, which generates 768-dimension feature tokens with a patch size of 14. The point decoder and triplane decoder are both 10-layer transformer networks with hidden dimension 512. The positional embeddings of the point decoder consist of 2048 tokens with 512 dimensions, corresponding to 2048 points. The positional embeddings of the triplane decoder comprise three 32×32 tokens, each with 512 dimensions, representing three axis-aligned planes. To ensure better convergence during training, we divide the training process into two stages. Initially, we train the point decoder using $\lambda_c = 10$ and $\lambda_e = 10$ only with CD loss and EMD loss, enabling the network to deliver high-quality geometry. Subsequently, we train the triplane decoder and the Gaussian decoder with the point decoder frozen, and set the weights of the losses with $\lambda_m = 1$, $\lambda_s = 1$, $\lambda_l = 2$, $\lambda_c = 10$, $\lambda_e = 0$.

4.2. Baselines, Dataset, and Metrics

Baselines. We compare our method with the previous state-of-the-art single-image reconstruction and generation methods. There are three kinds of them: (1) native 3D generative models, including Point-E [45] and Shap-E [26]; (2) 2D diffusion model, i.e., Zero-1-2-3 [36] and (3) feed-forward model based on 2D diffusion model output, i.e.,

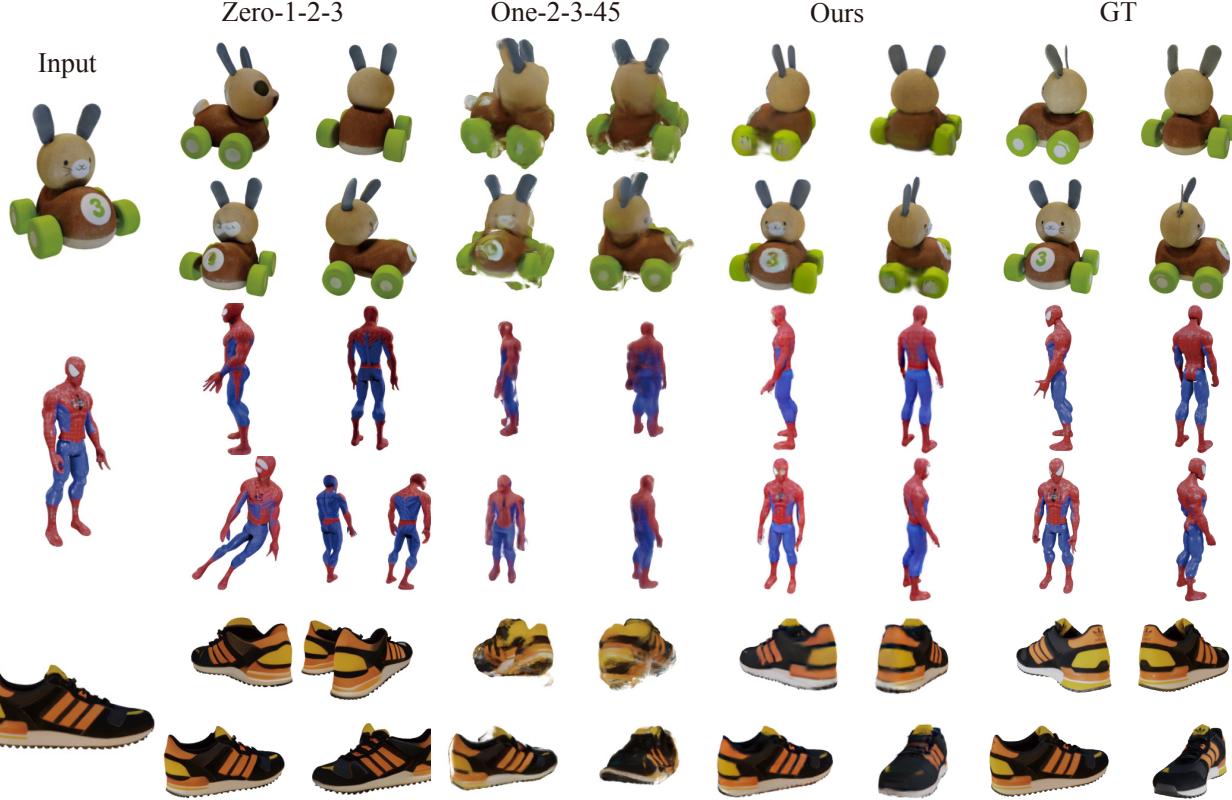


Figure 3. Qualitative comparisons of novel view synthesis from reconstructed object between our method and other baselines on the GSO dataset. Our approach achieves both quality and consistency across different novel views.

