

A Survey of Single-Image 3D Reconstruction for Low-Latency Interactive Applications and Engineering Deployment

Zihan Chen

Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, 999078, China

1230026951@student.must.edu.mo

ABSTRACT

Single-image 3D reconstruction aims to recover or generate the three-dimensional geometry and appearance of a target object from a single two-dimensional image. It has considerable application value in fields such as e-commerce, game development, virtual reality, augmented reality, and digital content creation. In recent years, relevant techniques have evolved from explicit representations, including voxels, point clouds, and meshes, toward neural implicit representations, multi-view diffusion models, large feed-forward reconstruction models, and general-purpose 3D generative models. Existing studies generally focus on geometric accuracy and visual quality; however, different methods exhibit substantial differences in inference latency, GPU memory requirements, input preprocessing, output asset formats, and post-processing costs. This paper reviews the major technical pipelines of single-image 3D reconstruction, with a focus on representative methods including 3D-R2N2, PSGN, Pixel2Mesh, Occupancy Networks, NeRF, Zero-1-to-3, Wonder3D, LRM, TripoSR, and CRM. These methods are compared from the perspectives of latency levels, end-to-end latency, computational resources, input robustness, asset usability, and platform integration difficulty. On this basis, this paper summarizes the major challenges associated with unobserved-region inference, generalization to real-world images, adaptation to different camera settings and materials, mesh quality, and evaluation criteria. It further discusses future directions, including hierarchical generation, 3D foundation models, sparse computation, unified latent spaces, interactive editing, and the generation of production-ready 3D assets.

Keywords: Single-Image 3D Reconstruction; 3D Generation; Feed-Forward Reconstruction; Triplane; Low-Latency Interaction; Engineering Deployment

1. INTRODUCTION

Single-image 3D reconstruction refers to the technique of inferring the three-dimensional geometry and appearance of an object from a single two-dimensional image. Compared with traditional multi-view reconstruction, this task reduces the requirements for image acquisition and camera calibration, making it more suitable for interactive scenarios where users upload a single image and obtain a 3D model. Learning-based approaches can further exploit shape priors learned from training data to predict the overall structure of objects when there are large viewpoint variations, insufficient textures, or failures of traditional feature matching methods ^[1].

However, a single image only provides a projection of an object from a specific viewpoint, while its depth information, back-side geometry, and occluded regions cannot be directly observed. The same input image may correspond to multiple plausible 3D interpretations, and models can only infer the missing parts by combining visible information with learned data distributions ^[2]. Therefore, modern approaches not only reconstruct visible regions but also generate unseen parts, gradually blurring the boundary between traditional “reconstruction” and data-driven “generation”.

This paper does not equate low-latency interaction with real-time operation in the sense of continuous video frame-rate processing. To facilitate the comparison of different methods, this paper categorizes the response process into three analytical levels based on the end-to-end generation time reported in the original studies. Sub-second outputs are mainly used for instant previews, candidate selection, and batch generation; generation processes taking approximately 1–10 seconds can support interactive generation, where users can continue adjusting inputs or parameters after receiving the results; minute-level pipelines are generally more suitable for offline or semi-offline asset production with higher priorities on geometric details and quality. This classification serves as an operational definition for engineering analysis in this paper rather than a unified industry standard.

2. DEVELOPMENT OF SINGLE-IMAGE 3D RECONSTRUCTION TECHNIQUES

2.1 Early Methods Based on Explicit Representations

“Voxel-based representation” divides 3D space into regular grids and predicts whether each grid cell is filled. Because of its regular structure, it is easy to be processed in 3D convolutional neural networks. However, if the resolution is improved, the amount of data saved and the cost required for calculation will increase on the basis of cubic. Low-resolution voxel representation sometimes disappears things like thin rods, flat structures, and tiny places like holes. 3D-R2N2 uses a 2D converter to extract image features, and uses a 3D creation unit and a decoder to create occupancy grids. Because of the repetitive structure, the reconstructed results can be updated bit by bit after adding the images of new viewpoints [1]. This method can be used to reconstruct a single image or multiple images, but the output resolution is limited by voxel representation.

