Meta 3D Gen

Raphael Bensadoun[1,*], **Tom Monnier**[1,*], **Yanir Kleiman**[1,*], **Filippos Kokkinos**[1], **Yawar Siddiqui**[1,2], **Mahendra Kariya**[1], **Omri Harosh**[1], **Roman Shapovalov**[1], **Benjamin Graham**[1], **Emilien Garreau**[1], **Animesh Karnewar**[1,2], **Ang Cao**[1,2], **Idan Azuri**[1], **Iurii Makarov**[1], **Eric-Tuan Le**[1], **Antoine Toisoul**[1], **David Novotny**[1,†], **Oran Gafni**[1,†], **Natalia Neverova**[1,†], **Andrea Vedaldi**[1,†]

[1]GenAI, Meta, [2]work done while doing internships at Meta
[*]Joint first authors, [†]Senior contributors

We introduce Meta 3D Gen (3DGen), a new state-of-the-art, fast pipeline for *text-to-3D asset generation*. 3DGen offers 3D asset creation with high prompt fidelity and high-quality 3D shapes and textures in under a minute. It supports physically-based rendering (PBR), necessary for 3D asset relighting in real-world applications. Additionally, 3DGen supports *generative retexturing* of previously generated (or artist-created) 3D shapes using additional textual inputs provided by the user. 3DGen integrates key technical components, Meta 3D AssetGen and Meta 3D TextureGen, that we developed for text-to-3D and text-to-texture generation, respectively. By combining their strengths, 3DGen represents 3D objects simultaneously in three ways: in view space, in volumetric space, and in UV (or texture) space. The integration of these two techniques achieves a win rate of 68% with respect to the single-stage model. We compare 3DGen to numerous industry baselines, and show that it outperforms them in terms of prompt fidelity and visual quality for complex textual prompts, while being significantly faster.

Date: June 25, 2024

Figure 1 **Meta 3D Gen** integrates Meta's foundation models for text-to-3D (Meta 3D AssetGen (Siddiqui et al., 2024)) and text-to-texture (Meta 3D TextureGen (Bensadoun et al., 2024)) generation in a unified pipeline, enabling efficient, state-of-the-art creation and editing of diverse, high-quality textured 3D assets with PBR material maps.

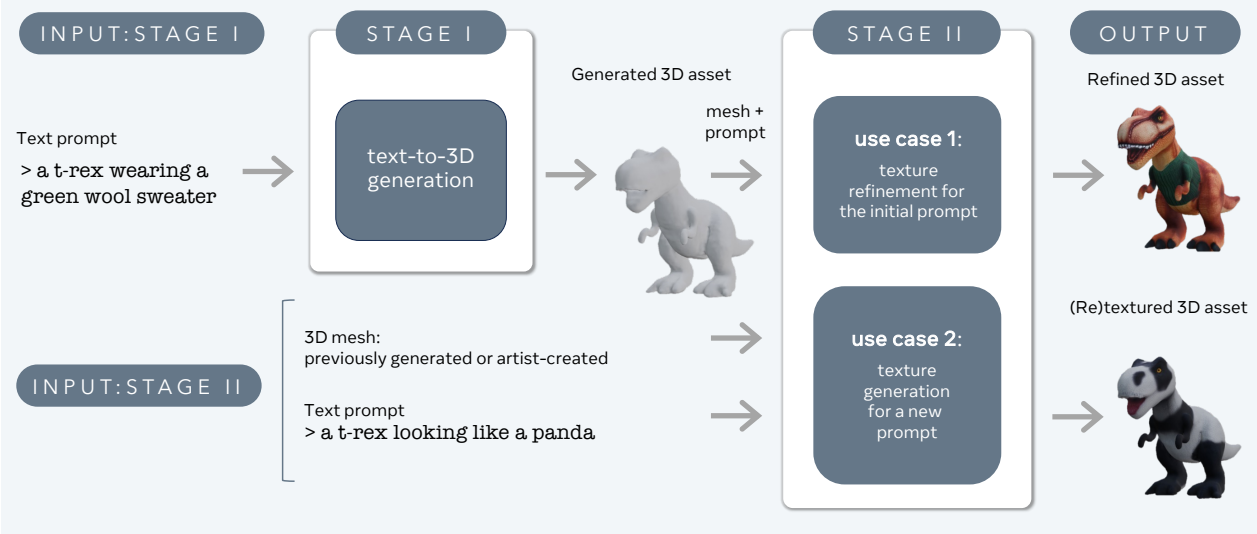

Figure 2 Overview of Meta 3D Gen. The pipeline takes a text prompt as an input and performs text-to-3D generation (Stage I, Siddiqui et al. (2024)), followed by texture refinement (Stage II, Bensadoun et al. (2024)). Stage II can also be used for *retexturing* of generated or artist-created meshes using new textual prompts provided by the user.

1 Introduction

We introduce Meta 3D Gen (3DGen), a new state-of-the-art solution for efficient text-to-3D generation. Text-to-3D is the problem of generating 3D content, such as characters, props and scenes, from textual descriptions. Authoring 3D content is one of the most time-consuming and challenging aspects of designing and developing video games, augmented and virtual reality applications, as well as special effects in the movie industry. By providing AI assistants which can double as a 3D artist, we can enable new experiences centred on creating personalized, user-generated 3D content. Generative 3D assistants can also support many other applications, such as virtual product placement in user-generated videos. AI-powered 3D generation is also important for building infinitely large virtual worlds in the Metaverse.

3D generation has unique and difficult challenges not shared by other forms of media generation such as images and videos. Production-ready 3D content has exacting standards in terms of artistic quality, speed of generation, structural and topological quality of the 3D mesh, structure of the UV maps, and texture sharpness and resolution. Compared to other media, a unique challenge is that, while there exist billions of images and videos to learn from, the amount of 3D content viable for training is three to four order of magnitude smaller. Thus, 3D generation must also learn from images and videos which are *not* 3D and where 3D information must be *inferred* from partial, 2D observations.

Meta 3D Gen achieves high quality generation of 3D assets in under a minute. It supports Physically-Based Rendering (PBR) (Torrance and Sparrow, 1967), necessary for enabling relighting of generated assets in applications. When assessed by professional 3D artists, Meta 3D Gen significantly improves key metrics for production-quality 3D assets, particularly for complex textual prompts. The faithfulness to the textual prompts is better than other text-to-3D approaches, commercial or not, outperforming techniques that take from three minutes to an hour for generation. The quality of the generated 3D shapes and textures is better or at least on par with these competitors, using a scalable system that is significantly faster and more faithful.