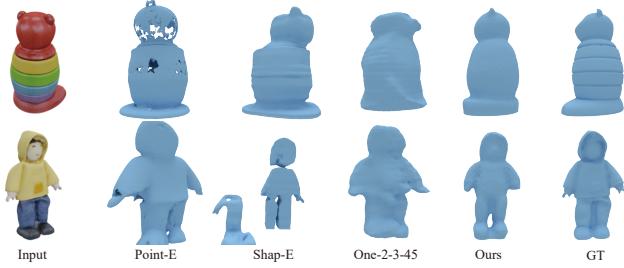


Figure 4. Qualitative comparisons of geometry reconstruction from a single image between our method and other baselines on the GSO dataset.

One-2-3-45 [35]. Point-E [45] and Shap-E [26] are trained on a large-scale internal 3D dataset to produce point cloud and implicit representation (NeRF) from single images. Zero-1-2-3 [36] is a 2D generative model to utilize view-conditioned diffusion for novel view synthesis. One-2-3-45 [35] trains a robust multi-view reconstruction model which takes multi-view images generated from a 2D diffusion model (e.g., Zero-1-2-3). It directly builds an SDF field by SparseNeuS [39], achieving state-of-the-art performance.

Dataset. Following One-2-3-45 [35], we train our model on the Objaverse-LVIS dataset [10], which contains 46K

3D models in 1,156 categories. We use Blender to render ground-truth RGB and depth with a circle path for an object. For Evaluation, we adopt the Google Scanned Objects (GSO) dataset [13], which includes a wide variety of high-quality scanned household items, to evaluate the performance of our method and other baselines. We randomly choose 100 shapes and render a single image per shape for evaluation.

4.3. Single View Reconstruction

We conduct the evaluation quality of geometry reconstruction from single-view images with other baselines, including Point-E [45], Shap-E [26] and One-2-3-45 [35]. Table 1 demonstrates the quantitative results on CD and Volume IoU, and Figure 4 illustrates some visual comparisons. Point-E tends to produce holes in the reconstructed meshes due to the sparsity of the generated point cloud. Meanwhile, Shap-E is susceptible to collapsing during the generation process, resulting in an unpredictable outcome. The geometric output generated by One-2-3-45 [35] is characterized by a coarse representation, lacking intricate details. In contrast, our method reconstructs geometry based on the explicit point clouds, which can preserve finer details and maintain consistency with the ground truth.

	CD ↓	IoU ↑	Time (s) ↓
Point-E	68.56	0.181	78
Shape-E	79.84	0.218	27
One-2-3-45	75.56	0.338	40
Ours	21.04	0.401	0.14

Table 1. Quantitative Comparison for single view 3D reconstruction on the GSO dataset, in terms of Chamfer Distance $\times 10^{-3}$, Volume IoU and runtime efficiency.

	PSNR ↑	SSIM ↑	LPIPS ↓	Time (s) ↓
One-2-3-45	15.55	0.76	0.25	14
Zero-1-2-3	17.57	0.78	0.19	1.70
PixelNeRF	20.64	0.82	0.29	4.09
Ours	23.15	0.87	0.13	0.003

Table 2. Quantitative comparison on novel-view synthesis from single images on the GSO dataset, in terms of PSNR, SSIM, LPIPS, and runtime efficiency.

	PSNR ↑	SSIM ↑	LPIPS ↓
3DG	18.56	0.80	0.20
Triplane-NeRF	21.85	0.84	0.15
Triplane-Gaussian	23.15	0.87	0.13

Table 3. Quantitative comparison between different representations for novel view synthesis on GSO.