Point cloud representation only stores the separated 3D coordinates of the surface of the object, so there is no need to use the storage location of the data in a large space with nothing. PSGN directly predicts the unordered point set by fully connected network, and determines the corresponding relationship between predicted points and correctly answered points by using Chamfer Distance and Earth Mover’s Distance [2]. Although point cloud can represent fine structures better than low-resolution voxel, there is no relationship between points. Therefore, when importing point cloud into commonly used 3d software, surface reconstruction and texture mapping are needed.

Mesh is composed of vertices, edges and faces, which is widely used in the output of modeling, animation and asset. Pixel2Mesh starts from the initial oval mesh, and uses Graph Convective Network and image-aware feature to deform the vertices bit by bit, forming a triangular mesh from a rough state to a fine state [3]. Compared with the performance of voxel and point cloud, the output result is closer to the actual 3D asset. However, the mesh has the same connection method as the original template, so it is difficult to represent complex holes and many unconnected parts [3]. Moreover, if the adjustment of the side length is cancelled in paper’s ablation study, the numerical indicators will become better, but there are obvious shortcomings of mesh, and it is understood that the indicators based on point distance alone cannot judge the whole quality of mesh [3].

2.2 Continuous Implicit Representations and Neural Radiance Fields

In order to eliminate the certain voxel resolution limit, Occupancy Networks uses neural networks to predict the occupation probability of favorite 3D positions, and represents the object surface as the continuous decision boundary of classifier [4]. The model can infer continuous occupancy function from single image, point clouds and low-resolution voxel input, and make watertight mesh by multi-resolution isosurface extraction and Marching Cubes algorithm [4]. This method can reconstruct the surface with high resolution even if the cost is low. However, making mesh still needs a lot of spatial query, isosurface extraction, simplification and surface optimization. According to the paper, it takes about 3 seconds to make a mesh during all the speculation [4].

Nerf (Neural Radiance Fields) also uses multilayer perceptrons to represent colors that change according to the density and viewpoint of continuous space, and uses differentiable volume rendering to make images of new viewpoints [5].

The original NeRF needs multiple images with known camera postures, and optimizes each new scene independently. Therefore, it is not a direct single-image mesh reconstruction method. The most important function is to establish the structure of continuous neural fields and volume rendering, which provides a technical basis for the NeRF method using single-image and the model based on Triplane mentioned later. Instant-NGP accelerates the training and rendering efficiency of neural graphics primitives by introducing a multi-resolution hash table encoding for spatial coordinates, and combined with small neural networks, fully integrated CUDA kernels, space flight methods and other paper, it writes that high-quality neural fields begin to form in a few seconds, and 1920×1080 images can be drawn in tens of milliseconds [6]. However, this speed is attributed to the implementation of hash encoding, GPU memory cache and CUDA [6]. In the work of using NeRF, Instant-NGP also needs multi-view input and adjustment of each scene, so it is not so much a direct single-image feed-forward reconstruction model as an implicit expression of engineering.

2.3 Reconstruction Methods Based on Multi-view Diffusion

Two-dimensional diffusion models have learned rich semantic and appearance priors from large-scale image datasets, providing a new approach for completing invisible regions in single images. Zero-1-to-3 formulates single-image novel-view synthesis as a camera-conditioned image translation task. Given an input image and relative camera rotation and translation parameters, it generates images from specified viewpoints [7]. The method itself mainly produces 2D novel

views rather than complete 3D models. Full 3D reconstruction requires using the generated views as priors to supervise neural fields and performing per-object optimization through frameworks such as SJC, followed by mesh extraction from density fields [7]. This approach improves open-category generalization and can generate multiple plausible results for the same invisible viewpoint. However, stochastic optimization introduces additional computational overhead and may lead to inconsistencies between different views.

Wonder3D extends this approach by jointly generating multi-view color and normal information. Through cross-view and cross-domain attention mechanisms, the model simultaneously generates six-view color images and normal maps, then optimizes a signed distance field and obtains textured meshes through geometry-aware normal fusion [8]. Normal maps provide more direct surface information for geometric reconstruction compared with color images alone. However, the complete pipeline still requires approximately 2–3 minutes [8]. The paper also indicates that the current six-view configuration remains insufficient for accurately recovering extremely thin structures and severely occluded regions [8]. Therefore, multi-view diffusion approaches can improve detail quality and generalization capability, but their sampling procedures and subsequent geometric optimization continue to limit online response speed.

