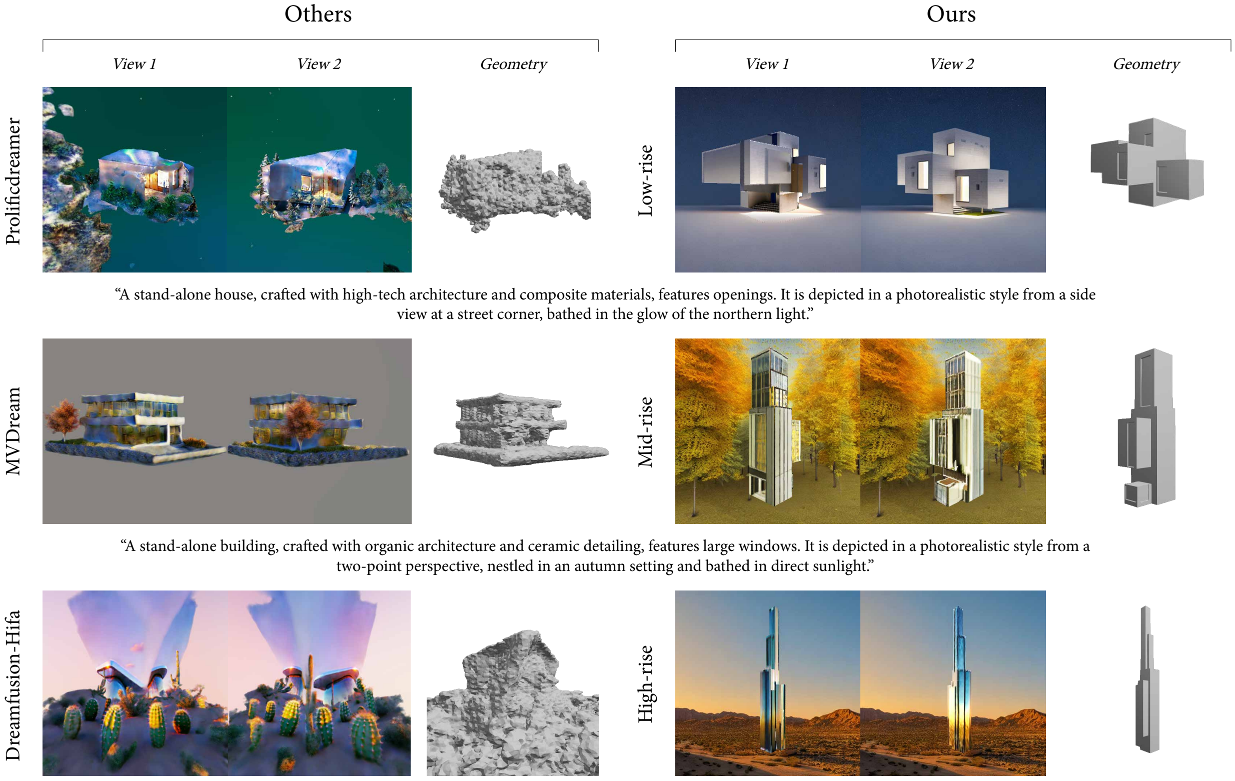
3D Synthesis for Architectural Design

I-Ting Tsai
Cornell University
it84@cornell.edu

Bharath Hariharan
Cornell University
bharathh@cs.cornell.edu



"A stand-alone house, crafted with high-tech architecture and composite materials, features openings. It is depicted in a photorealistic style from a side view at a street corner, bathed in the glow of the northern light."

"A stand-alone building, crafted with organic architecture and ceramic detailing, features large windows. It is depicted in a photorealistic style from a two-point perspective, nestled in an autumn setting and bathed in direct sunlight."

"A futuristic building with a prominent large glass window, set in a desert landscape surrounded by a variety of cacti and desert flora, bathed in sunset."

Figure 1. Existing methods for 3D synthesis blend background into foreground textures and create overly complex geometries (left) [48, 53, 61]. In contrast, our method (right) creates varied, consistently styled designs with clean geometries, facilitating idea communication and rapid iteration for the early design stage.

Abstract

We introduce a 3D synthesis method for architectural design to allow for the efficient generation of diverse and realistic building designs. In spite of advances in 3D synthesis, current off-the-shelf 3D synthesis techniques are inappropriate for architectural design: they are trained primarily on isolated objects, have limited diversity, blend building facades with background and produce overly complex geometry that is difficult to edit or manipulate, a major issue in an iterative design process. We propose an alternative pipeline that integrates auto-generated coarse models with segment-wise texture inpainting and semantics-based editing, resulting in diverse, style-consistent, and shape-precise designs. We show through qualitative and quantitative experiments that our pipeline generates more diverse, visually appealing architectures with clean geometries without the need for any extensive training. Project page: https://itingtsai.github.io/syn_arch_2025/

1. Introduction

With the advent of diffusion models, we have seen rapid advances in 3D synthesis [1, 9, 18–20, 28–34, 39, 40, 45, 48, 49, 51, 53, 61], paving the way for a variety of downstream creative applications. However, much of the focus has been on entertainment applications and virtual reality. Here, we focus on untapped potential in a novel domain: architectural design.

In architectural design, AI in general can be extremely useful in uncovering patterns and generating innovative solutions that might not be apparent to human designers [26]. However, past work has mostly focused on the generation of 2D layouts [4, 37, 42]. With the advent of more capable image diffusion models, there is interest in using these new models in other aspects of the design process [16]. But buildings are 3D artifacts, not 2D images. How these 2D generative models can produce useful 3D designs in an architectural design pipeline remains an open research question.

