ACT-R: Adaptive Camera Trajectories for Single-View 3D Reconstruction

Yizhi Wang*[1] Mingrui Zhao*[1] Hao Zhang[1]

[1]Simon Fraser University

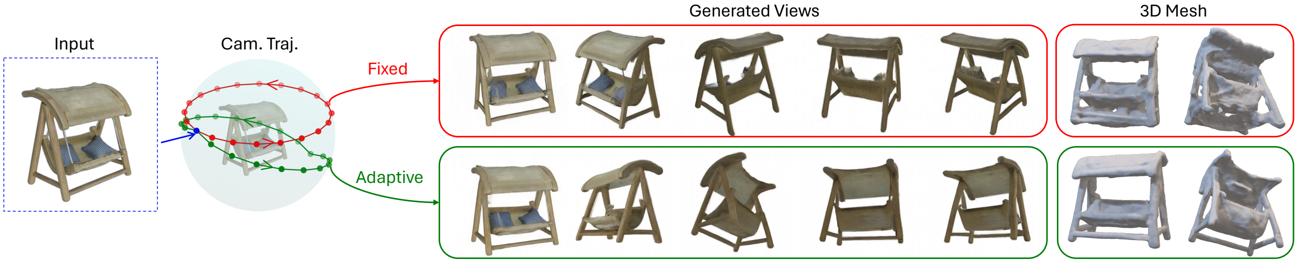

Figure 1. ACT-R, for single-view 3D reconstruction, predicts an *adaptive camera trajectory* (green) to maximize the visibility of occluded object parts over a fixed sequence length (20 views). The trajectory, obtained in under 10s, is then used by a video generator (e.g., SV3D [59]) to produce a *sequence* of novel views for multi-view 3D reconstruction, here using NeUS [61]. Compared to a generic trajectory (red), at fixed elevation, ACT-R yields much cleaner results with more faithful recovery of occluded regions.

Abstract

We introduce the simple idea of adaptive view planning *to multi-view synthesis, aiming to improve both occlusion revelation and 3D consistency for single-view 3D reconstruction. Instead of producing an unordered set of views independently or simultaneously, we generate a* sequence *of views, leveraging temporal consistency to enhance 3D coherence. Importantly, our view sequence is not determined by a pre-determined and fixed camera setup. Instead, we compute an* adaptive camera trajectory *(ACT), to maximize the visibility of occluded regions of the 3D object to be reconstructed. Once the best orbit is found, we feed it to a video diffusion model to generate novel views around the orbit, which can then be passed to any multi-view 3D reconstruction model to obtain the final result. Our multi-view synthesis pipeline is quite efficient since it involves no run-time training/optimization, only forward inferences by applying pre-trained models for occlusion analysis and multi-view synthesis. Our method predicts camera trajectories that reveal occlusions effectively and produce consistent novel views, significantly improving 3D reconstruction over SOTA alternatives on the unseen GSO dataset.*

1. Introduction

Single-view 3D reconstruction has been one of the most intensively studied problems in computer vision. One class of modern approaches directly generate or regress 3D representations of objects from input images [28, 57, 69, 70, 76, 79], often resorting to 3D supervision which requires large 3D datasets for training. With significant advances in novel view synthesis [22, 40], the second line of popular approaches to single-view 3D reconstruction first perform a multi-view synthesis, which typically generates an unordered set of views either *independently* [31] or *simultaneously* [33, 35, 50, 55] with *fixed camera setups*. This is followed by multi-view 3D reconstruction via differentiable rendering (e.g., NeUS [61] and variants [34, 60]) so that the entire solution pipeline can avoid direct 3D supervision.

The main challenges to multi-view synthesis are twofold. First, the produced images should reveal *occluded* structures of the target 3D object that are hidden from the input image. Second, the generated views must attain *3D consistency* to ensure that a plausible and coherent 3D model can be reconstructed. Upon close examination of state-of-the-art multi-view synthesis methods, as well as single-view reconstruction methods which combine such syntheses with direct 3D prediction [24, 30, 53, 65], we find that there is still much room for improvement.

In this paper, we introduce *adaptive view planning* to multi-view synthesis, so as to improve both occlusion revelation and 3D consistency for single-view 3D reconstruction. Instead of synthesizing an unordered set of views as in prior works, we generate a *sequence* of views, leveraging the inherent temporal consistency to enhance 3D coherence. More importantly, our view sequence is not constructed by a

*Equal contribution.

We use the term "direct 3D supervision" in the context of 3D reconstruction to refer to models trained using paired images and ground-truth 3D models.

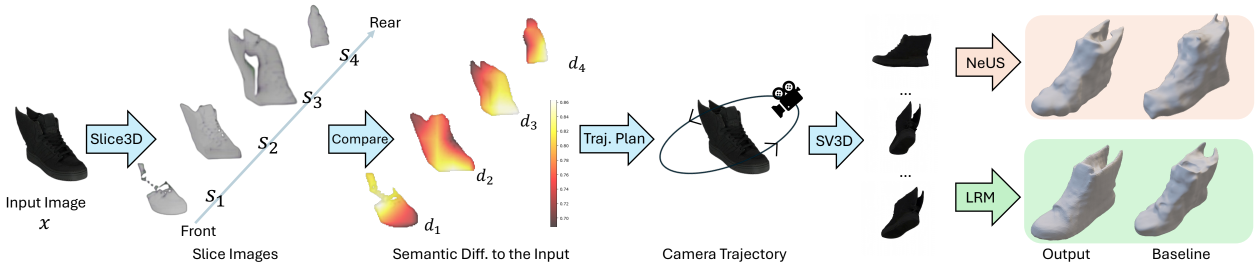

Figure 2. Pipeline of our single-view reconstruction method, ACT-R, with adaptive camera trajectories (ACT). We first employ Slice3D [63] to produce the slice images of the input object, with the slicing direction from the camera to the object center. Then we compute the semantic difference between the input and its slices by comparing their $512 \times 7 \times 7$ feature maps extracted from VGG16 [51]. Each difference map $d_i \in [0, 1]^{7\times 7}$ is up-scaled and overlaid onto slice images for beter visualization. Next, we identify the regions that have significant semantic differences (Sec. 3.2), and plan the camera trajectories based on them (Sec. 3.3). Finally, we condition SV3D [59] on our planned trajectories, yielding a sequence of views, which can be fed into NeUS [61] or InstantMesh (IM) [72] for multi-view 3D reconstruction.

pre-determined setup [33, 35, 39, 55, 73]. Instead, we compute an *adaptive camera trajectory* (ACT), which is specific to the target 3D object and its input view, to generate the sequence of novel views. Note that several recent works have shown benefits of camera perturbation [59] and stochastic conditioning [66] for novel view synthesis. Our approach takes these further with *judicious view planning*.

