Zero-1-to-3: Zero-shot One Image to 3D Object [1]

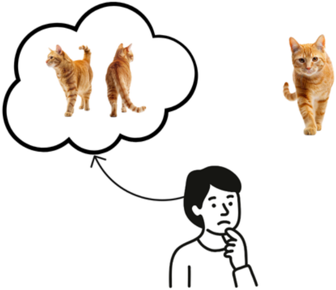
Nemanja Vujadinović Mohammed El Hassan Ayoubi
ENS Paris-Saclay

December 26, 2025

Outline

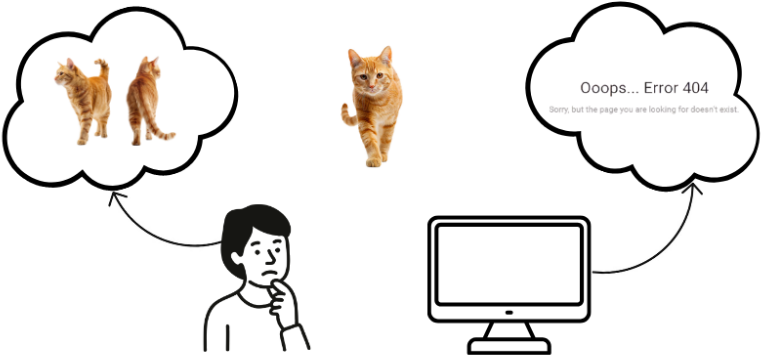
3D Shape and Appearance

- Humans can imagine 3D from a single view using prior experience
- Motivation:
  - Object manipulation
  - Navigation
  - Visual art & creativity



3D Shape and Appearance

- How do we model the same ability to machines?

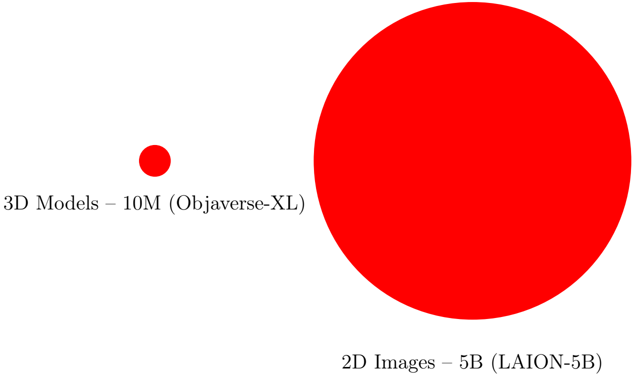


Traditional 3D reconstruction methods

- Closed-world models
- Limited scale & diversity
- Expensive 3D anottations (CAD)
- Geometry requirement (CO3D dataset [2])

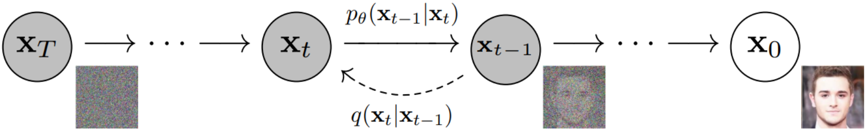
Size of Visual Datasets

- 3D datasets still much smaller than 2D datasets
- Can we exploit rich 2D pretrained models for 3D tasks?



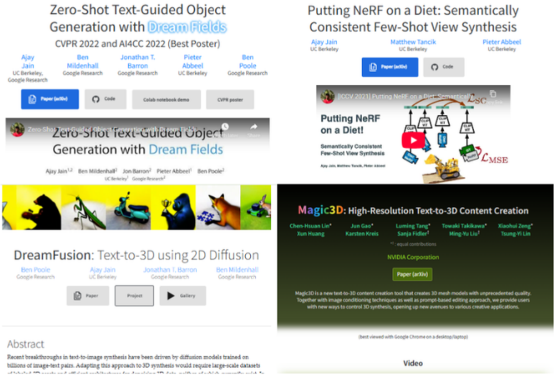
2D diffusion models

- Rich semantic priors
- High-fidelity synthesis
- Broad diversity of scenes & objects
- ... but comes with limitations:
  - No true geometric understanding
  - No 3D pose control
  - Canonical view bias