Unfortunately, current 3D synthesis techniques are unsuitable for architectural design. For one, they are trained on object datasets and so cannot produce buildings [11]. The alternative is to use 2D generative models trained on more general datasets, and leverage *Score Distillation Sampling* (SDS) [39] and neural fields to synthesize 3D shapes [9, 28, 29, 34, 48, 51, 53, 61]. However, this line of approaches suffers from three fundamental limitations. First, SDS is prone to mode collapse and as such limits diversity, undermining the design process (Fig. 1 left) [52]. Second, the resulting generations often feature textures inconsistent with the underlying building structure, such as background textures painted on building facades (Fig. 1 left). Third, neural field optimization results in overly complex geometries with bumps and other artifacts (Fig. 1 left). These unnecessary details not only complicate edits but also get in the way of communication, decision-making and rapid iteration, which are vital in the early stages of design.

In this paper, we address these limitations and introduce the first 3D synthesis method specifically tailored for architectural design. Our approach begins by first creating a coarse 3D model (called a "massing model" in architecture [3]) by aggregating a sequence of randomly generated primitives. This approach leads to diverse generations while incorporating domain knowledge about building designs.

We then paint textures on the 3D massing model using 2D generative models and user-defined prompts. We propose a novel facade-by-facade approach that ensures that the generated facade matches the building structure and is stylistically consistent. Specifically, we use ControlNet [58] to produce the first facades conditioned on the depth map, then generate the other facades one-by-one using an inpainting model [43] with a novel visual prompt to ensure stylistic consistency. This approach produces a more realistic facade than SDS-based approaches.

Finally, we propose a new semantically guided approach to refine the original coarse 3D geometry based on the generated facades. Concretely, we use open-vocabulary object detection models [36] to add facade details to the underlying coarse models only where necessary in specific semantic regions (e.g., windows and doors). Unlike SDS-generated 3D models with unnecessary complexity, our approach enhances facades selectively, maintaining clean geometry and ensuring easier downstream edits and design development.

In summary, our contributions are as follows:

- We introduce the problem of 3D synthesis for architectural design, and demonstrate that existing 3D synthesis approaches are not up to the task.
- We propose a novel pipeline that processes geometry, texture, and detail elements separately resulting in superior architectural design outcomes compared to a single-step process using SDS optimization. This separation allows for more precise control and higher quality results in each aspect of the design.
- We demonstrate through both qualitative and quantitative results that our pipeline produces diverse, innovative and useful designs that can be easily edited or manipulated by the designer.

2. Related Work

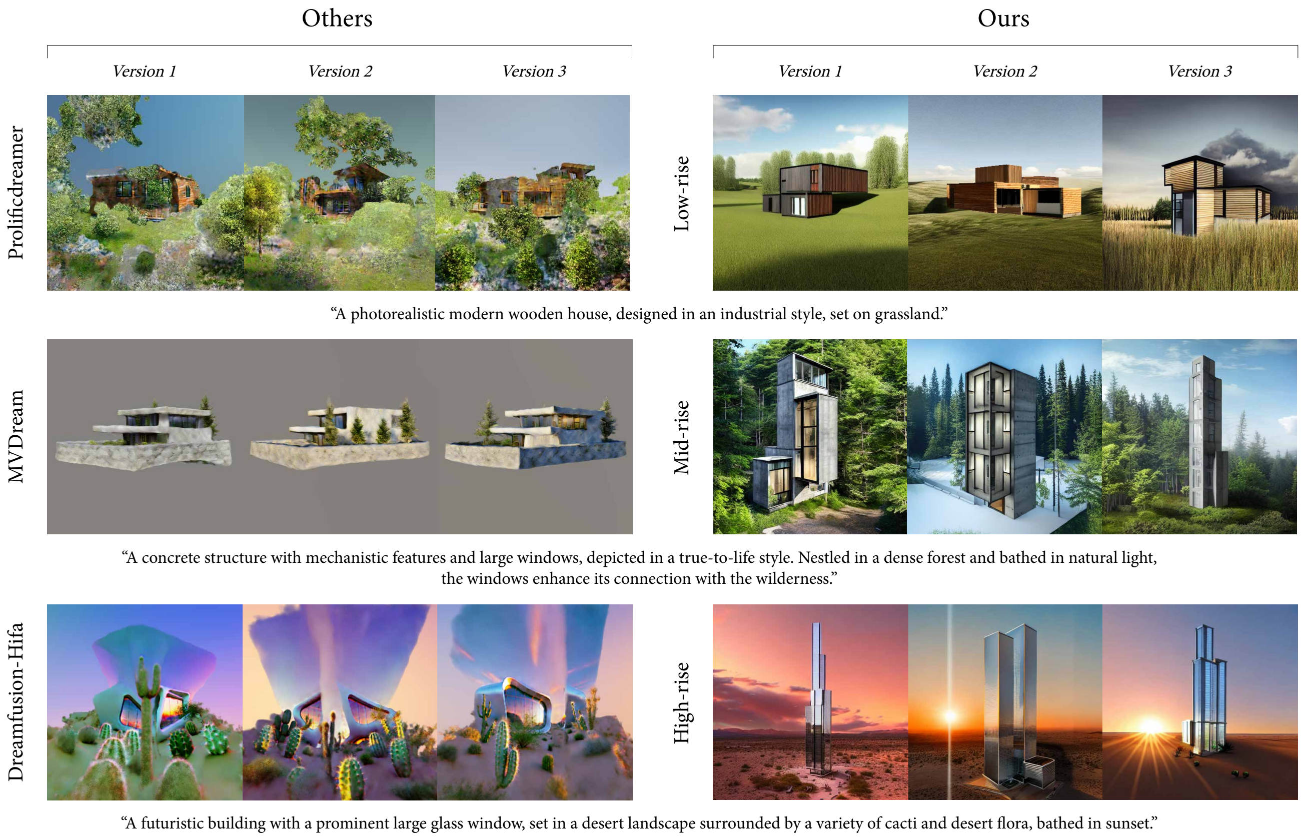
2.1. Generating Textured 3D Artifacts

The dominant approach to generating fully textured 3D shapes is to use *Score Distillation Sampling* (SDS) [39], where a 3D neural field is optimized to ensure that each rendered viewpoint is "high probability" according to a 2D diffusion model. More advanced 3D synthesis models have further refined SDS, leveraging both 2D [20, 28, 29, 33, 53] and 3D [18, 19, 30–32, 40, 45, 48, 49] priors to enhance text-guided and image-guided models. SDS-based generation can also leverage diffusion models that are conditioned on one view, as developed in novel view synthesis techniques [22, 23, 44, 60].