Given a single-view image of an object, we first search for an *orbit* of camera views to maximize the visibility of its occluded regions. Since occlusion prediction from a single image is ill-posed and searching over all orbits is intractable, we resort to a heuristic sampling of a manageable number of candidate orbits and utilize a neural model to analyze occlusion revelation by the camera orbits.

As our neural model, we employ Slice3D [63], a recent method for single-view 3D reconstruction which excels at recovering occluded 3D object structures. We apply a pretrained Slice3D to predict a stack of images capturing parallel volumetric slices of the 3D object in the input image. We then rank the candidate camera orbits based on how well they reveal occluded regions of the 3D object, which can be localized over the slice images by examining semantic differences between them and the input image.

Once the best orbit is found, we feed it to a video diffusion model to generate a sequence of novel views that are adapting to the 3D object. The multi-view images obtained are finally passed to a 3D reconstruction model to obtain the final result. Note that these last two steps can employ a variety of state-of-the-art video diffusion and multi-view 3D reconstruction models. In our current work, we employ Stable Video 3D (SV3D) [59] for the former, while for the latter, either NeUS [61], a well established method, or InstantMesh (IM) [72], a more recent one based on large reconstruction models (LRMs). As a result, our entire solution pipeline (see Fig. 2), which is coined ACT-R for using Adaptive Camera Trajectory for single-view 3D Reconstruction, does *not* use direct 3D supervision since none of Slice3D, NeUS, or InstantMesh does. Also, our multi-view generation is quite efficient since it involves no run-time training or optimization, only forward inferences by applying the pre-trained Slice3D and SV3D.

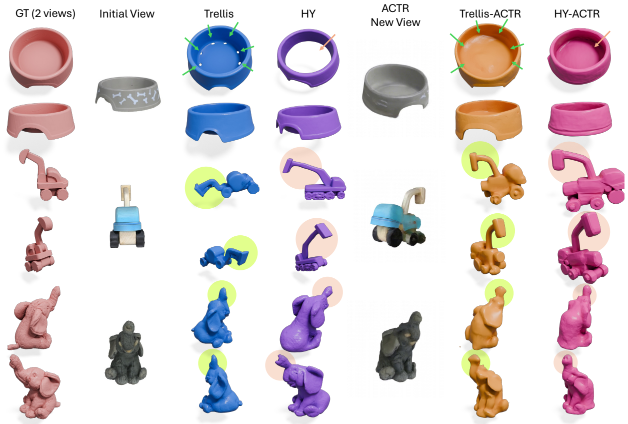

Figure 3. Initial single-view inputs to advanced methods such as Trellis and Hunyuan3D (HY) can be improved by one of the synthesized multi-view images from ACT-R, to improve reconstruction results, as highlighted in the colored regions.

Extensive experiments demonstrate that our method predicts camera trajectories tailored to each input example, effectively revealing occluded regions for higher-quality reconstruction. On the GSO benchmark [9], ACT-R outperforms, both qualitatively and quantitatively, state-of-the-art methods based on multi-view syntheses from static [59], fixed [35, 53], or randomly perturbed [59] camera trajectories, all without direct 3D supervision, as well as a recent method, Craftsman [25], which employs 3D supervision in the multi-view reconstruction step. On the other hand, ACT, which is best positioned to boost multi-view 3D reconstruction *without* direct 3D supervision, cannot beat some of the latest LRMs such as CLAY [76], Hunyuan3D [57], or Trellis [70], which all use direct 3D supervision trained by closed-source (CLAY) or carefully curated (Trellis) large-scale 3D assets. Directly comparing ACT-R with these models can be unfair due to the different training setups. Nevertheless, as shown in Fig. 3, ACT can benefit these supervised LRMs by improving the input views.

2. Related Work

As single-view 3D reconstruction has been extensively studied, we only focus on closely related works on video generation, multi-view synthesis, and view planning.

2.1. Video Generation in 2D and for 3D

Early works [2, 11, 15, 16, 52, 67, 78, 80] extend image diffusion to video generation. Stable Video Diffusion (SVD) [3] adapts latent diffusion methods [4, 48] to large-scale video datasets for temporally coherent generation. A stream of work focuses on keyframe conditioning [12, 26, 62, 75], where initial frames are generated to anchor subsequent video synthesis, with latent-consistency networks ensuring temporal and appearance coherence. Training-free approaches [17, 19] utilize Large Lanague Models (LLMs) for generation guidance. Although video generation models typically lack explicit 3D representations, they can still achieve 3D consistency by producing temporally coherent videos. In our approach, we use video diffusion models as the backbone to generate view sequences.

The temporal coherency achieved by video generation models such as SVD [3] can be utilized to enhance 3D consistency. CameraCtrl [14] trains a camera encoder upon a pre-trained T2V (Text-to-Video) model (such as [13]) to allow precise and customizable camera pose control. ViewCrafter [74] uses partial point clouds to render images as the condition of video diffusion models. CamCo [71] employs Plücker coordinates to control camera poses and leverages epipolar constraints to enhance 3D consistency in generated videos. Finally, SV3D [59] fine-tunes a video generation model [3] to create orbital videos around a 3D object given a camera trajectory. By adding random perturbations to the camera orbit, it can reveal structures that a standard orbit cannot. However, these randomized orbits do not reliably reveal additional occlusions. As more views are generated, the process slows down considerably.