2.4 Large Feed-forward Reconstruction Models

LRM (Large Reconstruction Model) advances single-image reconstruction from per-object optimization to single-pass forward inference after large-scale pre-training. It uses DINO to extract local and global image features, and a Transformer maps 2D features into a Triplane-NeRF representation through cross-attention. The model learns general 3D priors through multi-view rendering supervision [9]. The model contains approximately 500 million parameters and is trained on around one million 3D objects and video data. It completes Triplane prediction, 3D field querying, and mesh extraction in approximately 5 seconds on an A100 GPU [9]. However, the fixed camera assumption may introduce shape distortions when input cropping or field-of-view changes occur. The model mainly focuses on objects without backgrounds and does not explicitly model complex view-dependent appearance [9].

TripoSR inherits the DINO encoder, Image-to-Triplane Transformer, and Triplane-NeRF architecture from LRM. It improves efficiency and stability through data filtering, rendering strategies closer to real-image distributions, Mask Loss, and local high-resolution supervision [10]. It sets the output channel number of Triplane features to 40 to balance GPU memory consumption in volume rendering and representation capacity. Meanwhile, it removes explicit camera parameter conditioning, enabling the model to implicitly estimate camera relationships during both training and inference [10]. The paper reports that TripoSR can generate meshes within 0.5 seconds on an A100 GPU and releases its source code and pre-trained models under the MIT license [10].

CRM combines diffusion-based completion with feed-forward reconstruction. In this method, firstly, six orthogonal color views and canonical coordinate maps (CCMs) are made, and then the information of the aligned view is mapped three times by convolutional U-Net into a Triplane representation, followed by FlexiCubes to generate textured meshes [11]. According to the paper, the whole processing of A800 GPU takes about 10 seconds, and the two diffusion sampling steps take about 3 seconds respectively. The remaining time is used for the input and output of 3D reconstruction, surface querying, texture processing and file, and the forward pass of U-Net itself is less than 0.1 second [11]. It can be seen that the bottleneck of the whole system comes not only from the reconstruction network, but also from the generative preprocessing and asset export processes.

3. COMPARATIVE ANALYSIS FROM THE PERSPECTIVE OF REAL-TIME PERFORMANCE AND ENGINEERING DEPLOYMENT

3.1 Inference Latency and Computational Resources

The reaction speed of various technical pipeline is different. Zero-1-to-3 needs to prepare a neural field for everything using generation hints. Wonder3D provides multi-view diffusion and signed distance field (SDF) fusion. LRM and TripoSR directly predict the representation of Triplane in the learned model. CRM integrates multi-view diffusion, convolutional reconstruction and mesh texturing into a pipeline. Due to these differences, the generation time of each method is divided into minutes, seconds, less than one second and groups with different lengths.

Low-latency preview is suitable for 3D prototype visualization of e-commerce products, selection of candidate model, generation of large content and other applications. The method of interactive per second is suitable for 3D creation platform, in which users can upload an image, view the results, and then adjust the input. Minute method can be used for offline asset production, in which geometric accuracy and quality take precedence over immediate results. Therefore, the selection

of methods should not only rely on the shortest inference time, but also consider the output quality, hardware requirements and post-processing costs. The execution time and main deployment constraints of representative method reports are summarized in Table 1.

3.2 Input Robustness and Invisible Regions

Many experimental settings use inputs with clean background, centered target, uniform size and uniform viewpoint. However, real user images may contain complex backgrounds, multiple objects, occlusions, blur, and extreme viewpoints. Occupancy Networks still require instance-level mask cropping to extract the target region in real-image experiments [4]. LRM relies on a fixed camera assumption, Wonder3D is sensitive to limited viewpoints and severe occlusions, and the diffusion-based outputs generated by CRM may also introduce part or color inconsistencies across different views.