Once the object is generated, its texture can be further edited and customised in 20 sec, with higher quality and at a fraction of the cost compared to alternatives. The same approach can be applied to texturing of artist-created 3D meshes without modifications.

The rest of this technical report describes the Meta 3D Gen pipeline as a whole, discussing how Meta 3D AssetGen and Meta 3D TextureGen are integrated, and conducts extensive evaluation studies against the most prominent industry baselines for text-to-3D generation.

Key capabilities. Meta 3D Gen is a two-stage method that combines two components, one for text-to-3D generation and one for text-to-texture generation, respectively. This integration results in higher-quality 3D generation for immersive content creation. In particular:

- **Stage I: 3D asset generation.** Given a text prompt provided by the user, Stage I creates an initial 3D asset using our Meta 3D AssetGen (Siddiqui et al., 2024) model (AssetGen for short). This step produces a 3D mesh with texture and PBR material maps. The inference time is approximately 30 sec.
- **Stage II, use case I: generative 3D texture refinement.** Given a 3D asset generated in Stage I and the initial text prompt used for generation, Stage II produces a higher-quality texture and PBR maps for this asset and the prompt. It utilizes our text-to-texture generator Meta 3D TextureGen (Bensadoun et al., 2024) (TextureGen for short). The inference time is approximately 20 sec.
- **Stage II, use case 2: generative 3D (re)texturing.** Given an untextured 3D mesh and a prompt describing its desired appearance, Stage II can also be used to generate a texture for this 3D asset from scratch (the mesh can be previously generated or artist-created). The inference time is approximately 20 sec.

Technical approach. By building on AssetGen and TextureGen, 3DGen effectively combines three highly-complementary representations of the 3D object: the view spaces (images of the object), the volumetric space (3D shape and appearance), and the UV space (texture). This process begins in AssetGen by generating several fairly consistent views of the object by utilizing a multi-view and multi-channel version of a text-to-image generator. Then, a reconstruction network in AssetGen extracts a first version of the 3D object in volumetric space. This is followed by mesh extraction, establishing the object's 3D shape and an initial version of its texture. Finally, a TextureGen's component regenerates the texture, utilizing a combination of view-space and UV-space generation, boosting the texture quality and resolution while retaining fidelity to the initial prompt.

Each stage of 3DGen builds on Meta's series of powerful text-to-image models Emu (Dai et al., 2023b). These are fine-tuned using renders of synthetic 3D data (from an internal dataset) to perform multi-view generation in view space as well as in UV space, resulting in better textures.

Performance. Integration of the two stages (AssetGen and TextureGen) and their different representations results in a combined model winning 68% of the times in evaluations. In addition to the strength that comes from this new combination, the individual components outperform the state of the art in their receptive functionalities. Specifically, AssetGen advances text-to-3D in several aspects: it supports physically-based rendering, which allows to relight the generated object, it obtains better 3D shapes via an improved representation (based on signed distance fields), and develops a new neural network that can effectively combine and fuse view-based information in a single texture. Likewise, TextureGen outperforms prior texture generator approaches by developing an end-to-end network that also operates in mixed view and UV spaces. Remarkably and differently to many state-of-the-art solutions, both AssetGen and TextureGen are feed-forward generators, and thus fast and efficient after deployment.

2 Method

We start by giving a high-level view of the two components of 3DGen, namely AssetGen (Stage I) and TextureGen (Stage II), and we refer the reader to the original papers for more details. We start from Stage II as it simplifies setting out the notation.

TextureGen (Bensadoun et al., 2024): core of Stage II. TextureGen is a text-to-texture generator for a given 3D shape. Namely, given a 3D object M and a textual prompt y, it generates a texture T for the object that is consistent with the prompt y. The object $M = (V, F, U)$ consists of a 3D mesh (V, F), where $V \in \mathbb{R}^{|V|\times 3}$ is a list of vertices and $F \in \{1, \dots, |V|\}^{|F|\times 3}$ is a list of triangular faces. The object comes with a map assigning each vertex $v_i \in V$ to a corresponding UV coordinate $u_i \in U \in [0, 1]^{|V|\times 2}$. The texture T is a 2D image of size $L \times L$ supported on $[0, 1]^2$. The texture has either three or five channels, in the first case representing the RGB shaded appearance of the object (with baked light) and in the second case the RGB albedo (base color), roughness and metalness, respectively.

TextureGen comprises several stages. In the first stage, a network $\Phi^{\text{tex}}_{\text{mv}}$ is trained to generate, from the prompt y and the object M, several views $I_1, \ldots, I_K$ of the object M. The generator is joint, in the sense that it samples the distribution $p(I_1, \ldots, I_K | y, M)$. In the second stage, the views $I_1, \ldots, I_K$ are first re-projected on corresponding texture images $T_1, \ldots, T_K$. Then, a second generator network $\Phi^{\text{tex}}_{\text{uv}}$ takes these and the prompt y to output a final texture T sampled from the conditional distribution $p(T|y, T_1, \ldots, T_K)$. This step reconciles the view-based textures, which may be slightly inconsistent, and completes the parts of the texture that are not visible in any of the views. Finally, a third optional network $\Phi^{\text{tex}}_{\text{super}}$ takes the texture T and performs super-resolution (up to 4K). Networks $\Phi^{\text{tex}}_{\text{mv}}$, $\Phi^{\text{tex}}_{\text{uv}}$ and $\Phi^{\text{tex}}_{\text{super}}$ are *diffusion-based* generators, trained on a large collection of 3D assets starting from a pre-trained image generator in Emu family (Dai et al., 2023b).

AssetGen (Siddiqui et al., 2024): core of Stage I. AssetGen is a text-to-3D object generator: given a textual prompt y, it samples both a 3D mesh M and a corresponding texture T from a distribution $p(M, T|y)$. AssetGen also operates stage-wise. First, a network $\Phi^{\text{obj}}_{\text{mv}}$ takes the prompt y and generates a set of views $I_1, \ldots, I_K$ of the object. This is similar to TextureGen's first stage $\Phi^{\text{tex}}_{\text{mv}}$, except that the views are *not* conditioned on the geometry of the object M, which is instead a target for generation. Then, given the views $I_1, \ldots, I_K$, a second network $\Phi^{\text{obj}}_{\text{rec}}$ generates a 3D mesh M and initial texture T using a large reconstruction neural network. Differently from network $\Phi^{\text{obj}}_{\text{mv}}$, which models a distribution via diffusion and is thus aleatoric, the network $\Phi^{\text{obj}}_{\text{rec}}$ is *deterministic*. Images $I_1, \ldots, I_K$ contain sufficient information for the model to reconstruct the 3D object without too much ambiguity. For PBR material reconstruction, this is achieved by tasking the image generator to output the shaded appearance of the object as well as its albedo (intrinsic image), which makes it easier to infer materials. Finally, AssetGen refines the texture T, by first obtaining auxiliary partial but sharp texture by re-projecting the input views $I_1, \ldots, I_K$ into textures $T_1, \ldots, T_K$. Then, a network $\Phi^{\text{obj}}_{\text{uv}}$ maps $T, T_1, \ldots, T_K$ (defined in UV space) to a fused and enhanced texture T^*.