2.2. Multi-view Synthesis from Single View

Multi-view images are commonly adopted as an intermediate representation between single view and 3D, by which novel views are first synthesized from the input view, and then 3D entities are reconstructed via optimization [35, 59] or feed-forward networks [24, 53, 72]. Zero-1-to-3 [31] fine-tunes Stable Diffusion (SD) [49] to generate novel views conditioned on camera poses. One-2-3-45 [29] and its upgraded version [28] combines such multi-view image generator with a feed-forward network, achieving reconstruction speeds as few as 45 seconds. Such a pipeline is later improved in Instant3D [24], InstantMesh [72], and LGM [53] for faster and higher-quality inference by leveraging large reconstruction models (LRM). SyncDreamer [33] enhances 3D consistency through cross-attention mechanisms between different views. Wonder3D [35] produces depth and normal maps alongside with the RGB novel views to further support accurate 3D reconstruction. MVDiffusion [54] and MVDiffusion++ [55] utilize cross-view attention to improve multi-view consistency for generating panoramas and 3D structures.

Generating multiple views from a single image in one step presents significant consistency challenges, as establishing correspondences between substantially different viewpoints remains difficult. To address this issue, 3DiM [66] generates novel views in an auto-regressive manner, selecting a previously generated view as a condition for producing each subsequent view during the denoising process. ViewFusion [73] builds upon Zero-1-to-3 [31] to generate novel views using a similar auto-regressive approach as 3DiM [66]. IM-3D [39] employs a video generation model [12] to create novel views that are then processed by 3D G-Splat [22] for 3D reconstruction. This approach can be iteratively refined by feeding rendered objects back into the video diffusion model. However, these methods (3DiM, ViewFusion, IM-3D) rely on either random camera poses or predefined trajectories for novel view generation, without considering the specific structural properties of the objects being modeled.

2.3. View and Path Planning

Path planning has applications in navigation, scanning, and even computational fabrication [10, 21, 32, 36, 46, 47]. In particular, Next-Best-View (NBV) planning addresses the fundamental challenge of determining an optimal sequence of camera positions to maximize information gain during scene or object inspection [8, 38, 58]. Two popular traditional approaches to this problem include voxel-space methods that optimize coverage metrics [8, 37, 38, 58], and surface-based methods that analyze boundary characteristics to determine optimal viewpoints [6, 23, 43]. Recent advances in deep learning have transformed the NBV paradigm, introducing reinforcement learning [7, 42], reconstructability predictor [32] and uncertainty evaluation framework [20] to the scope. While traditional NBVs assume the availability of a complete 3D reference model, our work addresses a more challenging scenario where only a single image serves as input. This constraint fundamentally shifts the problem from pure coverage optimization to view prediction based on limited initial information, requiring novel strategies for trajectory planning.

3. Method

Given a single image $x \in \mathbb{R}^{3\times H\times W}$ of an object and a video generation model $\mathcal{G}$ conditioned on a camera trajectory (e.g., SV3D [59]), our goal is to determine an adaptive camera pose trajectory $(\pi_1, \pi_2, ..., \pi_N)$ that adapts to the object in its input view. It should condition $\mathcal{G}$ to generate a sequence of views that better reveals the geometry of the target object compared to a fixed generic trajectory, and leads to a more accurate 3D reconstruction.

Similar to SV3D [59], we assume that the camera always looks at the center of an object (origin of the world), and the distance between the camera and the center remains unchanged, so any viewpoint can be specified by only two parameters: $\pi_i = (e_i, a_i)$, where e_i, a_i are the elevation and azimuth angles.

When designing an adaptive trajectory in contrast with a generic one, we aim to avoid increasing its length (i.e., the number of views), as doing so would complicate both the multi-view generation and subsequent 3D reconstruction. To maintain consistency, we fix the number of generated views, N, as 21 and use a single closed orbit for the camera trajectory, the same as SV3D (u) [59] for a fair comparison as SV3D serves as our closest baseline.

3.1. Overview

Fig. 2 shows our pipeline. To guide camera trajectory generation, we first identify occluded regions from the input view. To this end, we employ Slice3D [63] to produce object slice images from the front to the rear of the object. These slices allow us to construct a coarse representation by voxelizing each slice and stacking them up. When comparing each slice image to the input image, significant semantic differences in certain areas often suggest that these regions in the slice are occluded from the input view. Therefore, we compute semantic difference maps between each slice image and input image by leveraging VGG16 features and incorporating this information into the voxels, forming a series of spatially-aware 3D semantic difference blocks.

From these blocks, we plan a camera trajectory aiming to maximize the visibility of the occluded regions, with each block weighted by its semantic difference. This optimized trajectory is fed into SV3D [59] to generate a sequence of novel views. The generated novel views can be used in a plug-and-play manner for any 3D reconstruction pipeline that takes novel views as intermediate representations. We used NeUS [61] and large reconstruction models trained from InstantMesh [72] as our two alternative 3D reconstruction backbones.

3.2. Building Semantic Difference Blocks

As illustrated in the first step in Fig. 2, Slice3D produces in total M slice images $\{s_1, s_2, \ldots, s_M\}$ from the input x. We quantify the semantic differences between slice image s_i and input x by comparing the features from the final pooling layer of VGG16 [51]. Specifically, we compute: $d_i = \langle \phi(x), \phi(s_i) \rangle$, where $\phi(\cdot) \in \mathbb{R}^{512\times7\times7}$ represents the VGG16 feature maps for a given input, with spatial resolution 7×7 and feature dimension 512 per position; $1 \leq i \leq M$; $\langle \cdot, \cdot \rangle$ denotes the cosine similarity between the 512-dimensional feature vectors at each corresponding spatial position from the two inputs, resulting in $d_i \in \mathbb{R}^{7\times7}$.