For engineering systems, object detection, background removal, image quality assessment, viewpoint estimation, and multi-object rejection can be designed as independent preprocessing modules. Since invisible regions do not have a unique solution, platforms can generate multiple candidate results or provide confidence scores and regeneration options. This prevents a single generated output from being interpreted as a deterministic recovery of the actual unseen surfaces of an object.

3.3 Usability of Output Assets

Voxel, point cloud, Occupancy Field, NeRF, and Triplane representations all require different levels of surface extraction before becoming commonly used mesh assets. Even when a model directly outputs meshes, the results may still contain holes, self-intersections, non-manifold edges, floating fragments, excessive polygon counts, or texture seams. Vertex colors are also not equivalent to complete UV textures and physically based rendering (Physically Based Rendering, PBR) materials.

Practical deployment should distinguish between preview-level results and production-level assets. Preview-level results mainly focus on generation speed and visual plausibility across multiple viewpoints, while production-level assets additionally require topology inspection, polygon count control, normal correction, UV unwrapping, texture baking, and format conversion. If these procedures still require substantial manual processing, even a fast core model inference process may not significantly reduce the overall cost of content production.

Different methods exhibit significant differences in intermediate representations, optimization strategies, reported runtimes, and final output formats. For a more concentrated comparison, the main engineering characteristics of representative methods are summarized in Table 1.

Table 1. Comparison of Engineering Characteristics of Representative Methods.

Method	Main Representation or Pipeline	Per-object Optimization	Reported Runtime in Paper	Main Engineering Limitations
3D-R2N2	Voxel occupancy grid	No	Not systematically reported	Resolution limited by voxel storage constraints [1]
PSGN	Point cloud	No	Not systematically reported	Lacks topology and requires additional mesh reconstruction [2]
Pixel2Mesh	Template-deformed mesh	No	Not systematically reported	Fixed template topology and incomplete texture representation [3]
Occupancy Networks	Continuous occupancy function and mesh	No	Approximately 3 s/mesh	Requires multi-resolution queries and isosurface extraction [4]
Zero-1-to-3	View-conditioned diffusion + neural field optimization	Yes	Not uniformly reported	Slow stochastic optimization and possible view inconsistency [7]
Wonder3D	Multi-view color and normal generation + SDF fusion	Yes	Approximately 2–3 min	Six-view configuration limits thin-structure and occlusion recovery [8]

Table 1. (continued) Comparison of Engineering Characteristics of Representative Methods.

Method	Main Representation or Pipeline	Per-object Optimization	Reported Runtime in Paper	Main Engineering Limitations
LRM	Triplane-NeRF and mesh	No	Approximately 5 s on A100	Large model size and strong background/camera assumptions ^[9]
TripoS	Triplane-NeRF and mesh	No	Approximately 0.5 s on A100	Limited reporting of complete memory usage and asset post-processing costs ^[10]
CRM	Multi-view diffusion + FlexiCubes textured mesh	No	Approximately 10 s on A800	Two diffusion stages increase latency and may propagate errors ^[11]

As shown in Table 1, explicit representation methods generally have relatively straightforward output pipelines, but they are limited by resolution, topology, or texture quality. Multi-view diffusion methods can leverage 2D generative priors to complete invisible regions, but they often require diffusion sampling and subsequent geometric optimization. Large-scale feed-forward models shift the primary computational cost to the offline training stage and achieve lower latency during inference. Since different papers adopt different hardware configurations, output formats, and timing scopes, the runtimes listed in Table 1 are only intended to illustrate the response levels of different approaches and should not be regarded as a strict unified speed ranking.

4. INSUFFICIENT CROSS-METHOD EVALUATION AND DEPLOYMENT VALIDATION

Existing studies lack unified settings in terms of datasets, input preprocessing, coordinate normalization, and geometric alignment strategies, making quantitative results across different papers difficult to compare directly. For example, Wonder3D aligns the generated shapes with ground-truth shapes before calculating Chamfer Distance and volume IoU ^[8]. TripoS first aligns predicted results using rotation angle search and the Iterative Closest Point (ICP) algorithm before computing geometric metrics ^[10]. Occupancy Networks samples 100,000 points from both predicted and ground-truth meshes to calculate Chamfer-L1 and estimates IoU based on high-resolution watertight meshes ^[4]. These evaluation settings differ in alignment methods, sampling numbers, and mesh processing procedures. Therefore, even when the same metric name is used, the resulting values may not be strictly comparable.