We find that SDS-based techniques are often insufficient for architectural design needs: Generated buildings lack diversity, appear unrealistic due to background blending with foreground, and produce overly complex geometry with unnecessary facade bumps, complicating downstream edits (Figs. 1 and 2). Our pipeline also uses 2D diffusion models, but eschews the SDS-based optimization in favor of a facade-by-facade texture painting approach combined with a separate process for generating coarse geometries.

2.2. 3D Shape Synthesis

Some methods focus solely on synthesizing high-resolution 3D shapes without textures [10, 12, 27, 46, 50,



"A photorealistic modern wooden house, designed in an industrial style, set on grassland."

"A concrete structure with mechanistic features and large windows, depicted in a true-to-life style. Nestled in a dense forest and bathed in natural light, the windows enhance its connection with the wilderness."

"A futuristic building with a prominent large glass window, set in a desert landscape surrounded by a variety of cacti and desert flora, bathed in sunset."

Figure 2. **Diverse Designs.** In contrast to prior work [48, 53, 61] (left) that produces repetitive design, our approach (right) generates diverse and varied building designs.

54, 59]. While these approaches produce complex geometries, they lack the capability to visualize design ideas with textures, limiting their effectiveness in enhancing design communication and inspiration. Our approach addresses this limitation by producing both diverse geometries and detailed textures, making it more suitable for visualizing and communicating design ideas and aiding in the decision-making process. Additionally, 3D generative models must be trained on 3D data, which is difficult to acquire for buildings. Our work avoids this limitation by leveraging 2D diffusion models instead.

2.3. Texture Synthesis

Another line of research focuses solely on texture synthesis, utilizing 2D diffusion models to texture input 3D objects [8, 14, 55–57]. While these methods enhance visual fidelity, they typically apply to isolated objects with well-defined geometries. Furthermore, these approaches usually require the complete, complex geometry as input. In contrast, our approach synthesizes both texture and geometry for the new space of buildings.

2.4. Shape-guided 3D Synthesis

Closely related to our work is the line of work on shape-guided synthesis, which starts with a basic 3D structure and uses both text and geometry to influence the final result. This method allows users to input custom geometry and offers editable controls through text-based guidance.

The sketch-shape method in Latent-NeRF [34] employs coarse geometry and text inputs to generate 3D geometries, using NeRF [35] to define surface details based on point-to-surface distances. Fantasia3D [9] employs DMTet [47] for geometry optimization and models appearance with a BRDF, with SDS supervising both processes. These approaches are based on optimization and thus inherit the limitations of SDS-based approaches, including overly complex geometries and unrealistic textures (Fig. 6). Furthermore, the use of global optimization precludes the ability to modify specific regions or elements. In contrast, we propose a pipeline that avoids SDS optimization and introduces a novel approach for adding geometric details to specific regions and elements.

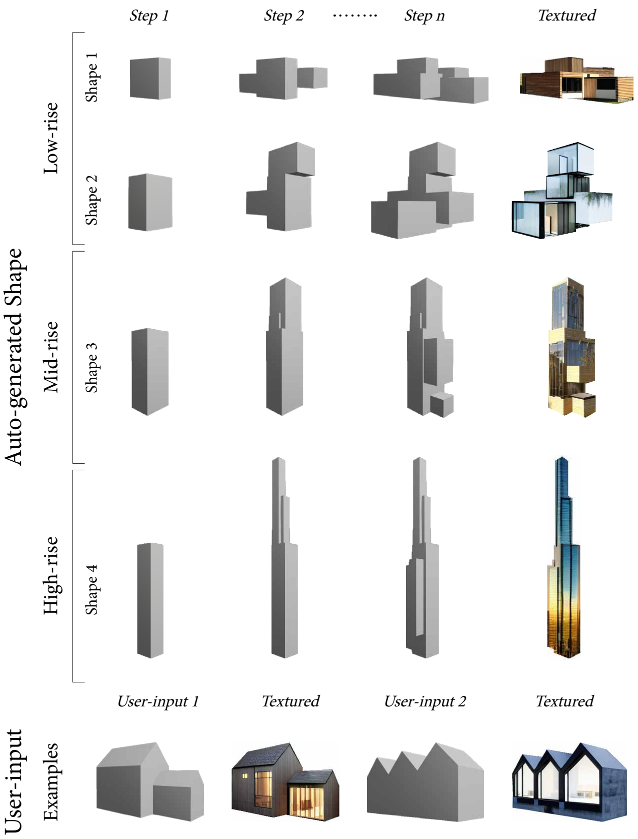


Figure 3. **Generating Coarse Geometry.** We generate 3D geometry using Constructive Solid Geometry (CSG), starting with an initial cuboid and iteratively adding cubes to create a "building massin" model. Each column shows a step in the generation. The last column shows the final generation after texturing.

3. Methodology

Our goal is to streamline the *schematic design phase* of the architectural design workflow [2]. During this months-long phase, involves iterative design changes as architects create preliminary drawings or visual materials to illustrate design concepts, including spatial relationships, scale, and form and communicate with owners to define project goals and requirements. Typically, even generating a single design proposal can take weeks. Our goal is to speed up this early decision-making process by allowing architects to generate numerous design options in a short period. Clients can also use our system to generate the designs they like and share them with architects, fostering collaboration and alignment.