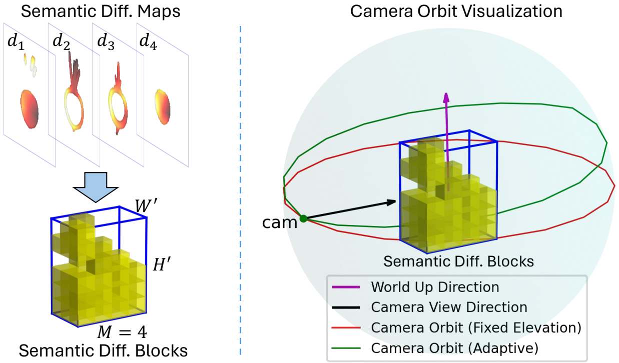

Figure 4. Illustration of camera trajectory planning. Left: Transforming the semantic difference maps into 3D blocks, where lighter yellow indicates greater differences. Right: Visualization of different camera orbits. Red: fixed elevation; Green: variable elevations that capture greater visibility. "diff" and "cam" denote difference and camera, respectively.

As shown in Fig. 4 (left), we transform the semantic difference maps $\{d_1, ..., d_M\}$ into 3D blocks (shown in yellow), approximating the object shape and indicating semantic differences to the input view. Specifically, we first identify the 2D mask of the object in the input image, and apply this mask to crop each map d_i into $d'_i \in \mathbb{R}^{H'\times W'}$ ($H' \leq 7$, $W' \leq 7$). Next, we estimate the object's 3D bounding box in the camera coordinate system (detailed in the supp. material) and divide this space into $M \times H' \times W'$ blocks. Each block at position (i, j, k) corresponds to the element d'_{ijk}, where $1 \leq i \leq M$, $1 \leq j \leq H'$ and $1 \leq k \leq W'$.

We retain a block only when its corresponding slice pixel s_{ijk} is not empty. The retained block is associated with the value of d'_{ijk} indicating the semantic difference of that region compared to the input view. This process yields a coarse 3D representation of the input object, highlighting which parts were obstructed from the input view and require further observation from subsequent views.

Compared with more recent feature encoders such as DINOv2 [41] and CLIP [45], we empirically find that VGG16 [51] works better in extracting semantic difference maps between slices and input images. Please see the supplementary materials for detailed ablation studies.

3.3. Camera Trajectory Planning

To accurately model the spatial relationship between the camera and the reconstructed object, we estimate:

1. r: distance from the camera to the object center, with the camera trajectory as an orbit on a sphere with radius r.
2. α: the elevation over the object in the input image.

The camera is initially positioned at the input view, i.e., $(a_0, e_0) = (0, \alpha)$. The generic camera trajectory from SV3D (u) [59] uses a fixed elevation α, shown as the red orbit in Fig. 4 (right). Here r and α are used to calculate the world-to-camera coordinate transform since SV3D [59] operates in a world frame while we use a camera coordinate system.

The reconstructed coordinate system is shown in Fig. 4. To maximize the observation of occluded regions, an ideal trajectory should aim to cover as many blocks as possible, prioritizing those blocks with greater semantic differences (denoted by lighter-coloured blocks). With the camera's known field of view (FOV) θ, we can determine the block visibilities from any given position.

Since the search space for camera trajectories is infinite, we discretize the camera's movement to facilitate the search. Specifically, we begin by sampling the azimuth steps at a constant interval of $18°$ (calculated as $360°/20$) between each frame from $0°$ to $360°$, which creates a closed-loop trajectory in terms of azimuth.

Next, we define the elevation angle changes at each step using the set $\{\pm5°, \pm4°, \pm3°, \pm2°, \pm1°, 0°\}$. We limit per-step elevation changes to within $5°$, as larger step sizes challenges frame-consistent video generation. Each orbit is divided into four segments based on azimuth angles. Within each segment, the elevation angle increments at a constant rate, with the step size selected from the set. To ensure a closed orbit, we enforce the total variation in elevation to be zero by mirroring and negating the elevation change in the second segment onto the third and the first onto the fourth. This approach results in a total of $11 \times 11 = 121$ candidate trajectories, which we denote as the set Π.

We choose the path π^* that maximizes the weighted visibility of difference blocks. For each camera position, we determine which blocks are visible, and our objective is to optimise culmulative visibility weight across all time steps:

$$\pi^* = \arg\max_{\pi \in \Pi} \sum_{t=1}^{N} \sum_{\substack{(i,j,k) \\ \in \psi(\pi(t))}} d'_{ijk}, \tag{1}$$

where $\psi(\pi(t))$ denotes the set of all visible blocks under camera $\pi(t)$. More details about the $\psi(\cdot)$ formulation and orbital camera trajectory justification can be found in the supplementary material.

3.4. View Generation and 3D Reconstruction

We fed the camera trajectory π^* to SV3D [59] to generate a video containing a sequence of novel views. Since the video could suffer from significant artifacts due to stochasticity in the generative models, we apply view-consistency as the primary metric to filter out low-quality results and regenerate with alternative random seeds when necessary. The final view sequence can be integrated into any 3D reconstruction pipeline that accepts posed-multiview-images as input. We demonstrate ACT-R's flexibility by reconstruction meshes through two different approaches: volumetric rendering with NeUS [61] using all 21 generated images, and processing through LRM [72] by feeding 6 uniformly sampled key frames from the 21 views.

4. Experiments

Implementation details. We used `rembg` [44] to remove the background of input images. The number of slices M is set to 4. The FOV θ of the camera is set to 33.8. SV3D produces a video consisting of 21 frames, with each frame at a resolution of 576×576 pixels. For volumetric rendering based reconstruction, we chose NeUS [61] as our reconstruction method because it is well-established and continues to be used in recent state-of-the-art methods (e.g., Wonder3D [35]). Our 3D reconstruction performance could potentially be enhanced with more recent 3D reconstruction methods, such as those in [18, 27]. We train the network of NeUS for 10k steps for each shape, which takes around 15 minutes in an NVIDIA 3090 GPU. For LRM based reconstruction, we uniformly sampled 6 views from the view sequence used pretrained checkpoint from InstantMesh [72] as it takes images from arbitrary camera pose.

Datasets. We conduct evaluation on GSO [9], a widely-used benchmark for novel view synthesis and 3D reconstruction which includes about 1K common household objects that were 3D scanned and represented as meshes.