Meta 3D Gen: integrated approach. Finally, we describe the combination of these two methods into a high-quality text-to-3D generator with retexturing capabilities. The idea is to utilize the texture generator in Stage II to significantly improve the quality of the texture obtained from the first-stage 3D object generator. The 3D object generator AssetGen does produce good quality textures, but has two limitations. First, it is not a model *specialized* for high-quality texture generation, but TextureGen is. Secondly, the texture generator TextureGen is conditioned on an existing 3D shape of the object, which makes it much easier to generate high-quality and highly-consistent multiple-views of the textured object. In other words, network $\Phi^{\text{tex}}_{\text{mv}}$ solves an easier task than network $\Phi^{\text{obj}}_{\text{mv}}$ (due to the additional geometric conditioning) and can thus generate better views, resulting in better high-resolution textures.

In principle, then, we could simply use network $\Phi^{\text{obj}}_{\text{mv}}$ from AssetGen to generate the 3D shape of the object and then network $\Phi^{\text{tex}}_{\text{mv}}$ and $\Phi^{\text{tex}}_{\text{uv}}$ to re-generate a better texture, with semantic consistency guaranteed by utilizing the same prompt y for conditioning the two steps. However, this approach does not work well by itself. The reason is that the texture fusion and enhancement network in TextureGen is trained on the basis of 'ground truth' UV maps by 3D artists; in contrast, the assets generated by AssetGen have automatically-extracted UV maps, that differ substantially from artist-created ones.

Fortunately, AssetGen comes with its own texture re-projection and fusion network $\Phi^{\text{obj}}_{\text{uv}}$ which is trained on the basis of automatically-extracted UV maps and can do a better job than network $\Phi^{\text{tex}}_{\text{uv}}$ on this task. Hence, our integrated solution is as follows:

- Given the prompt y, run networks $\Phi^{\text{obj}}_{\text{mv}}$ and $\Phi^{\text{obj}}_{\text{rec}}$ and mesh and UV extraction to obtain an initial mesh M and UV map U.
- Given the prompt y and the initial mesh M, run network $\Phi^{\text{tex}}_{\text{mv}}$ to generate a set of views $I_1, \ldots, I_K$ representing a new, better texture in view space. Using the UV map U, reproject these images into partial textures $T_1, \ldots, T_K$.
- Given the prompt y and the partial textures $T_1, \ldots, T_K$, run the network $\Phi^{\text{tex}}_{\text{uv}}$ from TextureGen to obtain a consolidated UV texture T.
- Given the partial textures $T_1, \ldots, T_K$ and the consolidated texture T, run network $\Phi^{\text{obj}}_{\text{uv}}$ from AssetGen to obtain the final texture T^*. This fixes any residual seams due to the non-human-like UV maps.

Method	Generation capabilities				Generation time	
	Mesh	Texture	PBR materials	Clean topology	Stage I only	Stages I+II
CSM Cube 2.0 (CSM, 2024)	✓	✓	✗	✗	15* min	1* h
Tripo3D (TripoAI, 2024)	✓	✓	✗	✗	30* sec	3* min
Rodin Gen-1 V0.5 (Deemos, 2024)	✓	✓	✓	✓	–	3*,† min
Meshy v3 (Meshy, 2024a)	✓	✓	✓	✗	1* min	10* min
Third-party T23D generator	✓	✓	✓	✗	10* sec	10* min
Meta 3D Gen	✓	✓	✓	✗	30 sec	**1 min**

* Averaged approximate estimates, as evaluated from corresponding public APIs.
† Depends on the complexity of geometry, can range from 2 to 30 min (in 7 % cases failed to converge).

Table 1 Overview of the industry baselines for the text-to-3D task. Comparison of generation capabilities and run times.

3 Experiments

We compare 3DGen against publicly-accessible industry solutions for the task of text-to-3D asset generation. We report extensive user studies to evaluate both the quality (for the baselines that are producing both textures and materials) and text prompt fidelity aspects of 3D generation, and provide qualitative results for both 3D generation and texturing.

3.1 Industry baselines

We compare performance of Meta 3D Gen with leading industry models for text-to-3D generation, which are currently accessible via web demos and public APIs. The summary of their capabilities, that are relevant to text-to-3D generation, and run times is provided in Table 1.

Common Sense Machines (CSM) Cube 2.0 (CSM, 2024). All results for comparisons were generated using the officially provided Cube API, with separate sequential calls for text-to-image and then image-to-3D generation (with the highest quality settings). Website: www.csm.ai.

Tripo3D (TripoAI, 2024). All results are generated using the official Tripo Platform, including both preview and refinement stages. Website: https://www.tripo3d.ai/app.

Rodin Gen-1 (0525) V0.5 (Deemos, 2024). The generations were obtained manually using the official web interface. The pipeline requires running several stages: text-to-image, image-to-shape, texture generation and material generation. To encourage prompt fidelity, we performed generations with the original text prompt at every stage. We also disabled the symmetry flag, as we found it to be hurtful for generating complex compositions. The rest of the settings were set to default. The method failed on 7 % of prompts (27 out of 404) during the meshing stage, likely due to the originally generated geometries being too complex. Website: hyperhuman.deemos.com/rodin.

Meshy v3 (Meshy, 2024a). The results were generated by the official API and with PBR materials, using the corresponding style setting. The rest of the settings were set by default. Website: www.meshy.ai.

Third-party text-to-3D (T23D) generator. We are providing additional quantitative comparisons with another industry-leading text-to-3D generator. The results were generated using the official web interface, including three stages: text-to-image, asset preview and asset refinement. Out of four image options proposed by the interface after the first stage, we always pick the top left one for consistency.

3.2 User studies

We conduct a series of user studies on *prompt fidelity* and *visual quality* of text-to-3D generations, produced by each of the models. Our pool of annotators consists of two groups: (1) representatives of a general population with no prior expertise in 3D, and (2) professional 3D artists, designers and game developers. We report aggregated results, as well as results obtained by the group with the strongest relevant expertise.