4.4. Novel View Synthesis

In this subsection, we evaluate the quality of novel view synthesis in comparison with baselines, including One-2-3-45 and Zero-1-2-3. The quantitative and qualitative results are shown in Table 2 and Figure 3, respectively. While Zero-1-2-3 can generate plausible results from a novel viewpoint, it doesn't ensure consistency between different views, and can even yield completely unreasonable results. For example, it may generate two different instances of Spider-Man and shoes when the input only contains one of each. Although One-2-3-45 is primarily designed for geometry reconstruction, we utilize it to render the novel view image from its generated feature volume. Due to its inherent geometry constraint in 3D volume, the rendering images are consistent across different views. However, its reliance on images generated by Zero-1-2-3 limits its ability to produce high-quality novel view synthesis. In contrast, Our method delivers both high quality and consistency, thereby attaining superior performance. Additionally, by leveraging the transformer architecture and local feature projection, our model exhibits robust generalization to unseen objects while preserving intricate textures.

4.5. Runtime Efficiency

We also evaluate the runtime efficiency of our method in comparison with other baseline approaches. Table 1 and Ta-

	CD ↓	IoU ↑
PCL tokens	22.16	0.385
Global image feature	22.22	0.401
Projection-aware condition	21.04	0.401

Table 4. Quantitative Comparison of different conditions for point cloud up-sampling in geometry reconstruction.

	P.C.	G.E.	GT 3D	PSNR ↑	SSIM ↑	LPIPS ↓
a	X	X	✓	22.04	0.85	0.16
b	✓	X	✓	22.47	0.86	0.15
c	X	✓	✓	22.76	0.86	0.14
d	✓	✓	✓	23.15	0.87	0.13
e	✓	✓	X	22.65	0.86	0.14

Table 5. Quantitative effect of projection-aware condition, geometry-aware encoding and ground-truth 3D supervision to novel view synthesis.

ble 2 demonstrate the runtime of reconstruction and rendering of each baseline, respectively. We evaluate the runtime of Zero-1-2-3's diffusion sampling process due to its lack of rendering process from 3D representation. We evaluate the volume rendering of One-2-3-45 from its generated implicit feature volume rather than the generated mesh. We can find that our method has achieved significant improvements in speed for both reconstruction and rendering processes compared to other baselines, benefiting from feed-forward fashion and efficient rasterization.

4.6. Ablation Study

3D representation. To evaluate the effectiveness of the design of our Triplane-Gaussian design, we conducted an experiment that compared it with two other representations, including (1) native generalizable 3D Gaussians (3DG) and (2) Triplane-NeRF. Table 3 and Figure 5 (left) present the quantitative and qualitative comparisons among these methods. The native generalizable 3D Gaussians, which are directly decoded from latent tokens like points, exhibit the poorest performance according to all metrics. Our Triplane-Gaussian, leveraging the projection-aware condition with explicit geometry, excels in producing more detailed texture compared to Triplane-NeRF, as illustrated in the red box of Figure 5 (left), and achieves the best quantitative results. Furthermore, once the 3D Gaussians are decoded, our rendering process demonstrates faster performance compared to Triplane-NeRF (48ms).

Projection-aware condition. We first conduct experiments with two ablation shape code settings to investigate the impact of different shape codes within the point upsampling module, including (1) the point cloud (PCL) tokens and (2) global image features from DINOv2. As shown in Table 4, image-based shape codes achieve higher IoU com-