Overview of approach: Our approach first builds a coarse 3D model (Section 3.1), where we achieve design complexity and variation by aggregating a sequence of randomly generated primitives. We then use segment-wise texture inpainting and projection to ensure consistent facade styles (Section 3.2). Finally, we apply object detection and segmentation to add details such as window openings to the facade (Section 3.3).

3.1. Generating Coarse Geometry

Current techniques optimize the geometry by minimizing the SDS loss [39] which penalizes unrealistic 2D projections (according to 2D generative models) of the 3D object being synthesized. However, SDS optimization is prone to mode collapse [52], which limits the diversity of the generated models. Furthermore, because the underlying 3D geometry being optimized is often represented as a NeRF [35], the resulting model has no prior for planar surfaces. This, in conjunction with slight imperfections in the rendered 2D views results in bumpy, non-planar surfaces that are often not perpendicular to each other (Fig. 1).

Instead of relying on SDS to directly yield a diverse set of meaningful 3D models, we use a simple procedural approach to create coarse 3D geometry, optimizing for diversity. Specifically, we start by randomly sampling an initial cuboid. Additional cuboids are added iteratively by randomly selecting a previously generated cuboid as an anchor and attaching the new cube to one of its faces. We inject some domain knowledge into the generation process by defining three kinds of buildings – low-rise, mid-rise, or high-rise – and constraining generated cuboid dimensions based on the user-defined building type. For instance, the height of the initial cuboid for a low-rise building is constrained to be between one and two times the width of the base dimensions. The precise constraints are defined in the supplementary material.

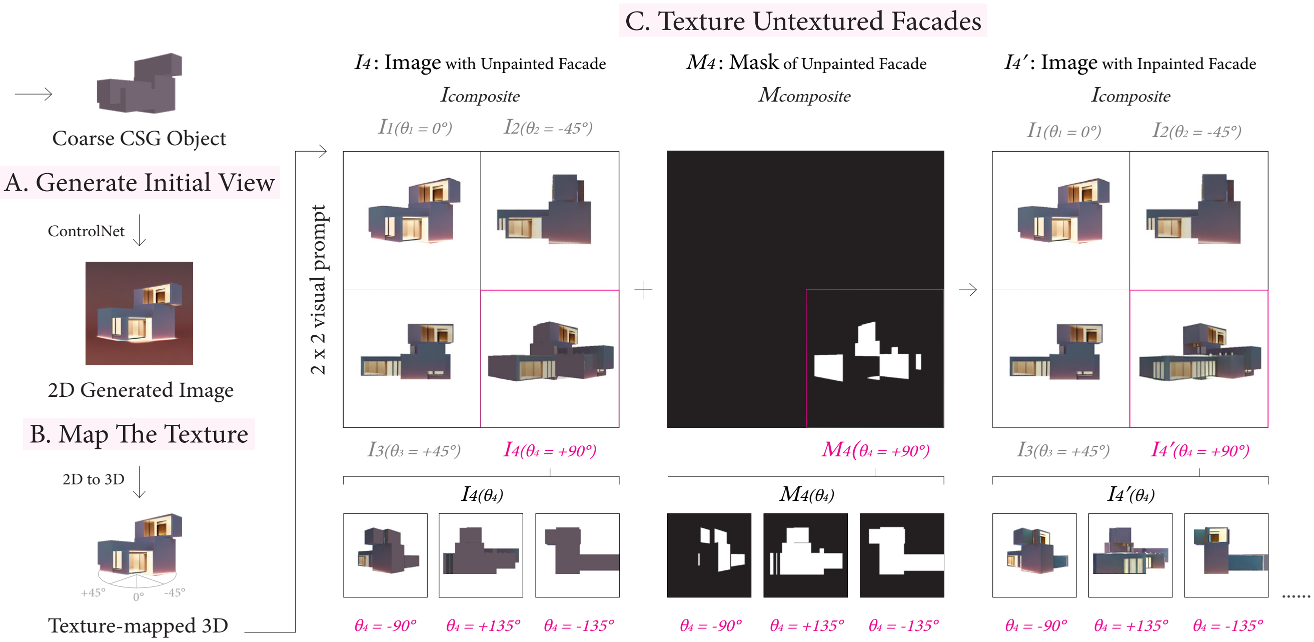
The final coarse 3D model is the union of the generated cuboids. Such a model, referred to as "building massing" in architecture, provides a simplified 3D representation that can be used to examine a building project's overall shape, form, and spatial organization during the early design stages [3].

Variations: Users can also input their own customized coarse 3D models for unique variations. The user-input examples in Fig. 3 were designed in Blender [7].

3.2. Synthesizing Textures

While SDS-based optimization yields good results for painting isolated objects with texture [8, 55–57], these fall short when applied to architecture. They project background textures onto building facades, or merge distinct facades into a single facades (Fig. 5). Unfortunately, the unrealistic building textures generated by these prior techniques are still photoconsistent and produce realistic images when viewed from different viewpoints. As such, SDS optimization-based techniques cannot improve upon these generations. To generate realistic building textures, we need to constrain the generation process further.

We propose an alternative approach where we synthesize



"A stand-alone house, crafted with high-tech architecture and composite materials, features openings. It is depicted in a photorealistic style from a side view at a street corner, bathed in the glow of the northern light."

Figure 4. **Synthesizing Textures.** To generate realistic building textures, we propose a novel facade-by-facade approach, constraining the generation process for each individual facade. We utilize a 2 × 2 visual prompt as input (Eq. (1)), providing a learnable context for the inpainting model to effectively learn and apply textures to the coarse 3D model. The facade-by-facade approach ensures cross-face continuity, mitigating the Janus problem [39] and producing consistent 3D-aware textures. The fully textured views of this example are provided in Fig. 7.