Evaluation benchmark. For evaluations, we use a set of deduplicated 404 text prompts that were initially introduced with DreamFusion (Poole et al., 2023). For our analysis, we split this set into a number of

Method	All prompts, per stage (↑)		Stage II, per prompt category (↑)		
	stage I	stage II	(A) objects	(B) characters	(A)+(B) compositions
CSM Cube 2.0 (CSM, 2024)	–	69.1 %	84.0 %	87.8 %	54.6 %
Tripo3D (TripoAI, 2024)	–	78.2 %	77.6 %	87.9 %	71.6 %
Rodin Gen-1 (0525) V0.5 (Deemos, 2024)	–	59.9 %	66.7 %	70.1 %	48.8 %
Meshy v3 (Meshy, 2024a)	60.6 %	76.0 %	97.2 %	83.2 %	63.5 %
Third-party T23D generator	73.5 %	79.7 %	95.0 %	**89.7 %**	67.9 %
Meta 3D Gen	**79.7 %**	**81.7 %**	**96.5 %**	84.1 %	**73.9 %**

Table 2 User studies: prompt fidelity. *Stage I* corresponds to the first-round text-to-3D generations, and *stage II* to the results of the final refinement. For simplicity, we consider Rodin Gen-1 to be a single-stage method.

Method	Q0: fidelity		Q1: quality		Q2: texture		Q3: geometry	
	Win (👍)	Loss (👎)	Win (👍)	Loss (👎)	Win (👍)	Loss (👎)	Win (👍)	Loss (👎)
All annotators								
Rodin Gen-1 (0525) V0.5 (Deemos, 2024)	67.6 %	32.4 %	66.2 %	33.8 %	70.9 %	29.1 %	60.3 %	39.7 %
Meshy v3 (Meshy, 2024a)	61.5 %	38.5 %	60.1 %	39.9 %	49.7 %	50.3 %	65.7 %	34.3 %
Third-party T23D generator	57.2 %	42.8 %	60.4 %	39.6 %	58.6 %	41.4 %	60.0 %	40.0 %
Professional 3D artists								
Rodin Gen-1 (0525) V0.5 (Deemos, 2024)	68.0 %	32.0 %	59.8 %	40.2 %	69.1 %	30.9 %	56.7 %	43.3 %
Meshy v3 (Meshy, 2024a)	60.0 %	40.0 %	65.3 %	34.7 %	53.7 %	46.3 %	66.3 %	33.7 %
Third-party T23D generator	59.1 %	40.9 %	61.3 %	38.78 %	60.2 %	39.8 %	60.2 %	39.8 %

Table 3 User studies: summary of A/B tests (for models producing textures and materials). The annotators were asked four questions: Q0 – "*which 3D asset is the better representation of the prompt?*", Q1 – "*which 3D asset has better quality overall?*", Q2 – "*which has better texture?*", Q4 – "*which has more correct geometry?*". Win and loss are measured for our method (Meta 3D Gen), with respect to each of the strongest baseline methods (stage II, where applicable).

categories, according to the described content complexity: objects (156), characters (106) and compositions of characters and objects (141). We report each model's performance on each of the categories separately, as well as the aggregated scores. In all studies, the annotators were shown fly-around 360° videos of rendered meshes. Text prompt fidelity, overall visual quality, as well as quality of geometries and textures are evaluated for every model either separately, or in randomized A/B tests.

Evaluation results. User studies results for text prompt fidelity are shown in Table 2. These were obtained independently for each model, by asking the annotators to decide whether or not the prompt correctly describes the generated content. 3DGen outperforms all considered industry baselines on this metric (in both stages), with the third-party text-to-3D (T23D) generators being the strongest competitor overall.

The A/B test user studies were designed to evaluate text prompt fidelity, overall visual quality, geometry visual quality, and texture details and artefacts for our model compared with baselines producing both textures and PBR materials. We do not perform exhaustive evaluations of our method versus models generating baked textures, due to significant perceptual differences between generations produces by the two classes of models at rendering time and due to practically limited usability of texture-only generations in real-world applications. The results are summarized in Table 3. We first report aggregated scores across all annotators, and then separately from the subset with a strong expertise in 3D. Overall, 3DGen performs stronger than the competitors according to most metrics, while also being significantly faster.

We observed that annotators with less experience in 3D tend to favour assets with sharper, more vivid, realistic, detailed textures and are not sensitive to presence of even significant texture and geometry artefacts. The professional 3D artists expressed a stronger preference for 3DGen generations across the whole range of metrics. We observed that their evaluations gave more weight to correctness of geometries and textures.

In Figure 3, we analyze performance rates for visual quality, geometry, texture details and presence of texture artefacts, as functions of the scene complexity as described by the text prompt. The plots show that, while some of the baselines perform on par for simple prompts, 3DGen starts outperforming them strongly as the prompt complexity increases from objects to characters and their compositions.

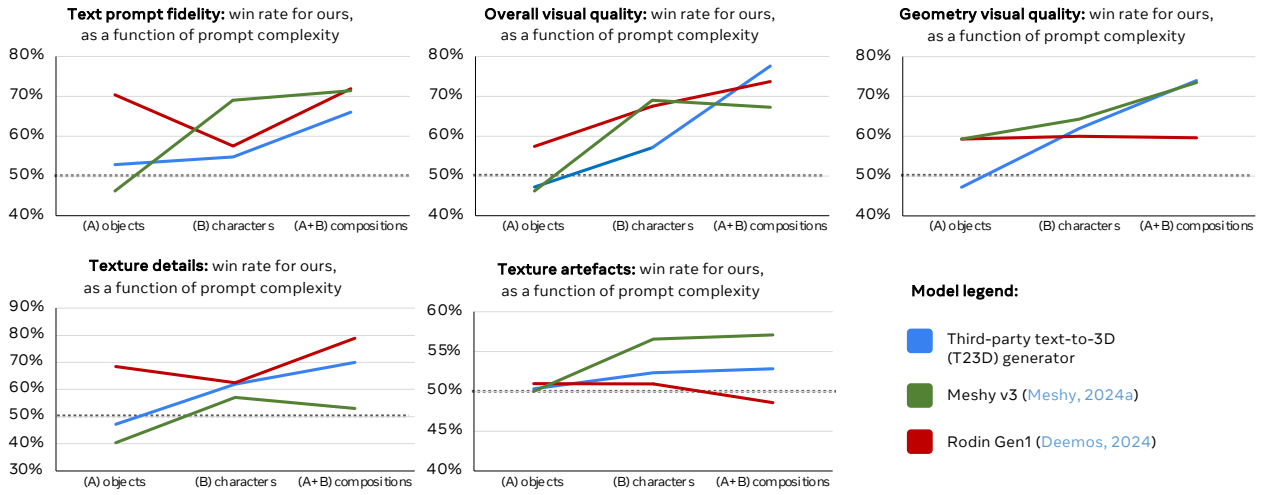

Figure 3 User studies: analysis of prompt fidelity, visual quality, geometry and texture parameters as functions of the scene complexity, as described by the text prompt (aggregated across all annotators). We report win rate for 3DGen against baselines and highlight the 50% threshold (dashed line) where our method is found to be as good as the baselines.