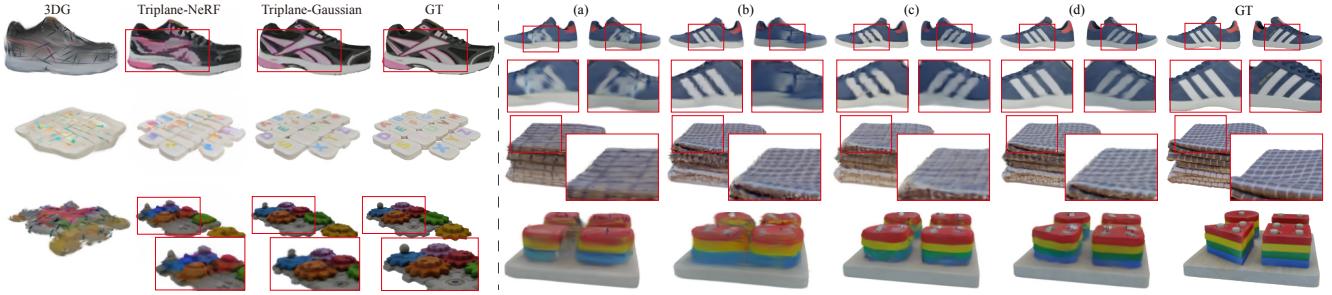


Figure 5. Qualitative comparison with 3DG and Triplane-NeRF (left) and qualitative effect of projection-aware condition and geometry-aware encoding (right), where the (a-d) are corresponding with (a-d) in Table 5.

pared to merely PCL tokens, which demonstrates the importance of incorporating structural information from the visual modality. Moreover, the projection-aware condition outperforms the other two code methods in terms of CD, suggesting that the local patterns of the image are effectively transferred to the structure of the point cloud.

We also assess the impact of the projection-aware condition on the 3D Gaussian Decoder for novel view synthesis. Table 5 demonstrates that using projection-aware condition (P.C.) improves the quality of novel view synthesis across all metrics. In Figure 5 (right), two examples illustrate its visual effects. The use of P.C. (Figure 5 (b) and (d)) significantly enhances the rendering image with sharper and more detailed texture on the same side as the input view. For example, in the first row, the left image of the shoe, and in the second row, the complex checkered texture of the coverlet.

Geometry-aware encoding. While the projection-aware condition improves rendering quality on the same side as the input view, achieving good texture on the backside remains challenging (see the first row of Figure 5 (b)). Introducing geometry-aware encoding (G.E.) enhances texture results, as evident in Figure 5 (c) and (d), particularly on the backside (right image of the shoe). Geometry-aware encoding enables the triplane decoder to predict backside texture with a shape prior, considering that shoes often have a similar texture on both sides. Since our point cloud decoder and triplane decoder are trained separately, the implicit field is sometimes not well-aligned with the explicit point cloud for complex geometry in the absence of geometry-aware encoding. This leads to blurry and non-sharp rendering results, as observed in the last row of Figure 5 (a) and (b). The geometry-aware encoding enhances the alignment between them, resulting in high-quality novel view synthesis. Table 5 also demonstrates the quantitative improvement achieved by geometry-aware encoding in novel view synthesis results.

3D supervision. We find it hard to train our model successfully only by rendering loss after some attempts, so we leverage the 3D supervision from the ground-truth point cloud for better convergence. However, for some multi-

view datasets, obtaining accurate and complete ground-truth 3D models is not an easy task. In the absence of ground-truth 3D models, we can address this issue by overfitting each object using Gaussian Splatting and leveraging them for pseudo-point-cloud supervision. Table 5 demonstrates that the pseudo-point-cloud supervision only leads to a slight decrease while maintaining high-quality results, which indicates that this solution can mitigate the reliance on ground-truth 3D models.

4.7. Limitations

The rendering quality of the 3D Gaussians largely depends on the initial geometry. If the predicted point cloud deviates severely from the ground truth, it becomes challenging for the 3D Gaussians to recover the missing regions. Additionally, since our model is not a probabilistic model, the backside tends to be blurry. One solution for this issue would be combined with GAN or diffusion model.

5. Conclusion

In this paper, we present *TGS*, a new framework for 3D reconstruction from a single image with only a feed-forward inference. Our approach constructs a hybrid Triplane-Gaussian representation using two transformer decoders from input images, and subsequently, a 3D Gaussian decoder generates Gaussian properties. To improve the consistency with the input observation, we further augment the generation process of these two representations by the application of projection-aware conditioning and geometry-aware encoding. Experimental results demonstrate that our method achieves high-quality geometry reconstruction and novel view synthesis while maintaining rapid reconstruction and rendering speeds.