Evaluation Metrics. For 3D metrics, we use Chamfer $\mathcal{L}_2$ (CD), Hausdorff (HD), and Light Field distances (LFD) [5], as well as F-score%1 (F1) [56], to evaluate reconstruction results. In addition, we also report 2D metrics such as PSNR, SSIM [64], and LPIPS [77] on 12 rendered views of the reconstructed meshes. Since no single metric is entirely informative, we provide a comprehensive evaluation to cover both object- and image-space, as well as distortion and visual similarity aspects of the 3D reconstruction.

4.1. Qualitative Visual Comparisons

We compare ACT-R to representative SOTA approaches for single-view 3D reconstruction, including Wonder3D [35], SV3D(u) [59] for volumetric-rendering-based reconstruction, LGM [53] based on LRMs, and CraftsMan [25] as an image-to-3D method with direct 3D supervision. We further present three ablated experiment setups:

1. Random trajectory: Generate novel views with a random orbital trajectory by using random elevation increments, with the mesh reconstructed by NeUS.
2. SV3D$_{\text{IM}}$: LRM with 6 views sampled uniformly from generic camera trajectories with constant elevation.

Qualitative results in Figure 6 show that different reconstruction backbones offer distinct advantages, while also

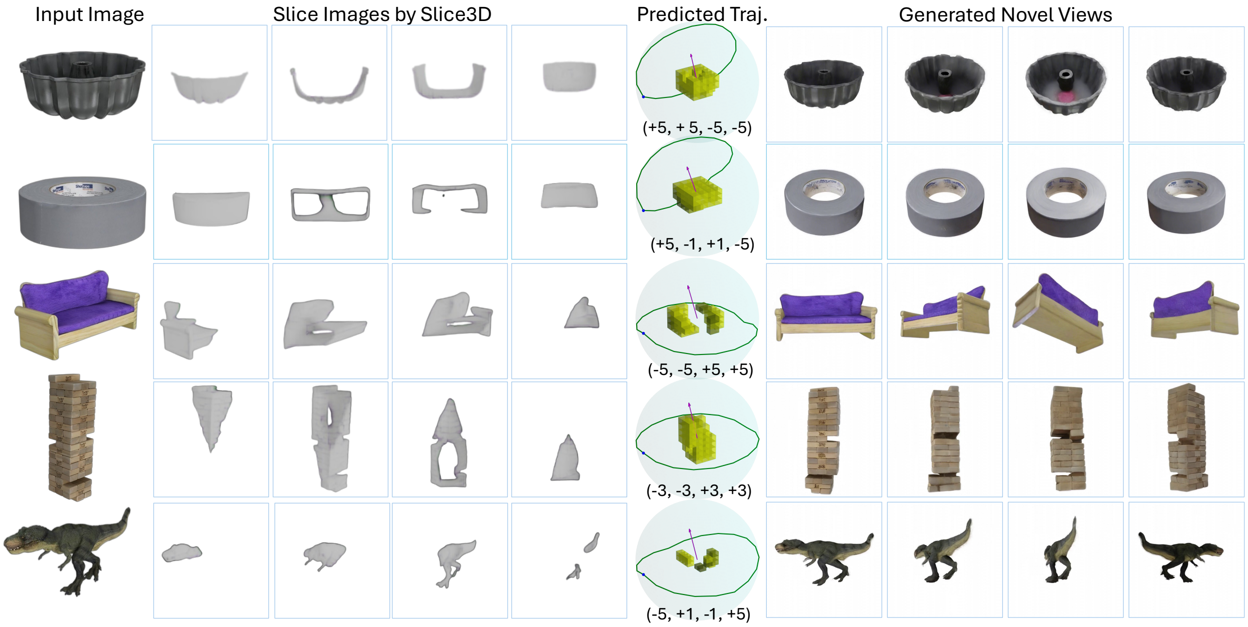

Figure 5. Visualization of slice images predicted by Slice3D [63] and planned trajectories (shown in green). Purple arrows indicate the world's up direction. The 3D blocks (i.e., semantics difference blocks) roughly represent the input shape, using colors to highlight their semantic differences from the input view. The numbers in brackets show the elevation changes in each segment of the orbit.

Method	CD↓	F1↑	HD↓	PSNR↑	SSIM↑	LPIPS↓	LFD ↓
Wonder3D [35]	5.17	2.66	18.9	15.8	8.17	17.9	1.33
SV3D(u) [59]	4.93	2.79	18.4	15.9	8.12	17.6	1.30
LGM [53]	7.78	1.53	28.9	13.3	7.56	24.1	1.79
Craftsman [25]	6.37	2.81	20.8	15.7	8.08	17.4	1.42
SV3D$_{\text{IM}}$	5.15	2.64	18.0	16.4	8.35	16.1	1.17
Random Traj.	4.79	3.15	17.2	16.8	8.32	15.5	1.04
Ours$_{\text{IM}}$	4.51	3.41	**16.2**	**17.4**	**8.49**	**14.0**	1.05
Ours	**4.47**	**3.78**	17.5	17.1	8.35	15.1	**1.00**

Table 1. Comparison of 3D reconstruction results. Cell colors indicate ranking: green (1st), blue (2nd), and amber (3rd) for each metric. Lower is better for CD, HD, LPIPS, and LFD. Higher is better for F1, PSNR, and SSIM.

expose their limitations. NeUS produces more faithful reconstructions that closely match the input, yet suffers from blobby surface artifacts. Effectively leveraging 3D priors, LRMs produce smoother surfaces, but respect less the details from the input image. Despite these inherent backbone-specific characteristics, ACT-R's adaptive trajectories *consistently* improve reconstruction quality, particularly over occluded regions. Our approach more faithfully reconstructs the solid bowl interiors (rows 2) that most competing methods overlooked due to occlusion in the input image and provides enhanced spatial awareness between objects (row 3). See supplementary material for more results.