3.3 Qualitative results

Text-to-3D. Visual comparisons of Stage I and Stage II generations are given in Fig. 4. The latter tend to have higher visual aesthetics, appear more realistic and have higher-frequency details. Our annotators prefer generations from Stage II in 68 % of the cases. More qualitative examples of text-to-3D generations produced by 3DGen Stage II are shown in Figure 5 (diverse classes) and Figure 11 (within one object class).

Next, we visually compare performance of our model with industry baselines on the same scenes (Figure 7), additionally on more challenging prompts (Figure 6) and in terms of most common failure cases of both our method and the baselines (Figure 8). Overall, these qualitative observations confirm that, while the alternative methods do well on simple objects, generation of more complex compositions and scenes becomes a bigger challenge for them. There is also a clear trade-off between generating high-frequency details in textures vs exposing visual artefacts. Meshy v3 Meshy (2024a), in particular, has a visually appealing style with highly detailed generations (which are often appreciated in user studies, in particular among non-professionals), but often suffers from Janus effects, inpainting problems and seams in texture maps. Geometry-wise, Rodin Gen1 Deemos (2024) produces quad meshes with correct topologies, but at cost of compromising prompt fidelity and sometimes failing to produce results for complex prompts altogether.

Figure 4 Visual comparison of text-to-3D generations obtained after Meta 3D Gen's Stage I (left) and Stage II (right). In our A/B user studies, the Stage II generations had a win rate of **68 %** in texture quality over the first-stage generations.

Figure 5 Qualitative results for text-to-3D generation. We show quality and diversity of text-to-3D generations produced by 3DGen, across different scene categories (single objects and compositions).

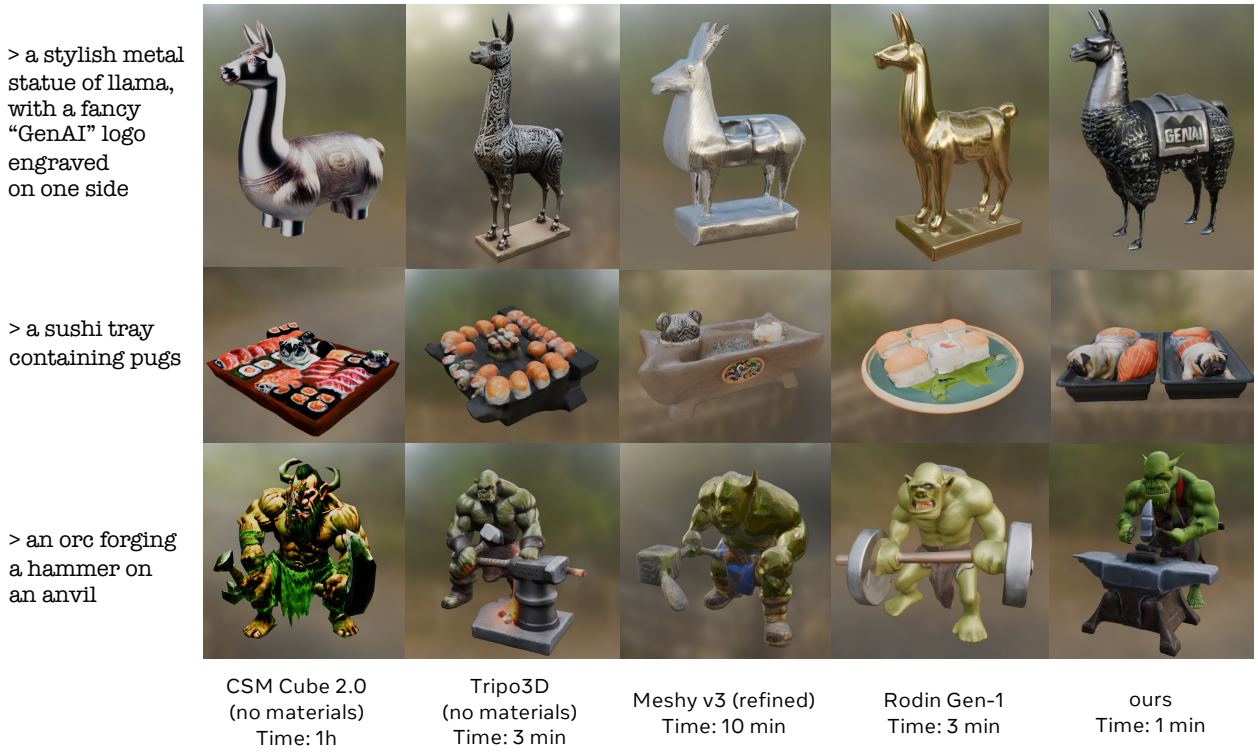

Figure 6 Qualitative comparison of text prompt fidelity with all industry baselines (on challenging prompts).

Figure 7 Qualitative comparison with industry baselines producing textures with PBR materials (on the same set of prompts).

Figure 8 Examples of typical failure modes of different methods.

Figure 9 (Re)texturing results for generated shapes. Examples of meshes produced by Stage I of the pipeline and textured in Stage II with various text prompts, different from the original ones.

Figure 10 (Re)texturing results for generated shapes. Examples of themed scenes produced by Stage II of Meta 3D Gen by augmenting object-specific texturing prompts with the new style information in a coherent manner.

3D asset (re)texturing. Figure 9 shows qualitative results for the task of asset retexturing: 3D meshes, generated in Stage I, are then passed to Stage II with textual prompts that are *different* than the original ones. This process allows us to create new assets with the same base shapes, but different appearances. The results show that in addition to implementing semantic edits and performing both global and localized modifications, 3DGen can successfully imitate different materials and artistic styles. Figure 10 shows how one can retexture whole scenes in a coherent manner, by augmenting object-level prompts used for retexturing with the style information. As discussed in Bensadoun et al. (2024), Stage II can be applied for retexturing of both generated and artist-created 3D assets with no significant changes to the pipeline.

4 Related Work

There is ample literature in both text-to-3D and text-to-texture. We point the readers to (Siddiqui et al., 2024; Bensadoun et al., 2024) for a more extensive discussion and provide here key pointers.

Text-to-3D. Some methods (Nichol et al., 2022; Jun and Nichol, 2023; Gupta et al., 2023; Yariv et al., 2023; Xu et al., 2024c) train 3D generators on 3D datasets, but the limited availability of such data is an obstacle to generalization. Hence, most recent approaches start from image or video-based generators trained on billions of data samples (Shue et al., 2022; Mercier et al., 2024).