Acknowledgments This work was supported by the National Key Research and Development Program of China (No.2023YFF0905104), the Natural Science Foundation of China (No.62132012, No.62361146854), Beijing Municipal Science and Technology Project (No.Z221100007722001) and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

- [1] Kara-Ali Aliiev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor S. Lempitsky. Neural point-based graphics. In *ECCV*, pages 696–712, 2020. 3
- [2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5835–5844, 2021. 3
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 2
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640, 2021. 2
- [5] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *CoRR*, abs/2304.02602, 2023. 2
- [6] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 2
- [7] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3
- [8] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G. Schwing, and Liangyan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *IEEE CVPR*, pages 4456–4465, 2023. 3
- [9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scanet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE CVPR*, pages 2432–2443, 2017. 2
- [10] 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 *IEEE CVPR*, pages 13142–13153, 2023. 6
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 2
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 4
- [13] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas Barlow McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, pages 2553–2560, 2022. 2, 6
- [14] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10766–10776, 2021. 2
- [15] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *IEEE CVPR*, pages 2463–2471, 2017. 2
- [16] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. GET3D: A generative model of high quality 3d textured shapes learned from images. In *NeurIPS*, 2022. 3
- [17] Yiming Gao, Yan-Pei Cao, and Ying Shan. Surfelnerf: Neural surfel radiance fields for online photorealistic reconstruction of indoor scenes. In *IEEE CVPR*, pages 108–118, 2023. 3
- [18] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien P. C. Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *ICCV*, pages 14326–14335, 2021. 3
- [19] Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, pages 484–499, 2016. 2
- [20] Yuan-Chen Guo, Yan-Pei Cao, Chen Wang, Yu He, Ying Shan, and Song-Hai Zhang. Vmesh: Hybrid volume-mesh representation for efficient view synthesis. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3
- [21] 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. *CoRR*, abs/2311.04400, 2023. 2, 4, 5
- [22] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. *arXiv preprint arXiv:2312.06725*, 2023. 2
- [23] Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Vitor Guizilini, Thomas Kollar, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9153–9164, 2023. 2
- [24] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 5
- [25] Chiyu "Max" Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas A. Funkhouser.

- Local implicit grid representations for 3d scenes. In *IEEE CVPR*, pages 6000–6009, 2020. 3
- [26] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *CoRR*, abs/2305.02463, 2023. 3, 5, 6
- [27] Animesh Karnewar, Andrea Vedaldi, David Novotný, and Niloy J. Mitra. HOLODIFFUSION: training a 3d diffusion model using 2d images. In *IEEE CVPR*, pages 18423–18433, 2023. 3
- [28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 2, 3, 4
- [29] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. Point-based neural rendering with per-view optimization. *Comput. Graph. Forum*, 40(4):29–43, 2021. 3
- [30] Christoph Lassner and Michael Zollhöfer. Pulsar: Efficient sphere-based neural rendering. In *IEEE CVPR 2021, virtual, June 19–25, 2021*, pages 1440–1449. Computer Vision Foundation / IEEE, 2021. 3
- [31] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *CoRR*, abs/2311.06214, 2023. 2
- [32] Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. Advances in 3d generation: A survey. *arXiv preprint arXiv:2401.17807*, 2024. 3
- [33] 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 *IEEE CVPR*, pages 300–309, 2023. 2, 3
- [34] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *NeurIPS 2020*, 2020. 3
- [35] 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. 3, 6
- [36] 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. 2023. 2, 3, 5, 6
- [37] 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. *CoRR*, abs/2309.03453, 2023. 2, 3
- [38] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. In *ICCV*, pages 12829–12838, 2021. 2
- [39] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, pages 210–227, 2022. 6
- [40] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion 360° reconstruction of any object from a single image. In *IEEE CVPR*, pages 8446–8455, 2023. 2, 3
- [41] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc^2 : Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *IEEE CVPR*, pages 12923–12932, 2023. 3, 4, 5
- [42] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE CVPR*, pages 4460–4470, 2019. 2, 3
- [43] 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. In *ECCV*, pages 405–421, 2020. 2, 3
- [44] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 3
- [45] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *CoRR*, abs/2212.08751, 2022. 3, 4, 5, 6
- [46] Maxime Oquab, Timothée Darct, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 5
- [47] Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE CVPR*, pages 165–174, 2019. 3
- [48] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 5
- [49] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*, pages 523–540, 2020. 5
- [50] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2, 3
- [51] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE CVPR*, pages 77–85, 2017. 5
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2
- [53] Ruslan Rakhimov, Andrei-Timotei Ardelean, Victor Lempitsky, and Evgeny Burnaev. NPBG++: accelerating neural point-based graphics. In *IEEE CVPR*, pages 15948–15958, 2022. 3