The reporting of computational efficiency also lacks unified standards. LRM includes image-to-Triplane prediction, 3D field querying, and Marching Cubes mesh extraction in its approximately 5-second runtime ^[9]. The approximately 10-second runtime reported by CRM includes two diffusion sampling stages, surface querying, texture processing, and file input/output ^[11]. Although TripoS reports approximately 0.5 seconds for mesh generation, it does not provide an equally detailed breakdown of individual processing stages ^[10]. Therefore, core network inference time, complete mesh generation time, and platform-level end-to-end latency should be reported separately to avoid directly ranking results measured with different timing boundaries.

Engineering-oriented evaluations should further include metrics such as peak GPU memory usage, model file size, latency on consumer-grade GPUs, cold-start time, concurrent throughput, mesh validity, output polygon count, and post-processing time. Future studies should jointly report geometric accuracy, visual quality, complete asset generation latency, and resource consumption under consistent datasets, input preprocessing procedures, hardware configurations, and output specifications, thereby improving the reproducibility and practical value of cross-method evaluations.

5. FUTURE DEVELOPMENT TRENDS

5.1 Hierarchical Generation and Quality–Latency Trade-offs

Generating the highest-resolution assets in a single step usually increases inference costs and GPU memory consumption. A more practical engineering approach is to first generate basic models for interactive previews, and then perform geometry super-resolution, local refinement, or high-quality texture generation according to user requirements. TripoS identifies 3D model super-resolution as one of its future directions and proposes further scaling model size, data volume, and latent space capacity ^[12]. This hierarchical mechanism can concentrate high-cost computation on user-confirmed results or critical regions.

5.2 3D Foundation Models, Data Governance, and Sparse Computation

3D generation is moving toward larger-scale models and datasets; however, data quality cannot be measured solely by the number of samples. The ablation study of TripoSG demonstrates that, under the Normal-FID metric, 180,000 samples processed through scoring, filtering, orientation correction, and model processing outperform 800,000 unprocessed Objaverse samples. After maintaining data quality, further scaling the dataset to 2 million samples provides additional improvements^[12]. Therefore, data governance, consistent orientation, topology repair, and high-quality rendering may become important foundations for future 3D foundation models.

Model scaling also needs to control the actual cost of the computer. TripoSG uses a mechanism called Mixture of Experts (MoE), which is rarely used. TRELLIS only stores local latent variables on the surface of object and similar active voxels to reduce the calculation of empty positions^[13]. Rectified Flow is TripoSG and TRELLIS, and it is found that it has increased power. However, based on the current results, it is difficult to decide to replace diffusion models. On the contrary, it is just right to regard Rectified Flow as one of the candidate methods for large-scale 3D generation.

5.3 Unified Representations and Interactive Editing

Various 3D representations are suitable for different jobs: meshes is easy to edit and output, 3D Gaussian representation is suitable for fast rendering, and radiance fields is suitable for synthesizing continuous new viewpoints. TRELLIS proposed a unified and structured latent space, and used different decoders to output messages, three-dimensional Gaussian and radial fields^[13]. This design allows platforms to share the same generation backend and select different output formats according to the needs of the APP, thus reducing additional deployment.

3D systems is expected to become editable many times, not just once. In the structure of a small part of TRELLIS, users can delete, add or exchange specific locations. You can also modify materials and details without changing the overall shape^[13]. If text, referenced image and partial position are used together, it may reduce the time of remaking the whole model many times, thus improving the usability of 3D generation platforms.

5.4 Production-level Assets and Automated Quality Evaluation

In the appearance made by the current method, sometimes the fixed lighting, shadow and highlight from the input image are still retained. TRELLIS pointed out that in the process of converting image into 3D, the influence of light was not completely distinguished, and suggested to predict PBR materials as the future development direction^[13]. Future models need to be able to distinguish the information of shape, basic color, roughness, metallic feeling and surface direction in more detail. In this way, the products can adapt to various lighting conditions in the new game engine and 3D design environment.