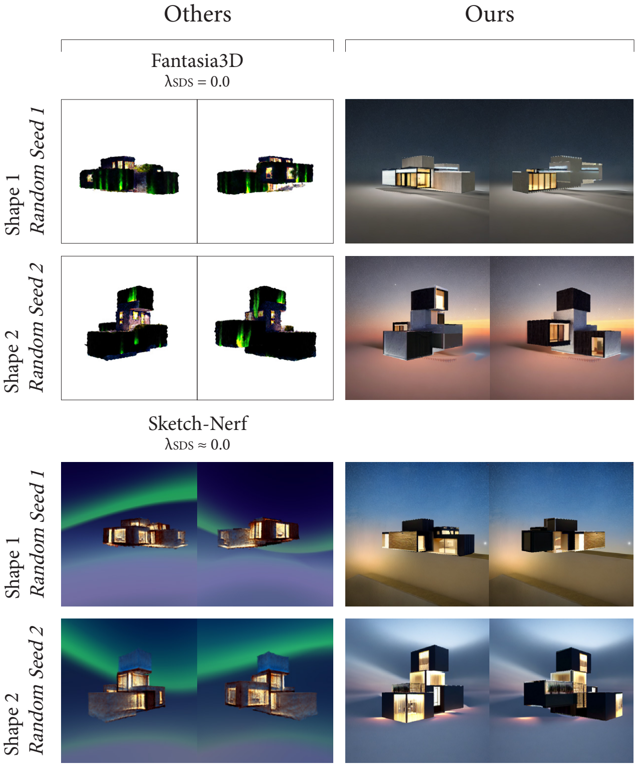
textures for the course model one facade at a time. This prevents the diffusion model from merging facades. It also provides enough geometric context at each step to the 2D diffusion model, ensuring that textures are applied accurately and appropriately, without arbitrary landscape elements interfering with the facades. In addition, the facade-by-facade approach mitigates the multi-face Janus problem [39] because each facade is painted in the context of other facades. Our texture synthesis process can be broken down into the following steps:

A. Generate Initial View. After creating the massing model, we automatically select a viewing direction at 45° to one of the facades (thus including two facades in the view). This viewpoint captures two neighboring facades, and renders it. The camera is positioned at a fixed distance from the mesh, using default intrinsic parameters with basic Lambertian shading and ambient lighting. The field of view is automatically adjusted to ensure the initial view is fully captured without any truncation. This rendered view is input to ControlNet [58], conditioning on depth estimation to generate the initial image. Including two facades in this initial view ensures that the output image contains sufficient design and style information for subsequent inpainting.

B. Map The Texture. The output image is subsequently mapped as a texture from 2D back onto the 3D massing model. We perform this texture mapping by dividing the facade into smaller faces logarithmically and mapping vertex colors to the nearest texture colors. The finer mesh enables automatic texture-dependent geometry editing later (Sec. 3.3).

C. Texture Untextured Facades. Given this initial texture, we next use an inpainting diffusion model to texture the remaining parts of the structure. To do so, we consider viewpoints at 45° increments in both directions from the initial viewpoint: while for simple cuboidal buildings, 90° increments are enough, for more complex geometries there may be extrusions that occlude parts of the facade, and so a finer increment is necessary. Concretely, we consider viewpoints at $45°, -45°, 90°, -90°, 135°, -135°$ and $180°$ in order from the initial viewpoint, and iteratively texture any untextured regions that are visible from each viewpoint.

For each viewpoint, we first identify the hitherto untextured parts of the image by comparing a render of the textured model with a render of the original model without texturing. Any pixels with identical colors in the two renders have likely not been textured. We then create a mask with this set of pixels and use an inpainting model to fill in these regions. Crucially, to give the inpainting model



"A stand-alone house, crafted with high-tech architecture and composite materials, features openings. It is depicted in a photorealistic style from a side view at a street corner, bathed in the glow of the northern light."

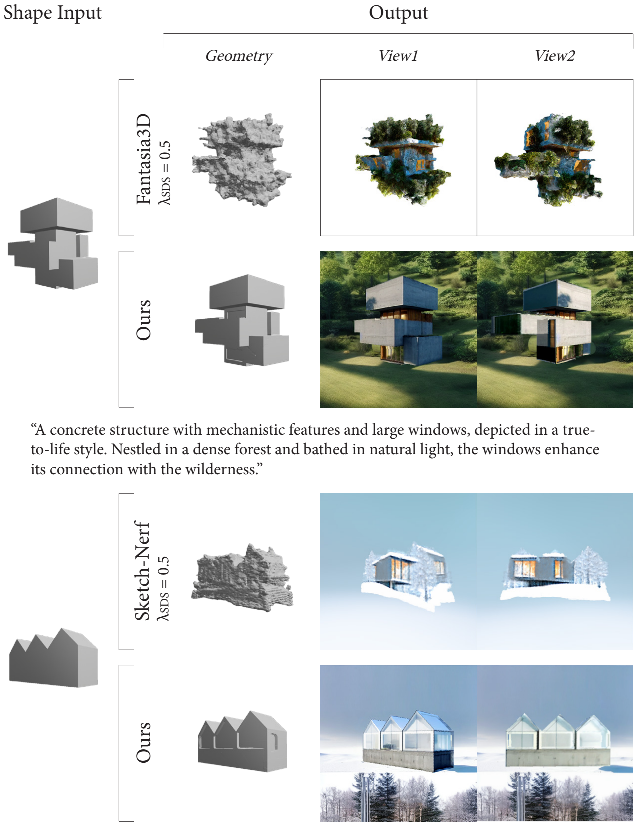
Figure 5. **Texture Synthesis Comparison.** Our approach (right) generates diverse design options from the same coarse geometry (Shape 1 and Shape 2, as shown in Fig. 3) and identical text prompts, while maintaining consistent styles. In contrast, existing methods [9, 34] (left) produce similar, low-quality, and unvaried results, often blending backgrounds into facades or rendering parts of the geometry invisible by blending into the sky. $\lambda_{SDS} \approx 0$ for these techniques implies that the initial geometry is preserved.