- [54] 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*, pages 8821–8831, 2021. 2, 3
- [55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 2
- [56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE CVPR*, pages 10674–10685, 2022. 2, 3
- [57] Bokui Shen, Xinchen Yan, Charles R. Qi, Mahyar Najibi, Boyang Deng, Leonidas J. Guibas, Yin Zhou, and Dragomir Anguelov. GINA-3D: learning to generate implicit neural assets in the wild. In *IEEE CVPR*, pages 4913–4926, 2023. 2
- [58] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *NeurIPS*, pages 6087–6101, 2021. 2
- [59] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *CoRR*, abs/2308.16512, 2023. 2, 3
- [60] 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 *ICCV*, pages 22819–22829, 2023. 2
- [61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2
- [62] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *IEEE CVPR*, pages 12619–12629, 2023. 3
- [63] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single RGB images. In *ECCV*, pages 55–71, 2018. 2
- [64] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. PF-LRM: pose-free large reconstruction model for joint pose and shape prediction. *CoRR*, abs/2311.12024, 2023. 2
- [65] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 5
- [66] 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. *CoRR*, abs/2305.16213, 2023. 3
- [67] Markus Worchsel, Rodrigo Diaz, Weiwen Hu, Oliver Schreer, Ingo Feldmann, and Peter Eisert. Multi-view mesh reconstruction with neural deferred shading. In *IEEE CVPR*, pages 6177–6187, 2022. 2
- [68] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *IEEE CVPR*, pages 9065–9075, 2023. 2
- [69] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5d sketches. In *NeurIPS*, 2017. 2
- [70] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. PQ-NET: A generative part seq2seq network for 3d shapes. In *IEEE CVPR*, pages 826–835, 2020. 2
- [71] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *arXiv preprint arXiv:2403.11134*, 2024. 2
- [72] Xiuzhe Wu, Peng Dai, Weipeng Deng, Handi Chen, Yang Wu, Yan-Pei Cao, Ying Shan, and Xiaojuan Qi. Cl-nerf: Continual learning of neural radiance fields for evolving scene representation. *Advances in Neural Information Processing Systems*, 36, 2023. 3
- [73] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. SnowflakeNet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 5
- [74] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflake point deconvolution for point cloud completion and generation with skip-transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6320–6338, 2023. 5
- [75] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomír Mech, and Ulrich Neumann. DISN: deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, pages 490–500, 2019. 2
- [76] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. DMV3D: denoising multi-view diffusion using 3d large reconstruction model. *CoRR*, abs/2311.09217, 2023. 2
- [77] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE CVPR*, pages 4578–4587, 2021. 2
- [78] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: latent point diffusion models for 3d shape generation. In *NeurIPS*, 2022. 3
- [79] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE CVPR*, pages 586–595, 2018. 5
- [80] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *IEEE CVPR*, pages 5439–5448, 2022. 3
- [81] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional SDF diffusion for controllable 3d shape generation. *ACM Trans. Graph.*, 42(4):91:1–91:13, 2023. 3

- [82] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *IEEE CVPR*, pages 12588–12597, 2023. 2
- [83] Zixin Zou, Weihao Cheng, Yan-Pei Cao, Shi-Sheng Huang, Ying Shan, and Song-Hai Zhang. Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7900–7908, 2024. 2