The method of automatic evaluation of quality is not only a form of distance, but also should be extended to evaluate from all aspects. Platform combines mesh shape rules, multi-view rendering indicators and multi-models, and can comprehensively evaluate the matching degree with conditions, naturalness, fineness, floating structure, texture and shape. The evaluation results can be used to determine the order of candidates, automatically re-create, modify parts or tell users to make other inputs. This reduces the trouble for people to check by hand.

5.5 From Single-object Generation Toward Scene-level Content Generation

If the AI quality of making a thing becomes better, the research goal may also be extended to a 3D scene composed of many materials. The generation of scene level should not only make everything, but also consider the difference of size, configuration, contact mode, invisible part and unity of style. TripoSG takes scene generation as the future direction, and TRELLIS shows the application examples of combining many materials to make building structures and street scenes^[12, 13]. If it is in this direction, the technology of making 3D from a painting may evolve from a tool for making materials to the structure of making games, virtual environment and digital twins scenes.

6. CONCLUSION

Single-image 3d reconstruction has changed from popular expressions such as voxels, point clouds and template-based meshes to continuous implicit surfaces, multi-view diffusion methods, and large feed-forward reconstruction models, and then it is gradually used with the general-purpose 3D basic model. The experimental results of LRM, TripoSR, and CRM demonstrate the practical feasibility of sub-second previews and second-level interactive mesh generation. However, the timing scopes and output specifications reported in different studies are not consistent, and the runtimes reported in papers

cannot be directly regarded as the complete user waiting time in practical platforms. Real-world deployment still needs to address challenges including input normalization, computational resource requirements, uncertainty in invisible regions, mesh repair, and material generation.

The methods of engineering APP should not only consider a single precision index, but also balance the generation quality, delay, GPU memory consumption, output editability and platform integration cost. Future research may focus on high-quality data governance, sparse large models, hierarchical generation, unified 3D latent space, local interactive editing, PBR production-level assets and automated quality assessment. In this way, 3D reconstruction of a single image can be promoted from research prototype to stable content production service.

REFERENCES

- [1] Choy, C. B., Xu, D., Gwak, J., Chen, K., & Savarese, S. (2016, September). 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision (pp. 628-644). Cham: Springer International Publishing
- [2] Fan, H., Su, H., & Guibas, L. J. (2017). A point set generation network for 3d object reconstruction from a single image. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 605-613)
- [3] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., & Jiang, Y. G. (2018). Pixel2mesh: Generating 3d mesh models from single rgb images. In Proceedings of the European conference on computer vision (ECCV) (pp. 52-67)
- [4] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A. (2019). Occupancy networks: Learning 3d reconstruction in function space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4460-4470)
- [5] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99-106
- [6] Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4), 1-15
- [7] Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., & Vondrick, C. (2023). Zero-1-to-3: Zero-shot one image to 3d object. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9298-9309)
- [8] Long, X., Guo, Y. C., Lin, C., Liu, Y., Dou, Z., Liu, L., ... & Wang, W. (2024). Wonder3d: Single image to 3d using cross-domain diffusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9970-9980)
- [9] Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., ... & Tan, H. (2024, May). Lrm: Large reconstruction model for single image to 3d. In International Conference on Learning Representations (Vol. 2024, pp. 50678-50702)
- [10] Tochilkin, D., Pankratz, D., Liu, Z., Huang, Z., Letts, A., Li, Y., ... & Cao, Y. P. (2024). Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*
- [11] Wang, Z., Wang, Y., Chen, Y., Xiang, C., Chen, S., Yu, D., ... & Zhu, J. (2024, September). Crm: Single image to 3d textured mesh with convolutional reconstruction model. In European conference on computer vision (pp. 57-74). Cham: Springer Nature Switzerland
- [12] Li, Y., Zou, Z. X., Liu, Z., Wang, D., Liang, Y., Yu, Z., ... & Cao, Y. P. (2025). Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- [13] Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., ... & Yang, J. (2025). Structured 3d latents for scalable and versatile 3d generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 21469-21480)