enough stylistic context, we propose a novel *visual prompt* where we lay four views of the building out in a 2×2 grid:

$$I_{\text{composite}} = \begin{array}{|c|c|}\hline I_1(\theta_1) & I_2(\theta_2) \\ \hline I_3(\theta_3) & I_4(\theta_4) \\ \hline \end{array} \tag{1}$$

Here, $\theta_1 = 0°, \theta_2 = +45°, \theta_3 = -45°$ and θ_4 is the viewpoint to be generated. The masks are laid out in a corresponding manner to produce a composite mask $M_{\text{composite}}$ (when texturing the $+45°$ (or $-45°$) viewpoint, we use $\theta_4 = 0°$). The composite image and mask are then fed into an inpainting model [43]. The output of this inpainting model is texture mapped back onto the 3D model as above before moving on to the next viewpoint. We found experimentally that this grid structure provided enough context to the inpainting model and led to more stylistically consistent generations.

In spite of the finer 45° increments, occlusion may still cause a challenge when it results in one viewpoint having multiple untextured adjacent faces with different orientations. In such cases, we found that the inpainting model struggles to distinguish faces with different orientations and may texture them in a manner inconsistent with their orientation. To address this, we group untextured faces that share the same orientation into *facade sets*. We then iterate through all the facade sets in each viewpoint, inpainting them separately.



"A concrete structure with mechanistic features and large windows, depicted in a true-to-life style. Nestled in a dense forest and bathed in natural light, the windows enhance its connection with the wilderness."

"A minimalist modern house with large glass windows and a concrete structure, elevated on stilts in a snowy forest. Bare trees covered in light snow create a serene winter scene."

Figure 6. **Specifying Facade Details.** We apply open-vocabulary recognition to automatically detect specific elements using keywords (here, "windows") allowing us to edit details (here, pushing windows inward) precisely where needed, compared to global edits of past work [9, 34].

Further details and handling of some edge cases are provided in the supplementary material.

3.3. Specifying Facade Details

Once our coarse building model has been textured, we add further details to the geometry conditioning on the generated texture. Existing shape-guided synthesis techniques [9, 34] are limited to whole-geometry modifications and cannot adjust specific regions or elements. Additionally, they produce bumpy surfaces, complicating further de-

Method (Text-guided)	FID↓	KID↓	Avg. CLIP ↑	CLIP Accuracy↑	Precision↑	Recall↑
Ours	**222.45**	**0.10±1.58**	**0.271**	**0.990**	**0.19**	**0.30**
DreamFusion-Hifa [61]	259.90	0.13±2.31	0.256	0.970	0.11	0.07
ProlificDreamer [53]	288.88	0.16±2.16	0.222	0.810	0.01	0.14
MVDream [48]	340.31	0.25±1.57	0.264	0.970	0.01	0.00
TextMesh [51]	315.98	0.14±1.79	0.203	0.680	0.00	0.01
Magic3D [29]	355.60	0.19±1.78	0.182	0.490	0.00	0.00

Table 1. **Quantitative Evaluation.** We quantitatively evaluate our model and other text-guided baselines using metrics including FID [17], KID [6], CLIP [41], CLIP Accuracy, and Precision-and-Recall [25].

Method (Text-guided)	More Diverse ↑	More Realistic ↑	More Fitting to the Prompt ↑
Ours	**95.0%**	**84.3%**	**50.3%**
DreamFusion-Hifa [61]	5.0%	15.7%	49.7%

Table 2. **User Study Results.** Preference percentages for diversity, realism, and prompt alignment were evaluated across 10 prompts, with a total of 20 comparisons assessed by 20 participants (designers and the general public), resulting in 400 instances. Our method significantly outperforms DreamFusion-Hifa, the strongest competitor in quantitative assessments.

sign manipulation. We propose an alternative that yields localized, semantics-based edits that preserve clean geometry.

In our proposed appproach, the user specifies which architectural elements to edit: windows, doors, balconies etc and how to edit them (e.g., intrusion or extrusion). We use off-the-shelf open-vocabulary object detectors [36] to detect these architectural elements on the rendered views. We use the midpoints of the detected boxes as prompts for SAM [21] to generate precise masks. Finally, we edit the geometry under these masks by intruding or extruding based on user input (intrusion/extrusion can be performed by moving the corresponding mesh vertices in the direction of the face normal). Fig. 6 shows an example of such edits, where we have automatically intruded the windows into the facade, in contrast to the global edits of prior work.

4. Experiments

To evaluate our proposed 3D synthesis pipeline, we conducted experiments to measure the quality and diversity of generated images, the alignment between images and text descriptions, and style consistency across different views of a building design.

We compared our method with **text-guided models** (MVDream [48], Prolificdreamer [53], Dreamfusion-Hifa [61], Magic3D [29], TextMesh [51]), **image-guided models** (Magic123-Hifa [61], DreamCraft3D [49], RealFusion [33], SyncDreamer [32], Stable Zero123 [1], ZeroNVS [44]), and **shape-guided models** (Latent-nerf [34], Fantasia3D [9]). Some of the baselines are from the threestudio library [15].

Prompts for generation were manually created with assistance from ChatGPT [38] to ensure that they have sufficient detail for good generation.

4.1. Quantitative Evaluation

Realism and diversity: We first evaluated the realism and diversity of generated models vis-a-vis the real world. To do so, we used 10 prompts to generate 10 examples each, selecting one viewpoint for each example, resulting in 100 images. For comparison, we searched for corresponding building images using the same set of prompts on Bing Image [5] and Google Image [13]. We then used the generated and real images and computed the Frechet Inception Distance (FID) [17] and Kernel Inception Distance (KID) [6] between the generated images and real buildings from the same prompt. Tab. 1 shows the average (across prompts) FID and KID scores for our approach and other text-guided baselines. Our FID and KID scores were the lowest among the models, indicating higher quality, realism, and diversity.