Fig. 8 compares the generated views of Wonder3D [35] and SV3D [59] to ours. Compared to SV3D, Wonder3D offers a wider range of visibility (e.g., bottom views of the objects) by varying both camera azimuths and elevations. However, it tends to suffer from multi-view inconsistency, as seen in the heavily distorted back view. SV3D maintains better consistency but cannot observe the back of the object, leading to inadequate information during 3D reconstruction. In contrast, our method captures the bottom of the object and achieves view consistency.

4.2. Quantitative Comparisons

Since the reconstructed 3D meshes from different methods exhibit different poses, we first normalize the output mesh and provide an initial transform to align it with the GT mesh, then optionally used Iterative Closest Point (ICP) to further tune the mesh poses, whichever leads to better quantitative measures for the results.

We report quantitative comparison results for all 1,030 objects from the GSO dataset in Table 1. Our method demonstrates clear advantages over the other methods across all metrics, especially in F1 [56]. The improvements on 2D metrics indicate that our method can generate objects that are aligned with the GT in pixel space. Overall, our method exhibits robustness in accurately modeling object geometry while preserving visual realism in image-space projections. Its superior performance on GSO further underscores its strong generalizability across diverse object categories, showcasing its capability in handling a wide range of shapes, sizes, and appearances effectively.

We further compared the coverage metric resulting from different trajectories. Quantitative numbers suggest that our

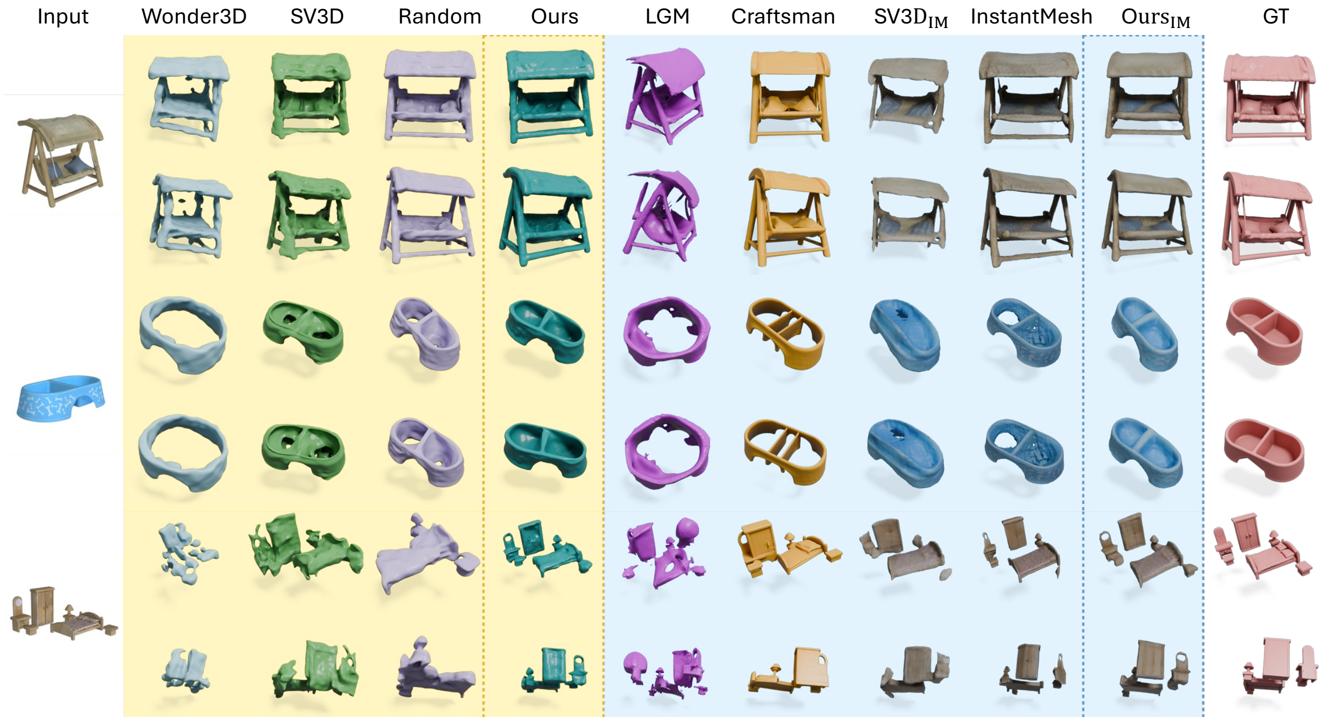

Figure 6. Qualitative visual comparisons between single-view 3D reconstruction methods on the GSO dataset. Please zoom in for a closer inspection. Meshes reconstructed with NeUS and LRM are in yellow and blue blocks, respectively. Our adaptive camera trajectories can be integrated into both reconstruction frameworks. From these examples, it is evident that our method can capture geometric and structural details better, especially over concavities and occluded regions, and it has less missing parts and geometric artifacts. For example, our method is the only one, whether with NeuS or LRM, that can faithfully reconstruct the inside region of the dog bowl (second example).

adaptive trajectory achieves a better coverage than static or random trajectories, please see the supplementary material for more details.

4.3. Visualizations of Trajectories

Fig. 5 shows our predicted slice images and trajectories. An naive view planning would negate the initial elevation angle for an even span, raising the camera for negative angles and lowering it for positive ones. In contrast, we base our trajectory prediction on a deeper understanding of object structures and occlusions.

For the Bundt cake pan, even with a positive elevation for the input view, we still raise the camera to reveal the inner tube. Although slice3D [63] failed to predict the inner tube in the sliced image, it still suggests that inner regions require additional attention. For the tape example, the camera highest elevation is lower that the pan example, since it has already fully observed the hole in the middle. This proves that our method is adaptive to each single object. For the sneaker example, our trajectory can better observe the pair of wings so that they will not occlude with each other in most of the frames. For the sofa, we lower the camera to fully capture its bottom. For the Jenga blocks, we planned a trajectory that can better observe the concave areas.

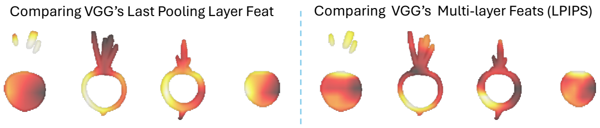
Comparing VGG's Last Pooling Layer Feat
Comparing VGG's Multi-layer Feats (LPIPS)

Figure 7. Computing semantic difference maps $\{d_i\}$ using different metrics. Brighter color indicates higher semantic difference.