Many approaches (Lin et al., 2022; Qian et al., 2023; Lin et al., 2022; Tang et al., 2023a; Yi et al., 2023; Chen et al., 2023e; Wang et al., 2023a,c; Zhu and Zhuang, 2023; Huang et al., 2023; Qian et al., 2023; Tang et al., 2023b; Yu et al., 2023a; Sun et al., 2023) are based on distillation (Poole et al., 2023). However, distillation is slow (Lorraine et al., 2023; Xie et al., 2024) and may lead to artifacts such as the Janus effect (Shi et al., 2024). Follow-up works have thus built on multi-view-aware image generators (Liu et al., 2023c; Shi et al., 2023; Jiang et al., 2023a; Chen et al., 2023d; Qian et al., 2023; Shi et al., 2024; Weng et al., 2023; Wang and Shi, 2024; Kim et al., 2024; Zhou et al., 2024).

More recent approaches focus on generating several consistent views (Liu et al., 2023b; Long et al., 2023; Liu et al., 2023d; Yang et al., 2023b,a; Chan et al., 2023; Tang et al., 2024c; Höllein et al., 2024; Gao et al., 2024; Melas-Kyriazi et al., 2024; Chen et al., 2024) from which direct 3D reconstruction is possible. However, these methods are susceptible to limitations in the multi-view consistency of the generated mages. Other approaches thus learn few-view robust reconstructors (Li et al., 2024; Hong et al., 2024; Liu et al., 2023a).

Multi-view to 3D. Many generators thus build on few-view 3D reconstruction. Methods like NeRF (Mildenhall et al., 2020) cast this as analysis by synthesis, optimizing a differentiable rendering loss. These approaches can use a variety of 3D representations, from meshes to 3D gaussians (Gao et al., 2020; Zhang et al., 2021a; Goel et al., 2022; Munkberg et al., 2022; Monnier et al., 2023; Kerbl et al., 2023; Guédon and Lepetit, 2023; Niemeyer et al., 2020; Mildenhall et al., 2020; Müller et al., 2022; Yariv et al., 2020; Oechsle et al., 2021; Yariv et al., 2021; Wang et al., 2021; Darmon et al., 2022; Fu et al., 2022).

When only a small number of views are available, authors train reconstruction models to acquire the necessary priors (Choy et al., 2016; Kanazawa et al., 2018; Mescheder et al., 2019; Liu et al., 2019; Wu et al., 2020; Monnier et al., 2022; Wang et al., 2023b; Hong et al., 2024; Vaswani et al., 2017; Chan et al., 2022; Chen et al., 2022; Xu et al., 2024a; Wei et al., 2024; Zou et al., 2023; Xu et al., 2024b; Tang et al., 2024a; Zhang et al., 2024; Wang et al., 2024; Wei et al., 2024; Tochilkin et al., 2024; Junlin Han, 2024).

PBR modelling. Several authors have considered reconstruction methods with PBR support too (Boss et al., 2021b,a; Xiuming et al., 2021; Zhang et al., 2021b; Munkberg et al., 2022; Hasselgren et al., 2022; Jiang et al., 2023b; Liang et al., 2023). This is also the case for 3D generators (Chen et al., 2023c; Qiu et al., 2023; Liu et al., 2023f; Xu et al., 2023; Poole et al., 2023).

Texture generation. Several works have tackled specifically the task of generating textures for 3D objects as well. For instance Mohammad Khalid et al. (2022); Michel et al. (2022) use guidance from CLIP (Radford et al., 2021) and differentiable rendering to match the texture to the textual prompt. Chen et al. (2023b); Metzer et al. (2023); Youwang et al. (2023) use SDS loss optimization Poole et al. (2022) and Siddiqui et al. (2022); Bokhovkin et al. (2023) use a GAN-like approach analogous to (Karras et al., 2019). Other methods use diffusion in UV space (Liu et al., 2024; Cheskidova et al., 2023), but focus on human character texturing. Yu et al. (2023b) uses point-cloud diffusion to generate a texture.

Richardson et al. (2023); Chen et al. (2023a); Tang et al. (2024b); Zeng (2023) combine texture inpainting with depth-conditioned image diffusion, but generate one image at a time, which is slow and prone to some artifacts. Liu et al. (2023e); Cao et al. (2023) improves consistency by alternating diffusion iterations and re-projections to combine them. Deng et al. (2024) generate four textured views jointly, but uses slow SDS optimization to extract the texture. Meshy (Meshy, 2024b) also provide a texture generator module, but its details remain proprietary.

Image generators. Our generators are based on image generators, which have been studied extensively starting from GANs (Goodfellow et al., 2014). Recent works use transformer architectures (Ramesh et al., 2021; Ding et al., 2021; Gafni et al., 2022; Yu et al., 2022; Chang et al., 2023). Several more operate in pixel space or

latent space using diffusion (Ho et al., 2020; Balaji et al., 2022; Saharia et al., 2022; Ramesh et al., 2022; Rombach et al., 2022; Podell et al., 2023). We build on the Emu class of image generators (Dai et al., 2023a).

5 Conclusions

We have introduced 3DGen, a unified pipeline integrating Meta's foundation generative models for text-to-3D generation with texture editing and material generation capabilities, AssetGen and TextureGen, respectively. By combining their strengths, 3DGen achieves very high-quality 3D object synthesis from textual prompts in less than a minute. When assessed by professional 3D artists, the output of 3DGen is preferred a majority of time compared to industry alternatives, particularly for complex prompts, while being from 3× to 60× faster.

While our current integration of AssetGen and TextureGen straightforward, it sets out a very promising research research direction that builds on two thrusts: (1) generation in view space and UV space, and (2) end-to-end iteration over texture and shape generation.

6 Acknowledgements

We are grateful for the instrumental support of the multiple collaborators at Meta who helped us in this work: Ali Thabet, Albert Pumarola, Markos Georgopoulos, Jonas Kohler, Uriel Singer, Lior Yariv, Amit Zohar, Yaron Lipman, Itai Gat, Ishan Misra, Mannat Singh, Zijian He, Jialiang Wang, Roshan Sumbaly.

We thank Manohar Paluri and Ahmad Al-Dahle for their support of this project.

References

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhng, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to0image diffusion models with an ensemble of expert denoisers. In *arXiv preprint arXiv:2211.01324*, 2022.

Figure 11 Quality and diversity of generations produced by 3DGen, for a *single* object class ("llama").