We also evaluated Precision (fraction of generations that are realistic) and Recall (how much of the real manifold is captured by generations) [25] using an InceptionV3 feature extractor [24]. Once again, on both metrics our approach performs the best, indicating that our approach produces more realistic and more varied generations.

Adherence to style/prompt: We used CLIP [41] to evaluate the alignment between images and the corresponding prompt. We report both the average CLIP score as well as a "CLIP Accuracy": the fraction of generated images that achieve CLIP score higher than a threshold. We used as a threshold the lowest CLIP score from our real image data. Both metrics measure how closely the generated images adhere to the provided prompt. Our approach yields the high-

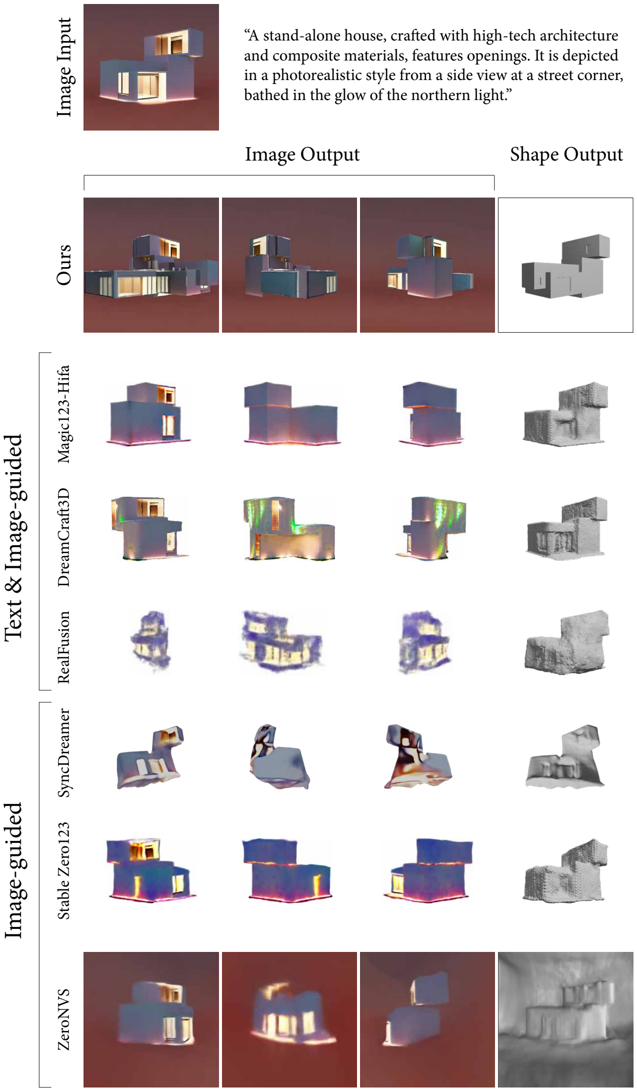


Image Input
"A stand-alone house, crafted with high-tech architecture and composite materials, features openings. It is depicted in a photorealistic style from a side view at a street corner, bathed in the glow of the northern light."
Image Output
Shape Output
Ours
Text & Image-guided
Magic123-Hifa
DreamCraft3D
RealFusion
Image-guided
SyncDreamer
Stable Zero123
ZeroNVS

Figure 7. **Image-guided Models Comparison.** We compare to image-guided methods [1,32,33,44,49,61] by providing them our first generated view as input. Existing image-guided models produce artifacts such as non-flat floors, sky blending into facades, or solid walls without openings, and overly simplistic forms. In contrast, our approach (top) generates realistic and varied architectural designs.

est score on both metrics, indicating that we match the requested style the best.

Finally, since the true arbiter of generated designs should be humans, we conducted a **user study**. We compared our approach to the strongest text-guided baseline, DreamFusion-Hifa [61]. In our user study, 20 participants, including designers and the laypeople, evaluated the diversity, realism, and prompt alignment of 20 pairs of generations from 10 prompts (2 pairs per prompt). For each comparison, participants reviewed 2 outputs from each method side by side and selected the result that best met the criteria. The order of all outputs were randomized to prevent bias. Overall, 400 evaluation instances were collected. As shown in Tab. 2, annotators overwhelmingly preferred our generations in terms of diversity and realism. There was less of a difference in terms of adherence to the prompt, which is understandable since our proposed approach is not aimed at improving prompt alignment per se.

4.2. Qualitative Evaluation

Qualitative comparisons to text-guided techniques [48, 53,61] are shown in Figs. 1 and 2. As already demonstrated by the quantitative results and the user study, our approach produces more diverse and more realistic buildings. Importantly, where prior work produces simplistic forms with bumpy surfaces and uneven walls, our approach yields sophisticated forms with clean geometry that can be edited through the design process.

We also compare our approach for texture synthesis and adding facade details to prior work on shape-guided synthesis [9, 34] in Figs. 5 and 6. Prior work produces repetitive, unrealistic textures with background artifacts (Fig. 5) and introduces bumps and uneven surfaces (Fig. 6). In contrast, our approach yields clean geometry, realistic and diverse textures, and allows for simple, localized editing.

Finally, we also compare to image-guided (e.g.,novel view synthesis) techniques [1, 32, 33, 44, 49, 61] by providing them as input the first textured view from our pipeline (Fig. 7). Even with a provided view, these techniques produce blurry, plain or unrealistic textures, and result in uneven, unrealistic geometry. Once again, we obtain sharp views as well as clean geometry.