4.4. Ablation Studies

Computation of semantic differences. The computation of semantic difference maps $\{d_i\}$ plays an critical role in our task. From the comparison between Ours and random trajectory result, it is evident that our current VGG features is guiding the camera in a meaningful way. Aside from VGG16, we also tried perceptual loss (LPIPS) [77] to compute the semantic differences by comparing multi-level features from VGG, rather than the last-pooling layer features we currently employ. As shown in Fig. 7, it appears that high-frequency (low-layer) features do not aid in locating the occluded regions. This is particularly evident in the third slice images, where LPIPS generates a dark map that barely highlights any differences from the input view.

Reconstruction from 6 views. Since Wonder3D only generates 6 views to obtain a 3D mesh, we also test our method

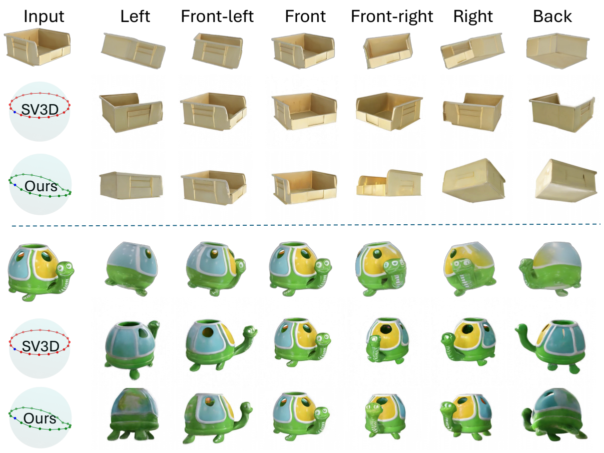

Figure 8. Generated views of Wonder3D [35] (1st row), SV3D [59] (2nd row) and our method (3rd row). The red and green orbits show the orbits of SV3D and our method, respectively.

to only rely on 6 frames from our generated videos that have nearly the same azimuths as Wonder3D. The results in Fig. 9 indicate that the quality of the views is more important than their quantity.

Robustness to camera pose estimation. Since our trajectory planning operates on a coarse 3D representation, i.e., the blocks, it is robust against errors from estimating the camera parameters and the bounding boxes. More details on how camera estimation affects our view planning can be found in the supplementary material.

5. Conclusion, Limitation, and Future Work

We propose adaptive view planning for synthesizing novel views from a single image. Our key insight is that slice images from Slice3D [63] can effectively reveal occluded structures from the input view and guide the camera's movement. By leveraging the capabilities of modern video generation models, the generated novel views along our planned trajectory tend to improve the visibility of occluded structures while maintaining multi-view consistency and the overall frame budget. Interestingly, this reveals how multi-slice and multi-view can work together to complement each other, where the former provides geometric insights on where the self-occlusion may be present while the later offers signals on the visible geometries. Combining the two, potential invisible regions are better exposed for improved single-view 3D reconstruction. In the end, our method ACT-R has been shown to outperform state-of-the-art reconstruction alternatives which take the single-to-multi view route without resorting to direct 3D supervision.

Limitations. Most limitations in our current implementation are inherited from SV3D [59], e.g., limited camera DoFs as only azimuth and elevation are changeable, and quality of the generated videos. Nonetheless, our view planning is applicable to other video generation models such as ViewCrafter [74], VD3D [1], and CameraCtrl [14] with even higher camera DoFs. ACT-R's performance is also limited by the accuracy of the estimated semantic differences between slice and input images. It is certainly possible to explore alternative occlusion-aware reasoning for view planning, or enlarge the scale of 3D data exposed to Slice3D — the version we employed for trajectory planning was trained on only 5% of the 880K Objaverse 3D dataset.

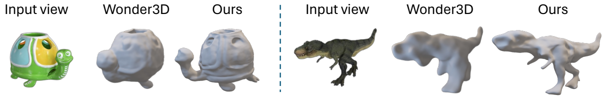

Figure 9. Reconstruction from 6 generated views.

Adaptive vs. fixed camera trajectories. While we firmly believe in the merits of adaptive view trajectories for occlusion revelation, view synthesis from such trajectories (e.g., via SV3D [59]) is more difficult compared to one that only works with fixed cameras, e.g., Zero123++ [50]. Even when trained on the same 3D dataset, SV3D is clearly outdone by Zero123++ in terms of the quality of the synthesized multi-view images. As a result, our method, which feeds adaptive trajectories to SV3D, cannot quantitatively beat InstantMesh [24], which employs Zero123++, across most metrics. Qualitatively however, we consistently observe that our method can outperform InstantMesh when the 3D objects have significant concavities or occlusions; see the two bowls in the fourth blue column in Fig. 6. We are motivated to resolve the above discrepancy by improving camera-adaptive view synthesis.

LRMs with direct 3D supervision. The most successful single-view reconstruction models of late have predominantly been LRMs [57, 70, 76] with direct image-to-3D and 3D supervision using large-scale 3D datasets. The strong 3D priors learned by these models appears to be diminishing the importance of multi-view synthesis. As we show in the supplementary material, multi-view Trellis does not outperform its single-view counterpart. This is also due in part to the blurriness of the synthesized multi-view images, regardless of the camera trajectories. That being said, hallucinations still abound either due to severe occlusions in the input view (see Fig. 3) or erroneous priors (e.g., symmetries by CLAY [76]), while potential overfitting to 3D training data remains a valid concern.

In future work, a more in-depth investigation into how to best integrate adaptive view planning into direct image-to-3D large reconstruction models (LRMs) is worth conducting. We would also like to explore other applications of ACT, e.g., for robotics and autoscanning [68].