Raphael Bensadoun, Yanir Kleiman, Idan Azuri, Omri Harosh, Andrea Vedaldi, Natalia Neverova, and Oran Gafni. Meta 3D Texture Gen: Fast and consistent texture generation for 3D objects. *arXiv preprint*, 2024.

Alexey Bokhovkin, Shubham Tulsiani, and Angela Dai. Mesh2tex: Generating mesh textures from image queries. *arXiv preprint arXiv:2304.05868*, 2023.

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. NeRD: Neural Reflectance Decomposition from Image Collections. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021a.

Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P. A. Lensch. Neural-PIL: Neural Pre-Integrated Lighting for Reflectance Decomposition. *arXiv preprint*, 2021b.

Tianshi Cao, Karsten Kreis, Sanja Fidler, Nicholas Sharp, and Kangxue Yin. Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4169–4181, 2023.

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, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proc. CVPR*, 2022.

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. In *Proc. ICCV*, 2023.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial radiance fields. In *arXiv*, 2022.

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023a.

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*, pages 22246–22256, 2023b.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv.cs*, abs/2303.13873, 2023c.

Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-Zero123: One image to highly consistent 3D with self-prompted nearby views. *arXiv.cs*, abs/2312.04424, 2023d.

Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3D using Gaussian splatting. *arXiv*, (2309.16585), 2023e.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3D: Video diffusion models are effective 3D generators. *arXiv*, 2403.06738, 2024.

Evgeniia Cheskidova, Aleksandr Arganaidi, Daniel-Ionut Rancea, and Olaf Haag. Geometry aware texturing. In *SIGGRAPH Asia 2023 Posters*, SA '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703133. doi: 10.1145/3610542.3626152. URL https://doi.org/10.1145/3610542.3626152.

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 *Proc. ECCV*, 2016.

CSM. CSM text-to-3D cube 2.0, 2024. URL https://www.csm.ai.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023a.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam S. Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh

Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack. *CoRR*, abs/2309.15807, 2023b.

François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proc. CVPR*, 2022.

Deemos. Rodin text-to-3D gen-1 (0525) v0.5, 2024. URL https://hyperhuman.deemos.com/rodin.

Kangle Deng, Timothy Omernick, Alexander Weiss, Deva Ramanan, Jun-Yan Zhu, Tinghui Zhou, and Maneesh Agrawala. Flashtex: Fast relightable mesh texturing with lightcontrolnet. *arXiv preprint arXiv:2402.13251*, 2024.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao. Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. In *NeurIPS*, 2022.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.

Jun Gao, Wenzheng Chen, Tommy Xiang, Clement Fuji Tsang, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Learning deformable tetrahedral meshes for 3D reconstruction. In *Proc. NeurIPS*, 2020.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *arXiv.cs*, 2024.

Shubham Goel, Georgia Gkioxari, and Jitendra Malik. Differentiable Stereopsis: Meshes from multiple views using differentiable rendering. In *CVPR*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Antoine Guédon and Vincent Lepetit. SuGaR: Surface-aligned Gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. *arXiv.cs*, abs/2311.12775, 2023.

Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oguz. 3DGen: Triplane latent diffusion for textured mesh generation. *corr*, abs/2303.05371, 2023.

Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *arXiv preprint*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. ViewDiff: 3D-Consistent Image Generation with Text-to-Image Models. *arXiv preprint*, 2024.

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. In *Proc. ICLR*, 2024.

Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3D content creation. *CoRR*, abs/2306.12422, 2023.

Yifan Jiang, Hao Tang, Jen-Hao Rick Chang, Liangchen Song, Zhangyang Wang, and Liangliang Cao. Efficient-3Dim: Learning a generalizable single-image novel-view synthesizer in one day. *arXiv*, 2023a.

Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. GaussianShader: 3D Gaussian splatting with shading functions for reflective surfaces. *arXiv.cs*, abs/2311.17977, 2023b.

Heewoo Jun and Alex Nichol. Shape-E: Generating conditional 3D implicit functions. *arXiv*, 2023.

Philip Torr Junlin Han, Filippos Kokkinos. Vfusion3d: Learning scalable 3d generative models from video diffusion models. *arXiv preprint*, 2024.

Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proc. ECCV*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *Proc. SIGGRAPH*, 42(4), 2023.

Seungwook Kim, Yichun Shi, Kejie Li, Minsu Cho, and Peng Wang. Multi-view image prompted multi-view diffusion for improved 3D generation. *arXiv*, 2404.17419, 2024.

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. *Proc. ICLR*, 2024.

Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. GS-IR: 3D Gaussian splatting for inverse rendering. *arXiv.cs*, abs/2311.16473, 2023.

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. *arXiv.cs*, abs/2211.10440, 2022.

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3D objects with consistent multi-view generation and 3D diffusion. *arXiv.cs*, abs/2311.07885, 2023a.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. In *Proc. NeurIPS*, 2023b.

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 *Proc. ICCV*, 2023c.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3D reasoning. *arXiv.cs*, abs/1904.01786, 2019.

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*, (2309.03453), 2023d.

Yufei Liu, Junwei Zhu, Junshu Tang, Shijie Zhang, Jiangning Zhang, Weijian Cao, Chengjie Wang, Yunsheng Wu, and Dongjin Huang. Texdreamer: Towards zero-shot high-fidelity 3d human texture generation. *arXiv preprint arXiv:2403.12906*, 2024.

Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. *arXiv preprint arXiv:2311.12891*, 2023e.

Zexiang Liu, Yangguang Li, Youtian Lin, Xin Yu, Sida Peng, Yan-Pei Cao, Xiaojuan Qi, Xiaoshui Huang, Ding Liang, and Wanli Ouyang. UniDream: Unifying Diffusion Priors for Relightable Text-to-3D Generation. *arXiv preprint*, 2023f.

Xiaoxiao Long, Yuanchen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3D: Single image to 3D using cross-domain diffusion. *arXiv.cs*, abs/2310.15008, 2023.

Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. ATT3D: amortized text-to-3D object synthesis. In *Proc. ICCV*, 2023.

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. IM-3D: Iterative multiview diffusion and reconstruction for high-quality 3D generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Antoine Mercier, Ramin Nakhli, Mahesh Reddy, and Rajeev Yasarla. HexaGen3D: Stablediffusion is just one step away from fast and diverse text-to-3D generation. *arXiv*, 2024.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *CVPR*, 2019.

Meshy. Meshy text-to-3D v3.0, 2024a. URL https://www.meshy.ai.

Meshy. Meshy 3.0. https://docs.meshy.ai/, 2024b. Accessed: 2024-05-01.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023.

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022.

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 *Proc. ECCV*, 2020.

Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022.