5. Limitations and Conclusion

We introduce the first approach for 3D synthesis for architectural design, with the goal to enable rapid generation of design options in the early stages of the design process. Our approach uses coarse 3D model generation from primitives to enhance design diversity, a facade-by-facade texturing approach to ensure stylistic consistency and realism, and open-vocabulary recognition to accurately detect and detail specific elements.

Our proposed approach significantly increases diversity and realism compared to baselines while maintaining clean geometry. In terms of limitations, the method's performance is sensitive to hyperparameters such as the thresholds for facade-element detection, and the prompts and outputs from ControlNet and the inpainting model. Lastly, our geometry generation is limited to non-free-form structures, which still encompass most architectural designs.

Acknowledgements: This work is partly supported by NSF IIS 2144117. We would also like to thank Luming Tang for the valuable discussions and feedback.

References

[1] Stability AI. Stable zero123: Quality 3d object generation from single images, 2024. Retrieved from Stability AI. 2, 7, 8

[2] American Institute of Architects. Defining the architect's basic services, n.d. 4

[3] Autodesk. Massing model architecture, 2024. Accessed: 2024-06-25. 2, 4

[4] Mathias Bank, Viktoria Sandor, Kristina Schinegger, and Stefan Rutzinger. Learning spatiality: A gan method for designing architectural models through labelled sections. *Legal Depot D/2022/14982/02*, page 611, 2022. 2

[5] Building images. https://www.bing.com/images. Accessed: 2024-05-28. 7

[6] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 7

[7] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Amsterdam, The Netherlands, 2024. 4

[8] Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors, 2023. 3, 4

[9] 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)*, pages 22246–22256, October 2023. 2, 3, 6, 7, 8

[10] Zhiqin Chen, Vladimir G. Kim, Matthew Fisher, Noam Aigerman, Hao Zhang, and Siddhartha Chaudhuri. Decor-gan: 3d shape detailization by conditional refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15740–15749, June 2021. 2

[11] 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. *arXiv preprint arXiv:2212.08051*, 2022. 2

[12] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10281–10291, 2021. 2

[13] Building images. https://images.google.com. Accessed: 2024-05-28. 7

[14] Yanhui Guo, Xinxin Zuo, Peng Dai, Juwei Lu, Xiaolin Wu, Li Cheng, Youliang Yan, Songcen Xu, and Xiaofei Wu. Decorate3d: Text-driven high-quality texture generation for mesh decoration in the wild. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[15] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 7

[16] Nervana Osama Hanafy. Artificial intelligence's effects on design process creativity: "a study on used a.i. text-to-image in architecture". *Journal of Building Engineering*, 80:107999, 2023. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *ArXiv*, abs/1706.08500, 2017. 7

[18] 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

[19] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 2

[20] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Popa Tiberiu. Clip-mesh: Generating textured meshes from text using pretrained image-text models. *SIGGRAPH Asia 2022 Conference Papers*, December 2022. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 7

[22] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. *arXiv preprint arXiv:2402.03908*, 2024. 2

[23] Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. *arXiv preprint arXiv:2312.01305*, 2023. 2

[24] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 7

[25] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *CoRR*, abs/1904.06991, 2019. 7

[26] Neil Leach. In the mirror of ai: what is creativity? *Architectural Intelligence*, 1(15), 2022. 2

[27] Manyi Li and Hao Zhang. D2im-net: Learning detail disentangled implicit fields from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10246–10255, June 2021. 2

[28] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023. 2

[29] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 7

[30] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[31] 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

[32] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 7, 8

[33] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 2, 7, 8

[34] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 2, 3, 6, 7, 8

[35] 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*, 2020. 3, 4

[36] Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *ArXiv*, abs/2205.06230, 2022. 2, 7

[37] Nelson Nauata, Kai-Hung Chang, Chin-Yi Cheng, Greg Mori, and Yasutaka Furukawa. House-gan: Relational generative adversarial networks for graph-constrained house layout generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 162–177. Springer, 2020. 2

[38] OpenAI. Introducing chatgpt. `https://openai.com/blog/chatgpt`, 2024. Accessed: May 20, 2024. 7

[39] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2, 4, 5

[40] 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. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 2

[41] 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 *International Conference on Machine Learning*, 2021. 7

[42] Morteza Rahbar, Mohammadjavad Mahdavinejad, Amir HD Markazi, and Mohammadreza Bemanian. Architectural layout design through deep learning and agent-based modeling: A hybrid approach. *Journal of Building Engineering*, 47:103822, 2022. 2

[43] 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 (CVPR)*, pages 10684–10695, June 2022. 2, 6

[44] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. *CVPR, 2024*, 2023. 2, 7, 8

[45] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 2

[46] 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 M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6087–6101. Curran Associates, Inc., 2021. 2

[47] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[48] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 1, 2, 3, 7, 8

[49] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 2, 7, 8

[50] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. 2021. 2

[51] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. In *International conference on 3D vision (3DV)*, 2024. 2, 7

[52] Peihao Wang, Dejia Xu, Zhiwen Fan, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, and Vikas Chandra. Taming mode collapse in score distillation for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9037–9047, June 2024. 2, 4

[53] 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 preprint arXiv:2305.16213*, 2023. 1, 2, 3, 7, 8

[54] Francis Williams, Matthew Trager, Joan Bruna, and Denis Zorin. Neural splines: Fitting 3d surfaces with infinitely-wide neural networks. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 9949–9958, June 2021. 2

[55] Yu-Ying Yeh, Jia-Bin Huang, Changil Kim, Lei Xiao, Thu Nguyen-Phuoc, Numair Khan, Cheng Zhang, Manmohan Chandraker, Carl S Marshall, Zhao Dong, et al. Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. *arXiv preprint arXiv:2401.09416*, 2024. 3, 4

[56] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 4

[57] 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, 2023. 3, 4

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 5

[59] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang YU, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[60] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. *arXiv*, 2023. 2

[61] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance, 2023. 1, 2, 3, 7, 8