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 8

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 3

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[5] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, pages 223–232. Wiley Online Library, 2003. 5

[6] SY Chen and YF Li. Vision sensor planning for 3-d model acquisition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(5):894–904, 2005. 3

[7] Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. Gennbv: Generalizable next-best-view policy for active 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16436–16445, 2024. 3

[8] Cl Connolly. The determination of next best views. In *Proceedings. 1985 IEEE international conference on robotics and automation*, pages 432–435. IEEE, 1985. 3

[9] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 2, 5

[10] Enric Galceran and Marc Carreras. A survey on coverage path planning for robotics. *Robotics and Autonomous systems*, 61(12):1258–1276, 2013. 3

[11] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 3

[12] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 3

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[14] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3, 8

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[17] Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. Large language models are frame-level directors for zero-shot text-to-video generation. In *First Workshop on Controllable Video Generation@ ICML24*, 2023. 3

[18] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 5

[19] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[20] Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11305–11312. IEEE, 2023. 3

[21] Karthik Karur, Nitin Sharma, Chinmay Dharmatti, and Joshua E Siegel. A survey of path planning algorithms for mobile robots. *Vehicles*, 3(3):448–468, 2021. 3

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, pages 1–14, 2023. 1, 3

[23] Simon Kriegel, Tim Bodenmüller, Michael Suppa, and Gerd Hirzinger. A surface-based next-best-view approach for automated 3d model completion of unknown objects. In *2011 IEEE International Conference on Robotics and Automation*, pages 4869–4874. IEEE, 2011. 3

[24] 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. *arXiv preprint arXiv:2311.06214*, 2023. 1, 3, 8

[25] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity

mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 2, 5, 6

[26] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 3

[27] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 5

[28] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024. 1, 3

[29] 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. 3

[30] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, et al. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024. 1

[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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 1, 3

[32] Yilin Liu, Liqiang Lin, Yue Hu, Ke Xie, Chi-Wing Fu, Hao Zhang, and Hui Huang. Learning reconstructability for drone aerial path planning. *ACM Transactions on Graphics (TOG)*, 41(6):1–17, 2022. 3

[33] 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. 1, 2, 3

[34] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. SparseNeuS: Fast generalizable neural surface reconstruction from sparse views. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 210–227. Springer, 2022. 1

[35] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1, 2, 3, 5, 6, 8

[36] Ali Mahdavi-Amiri, Fenggen Yu, Haisen Zhao, Adriana Schulz, and Hao Zhang. Vdac: volume decompose-and-carve for subtractive manufacturing. *ACM Trans. Graph.*, 39(6), 2020. 3

[37] Oscar Mendez Maldonado, Simon Hadfield, Nicolas Pugeault, and Richard Bowden. Next-best stereo: extending next best view optimisation for collaborative sensors. *Proceedings of BMVC 2016*, 2016. 3

[38] Nikolaos A Massios, Robert B Fisher, et al. *A best next view selection algorithm incorporating a quality criterion*. Citeseer, 1998. 3

[39] 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. *arXiv preprint arXiv:2402.08682*, 2024. 2, 3

[40] 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. 1

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[42] Daryl Peralta, Joel Casimiro, Aldrin Michael Nilles, Justine Aletta Aguilar, Rowel Atienza, and Rhandley Cajote. Next-best view policy for 3d reconstruction. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 558–573. Springer, 2020. 3

[43] Richard Pito. A solution to the next best view problem for automated surface acquisition. *IEEE Transactions on pattern analysis and machine intelligence*, 21(10):1016–1030, 1999. 3

[44] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020. 5

[45] 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. 4

[46] Mike Roberts and Pat Hanrahan. Generating dynamically feasible trajectories for quadrotor cameras. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 3

[47] Mike Roberts, Debadeepta Dey, Anh Truong, Sudipta Sinha, Shital Shah, Ashish Kapoor, Pat Hanrahan, and Neel Joshi. Submodular trajectory optimization for aerial 3d scanning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5324–5333, 2017. 3

[48] 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. 3

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[50] 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 preprint arXiv:2310.15110*, 2023. 1, 8

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4

[52] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[53] 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. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1, 2, 3, 5, 6

[54] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion, 2023. 3

[55] 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 preprint arXiv:2402.12712*, 2024. 1, 2, 3

[56] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, pages 3405–3414, 2019. 5, 6

[57] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025. 1, 2, 8

[58] J Irving Vasquez-Gomez, L Enrique Sucar, Rafael Murrieta-Cid, and Efrain Lopez-Damian. Volumetric next-best-view planning for 3d object reconstruction with positioning error. *International Journal of Advanced Robotic Systems*, 11(10): 159, 2014. 3

[59] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 1, 2, 3, 4, 5, 6, 8

[60] Aditya Vora, Akshay Gadi Patil, and Hao Zhang. DiViNet: Artistic typography via discriminated and stylized diffusion. In *Proc. of NeurIPS*, 2023. 1

[61] 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 preprint arXiv:2106.10689*, 2021. 1, 2, 4, 5

[62] Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, et al. Microcinema: A divide-and-conquer approach for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8414–8424, 2024. 3

[63] Yizhi Wang, Wallace Lira, Wenqi Wang, Ali Mahdavi-Amiri, and Hao Zhang. Slice3d: Multi-slice occlusion-revealing single view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9881–9891, 2024. 2, 4, 6, 7, 8

[64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[65] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 1

[66] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 2, 3

[67] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[68] Shihao Wu, Wei Sun, Pinxin Long, Hui Huang, Daniel Cohen-Or, Minglun Gong, Oliver Deussen, and Baoquan Chen. Quality-driven poisson-guided autoscanning. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 33: 203:1–203:12, 2014. 8

[69] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 1

[70] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, 2025. 1, 2, 8

[71] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3

[72] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 3, 4, 5

[73] Xianghui Yang, Yan Zuo, Sameera Ramasinghe, Loris Bazzani, Gil Avraham, and Anton van den Hengel. Viewfusion: Towards multi-view consistency via interpolated denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9870–9880, 2024. 2, 3

[74] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3, 8

[75] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 3

[76] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 1, 2, 8

[77] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 7

[78] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 3

[79] 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. *Advances in neural information processing systems*, 36:73969–73982, 2023. 1

[80] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3