Tom Monnier, Matthew Fisher, Alexei A. Efros, and Mathieu Aubry. Share With Thy Neighbors: Single-View Reconstruction by Cross-Instance Consistency. In *ECCV*, 2022.

Tom Monnier, Jake Austin, Angjoo Kanazawa, Alexei A. Efros, and Mathieu Aubry. Differentiable blocks world: Qualitative 3d decomposition by rendering primitives. *arXiv*, abs/2307.05473, 2023.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *Proc. SIGGRAPH*, 2022.

Jacob Munkberg, Wenzheng Chen, Jon Hasselgren, Alex Evans, Tianchang Shen, Thomas Muller, Jun Gao, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *CVPR*, 2022.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A system for generating 3D point clouds from complex prompts. *arXiv.cs*, abs/2212.08751, 2022.

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision. In *CVPR*, 2020.

Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction. *arXiv.cs*, abs/2104.10078, 2021.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Romach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *arXiv preprint arXiv:2307.01952*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *Proc. ICLR*, 2023.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv.cs*, abs/2306.17843, 2023.

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.cs*, abs/2311.16918, 2023.

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. *Zero-shot text-to-image generation (ICML spotlight)*, 2021. URL https://icml.cc/virtual/2021/spotlight/9430.

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

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv.cs*, abs/2310.15110, 2023.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *Proc. ICLR*, 2024.

J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3D neural field generation using triplane diffusion. *arXiv.cs*, abs/2211.16677, 2022.

Yawar Siddiqui, Justus Thies, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Texturify: Generating textures on 3d shape surfaces. In *European Conference on Computer Vision*, pages 72–88. Springer, 2022.

Yawar Siddiqui, Filippos Kokkinos, Tom Monnier, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. Meta 3D Asset Gen: Text-to-mesh generation with high-quality geometry, texture, and PBR materials. *arXiv preprint*, 2024.

Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. DreamCraft3D: Hierarchical 3D generation with bootstrapped diffusion prior. *arXiv.cs*, abs/2310.16818, 2023.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative gaussian splatting for efficient 3D content creation. *arXiv*, (2309.16653), 2023a.

Jiaxiang 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*, 2402.05054, 2024a.

Jiaxiang Tang, Ruijie Lu, Xiaokang Chen, Xiang Wen, Gang Zeng, and Ziwei Liu. Intex: Interactive text-to-texture synthesis via unified depth-aware inpainting. *arXiv preprint arXiv:2403.11878*, 2024b.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-It-3D: High-fidelity 3d creation from A single image with diffusion prior. *arXiv.cs*, abs/2303.14184, 2023b.

Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. MVDiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv*, 2402.12712, 2024c.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. TripoSR: fast 3D object reconstruction from a single image. *arXiv*, 2403.02151, 2024.

K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *J. Opt. Soc. Am.*, 57 (9), 1967.

TripoAI. Tripo3D text-to-3D, 2024. URL https://www.tripo3d.ai.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *CVPR*, 2023a.

Peng Wang and Yichun Shi. ImageDream: Image-prompt multi-view diffusion for 3D generation. In *Proc. ICLR*, 2024.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv.cs*, abs/2106.10689, 2021.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3D digital avatars using diffusion. In *Proc. CVPR*, 2023b.

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. *arXiv.cs*, abs/2305.16213, 2023c.

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*, (2403.05034), 2024.

Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. MeshLRM: large reconstruction model for high-quality mesh. *arXiv*, 2404.12385, 2024.

Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, C. L. Philip Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3D object synthesis. *arXiv*, 2023.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Kevin Xie, Jonathan Lorraine, Tianshi Cao, Jun Gao, James Lucas, Antonio Torralba, Sanja Fidler, and Xiaohui Zeng. LATTE3D: Large-scale amortized text-to-enhanced3D synthesis. In *arXiv*, 2024.

Zhang Xiuming, Srinivasan Pratul P., Deng Boyang, Debevec Paul, Freeman William T., and Barron Jonathan T. NeRFactor: neural factorization of shape and reflectance under an unknown illumination. In *Proc. SIGGRAPH*, 2021.

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*, 2404.07191, 2024a.

Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. MATLABER: Material-Aware Text-to-3D via LAtent BRDF auto-EncodeR. *arXiv preprint*, 2023.

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. GRM: Large gaussian reconstruction model for efficient 3D reconstruction and generation. *arXiv*, 2403.14621, 2024b.

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. In *Proc. ICLR*, 2024c.

Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. ConsistNet: Enforcing 3D consistency for multi-view images diffusion. *arXiv.cs*, abs/2310.10343, 2023a.

Yunhan Yang, Yukun Huang, Xiaoyang Wu, Yuan-Chen Guo, Song-Hai Zhang, Hengshuang Zhao, Tong He, and Xihui Liu. DreamComposer: Controllable 3D object generation via multi-view conditions. *arXiv.cs*, abs/2312.03611, 2023b.

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Proc. NeurIPS*, 2020.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *arXiv.cs*, abs/2106.12052, 2021.

Lior Yariv, Omri Puny, Natalia Neverova, Oran Gafni, and Yaron Lipman. Mosaic-SDF for 3D generative models. *arXiv.cs*, abs/2312.09222, 2023.

Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. GaussianDreamer: Fast generation from text to 3D gaussian splatting with point cloud priors. *arXiv.cs*, abs/2310.08529, 2023.

Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. *arXiv preprint arXiv:2312.11360*, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wangt, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Karagol Burcu Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Wangbo Yu, Li Yuan, Yan-Pei Cao, Xiangjun Gao, Xiaoyu Li, Long Quan, Ying Shan, and Yonghong Tian. HiFi-123: Towards high-fidelity one image to 3D content generation. *arXiv.cs*, abs/2310.06744, 2023a.

Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Zhengzhe Liu, and Xiaojuan Qi. Texture generation on 3d meshes with point-uv diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4206–4216, 2023b.

Xianfang Zeng. Paint3d: Paint anything 3d with lighting-less texture diffusion models. *arXiv preprint arXiv:2312.13913*, 2023.

Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild. In *NeurIPS*, 2021a.

Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting. *arXiv preprint*, 2021b.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: large reconstruction model for 3D Gaussian splatting. *arXiv*, 2404.19702, 2024.

Xiaoyu Zhou, Xingjian Ran, Yajiao Xiong, Jinlin He, Zhiwei Lin, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. GALA3D: Towards text-to-3D complex scene generation via layout-guided generative gaussian splatting. *arXiv.cs*, abs/2402.07207, 2024.

Junzhe Zhu and Peiye Zhuang. HiFA: High-fidelity text-to-3D with advanced diffusion guidance. *CoRR*, abs/2305.18766, 2023.

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.cs*, abs/2312.09147, 2023.