Transferring 2D Diffusion to 3D

- Naïve approach: scale 2D diffusion models to 3D domain
- Instead: reuse pretrained 2D diffusion models (+ NeRF)



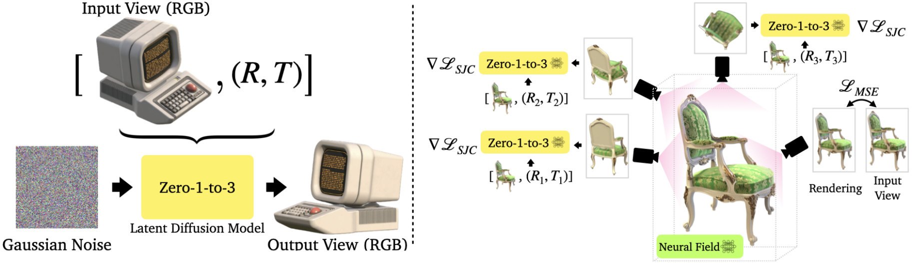
- Limitation of all: not designed for single-image novel view synthesis

Transferring 2D Diffusion to 3D from Single-View

- 3DiM [3]: Pose-conditioned image-to-image diffusion
  - Weak priors and low resolution images
  - No zero-shot generalization
- Mesh [4], voxel [5] or point cloud [6] methods
  - Require pose estimation
- Multiview Compressive Coding (MCC) [7]
  - Requires depth supervision

Zero-1-to-3 Approach

- Viewpoint-conditioned image-to-image translation
- Diffusion model trained with explicit camera control
- Strong semantic priors inherited from Stable Diffusion
- Learns geometric priors from synthetic multiview data
- No 3D supervision or depth required
- Zero-shot generalization to real-world images



Novel View Synthesis

3D Reconstruction

Objective

Given a single RGB image $x \in \mathbb{R}^{H \times W \times 3}$ of an object, goal is to synthesize an image of the object from a different camera viewpoint.

$$\hat{x}_{R,T} = f(x, R, T) \tag{1}$$

- $R \in \mathbb{R}^{3 \times 3}$, $T \in \mathbb{R}^3$

How to determine f ?

- Problem 1: No correspondences between viewpoints
- Problem 2: Biases in generative models

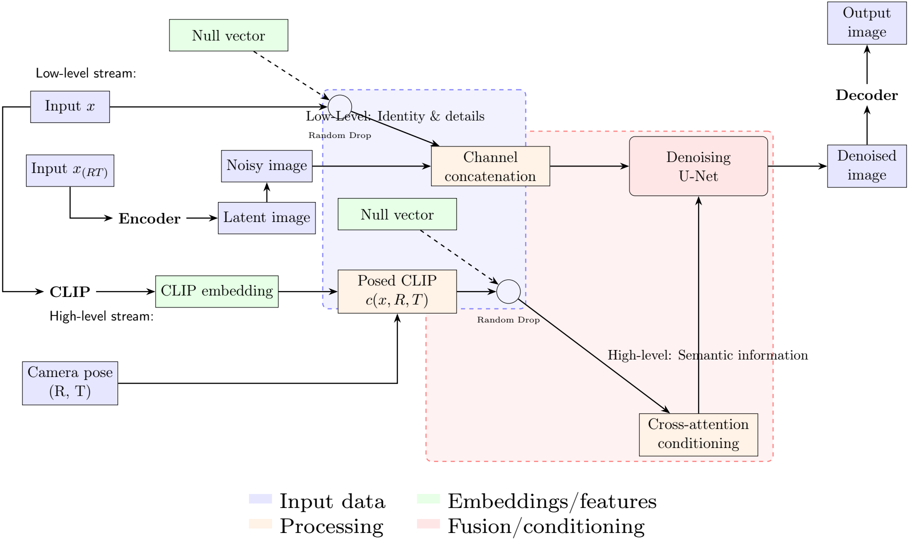
Learning to Control Camera Viewpoint

- $\{(x, x_{(R,T)}, R, T)\}$: image pairs with relative camera extrinsics
- We use a latent diffusion model with encoder $\mathcal{E}$, denoiser U-Net ϵ_θ, and decoder $\mathcal{D}$. During training, we minimize:

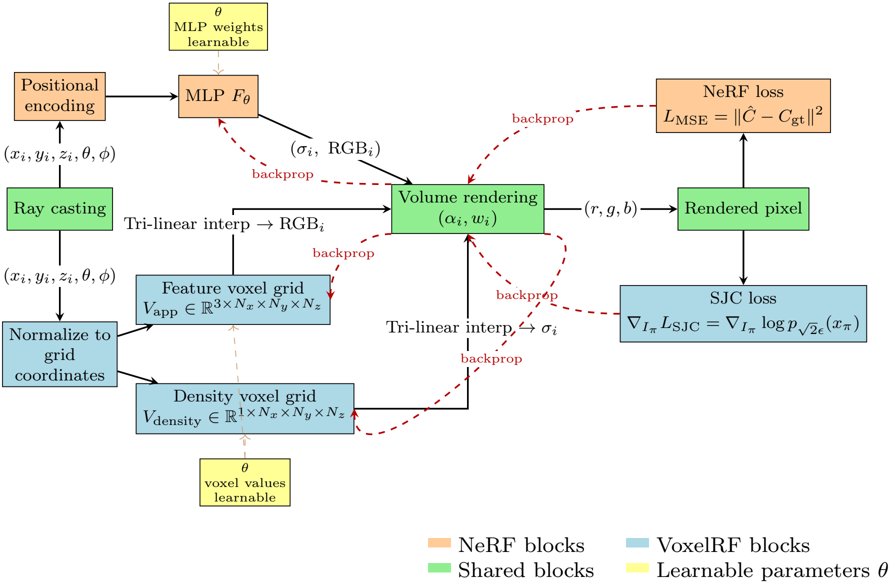
$$\min_\theta \mathbb{E}_{z\sim\mathcal{E}(x),\, t,\, \epsilon\sim\mathcal{N}(0,1)} \|\epsilon - \epsilon_\theta(z_t, t, c(x, R, T))\|_2^2,$$

- $c(x, R, T)$ is an embedding of the input view and camera extrinsics.

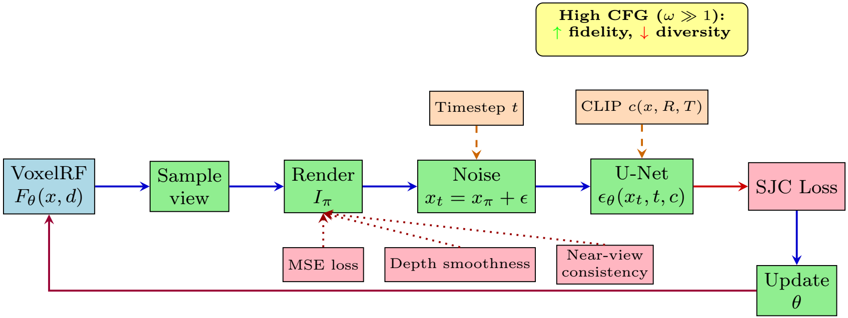
View-Conditioned Diffusion Architecture



NeRF vs VoxelRF

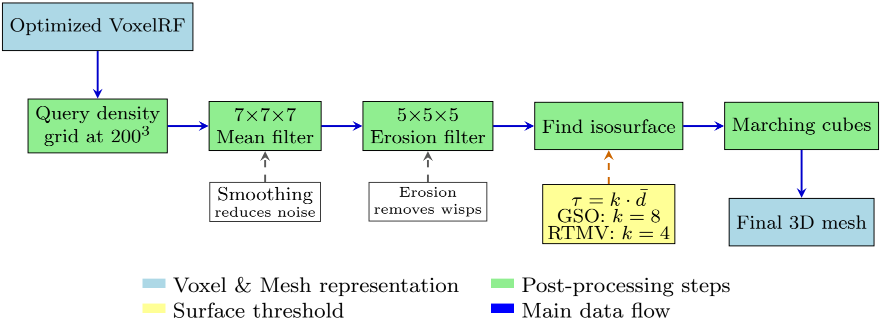


SJC [8] 3D Reconstruction Pipeline

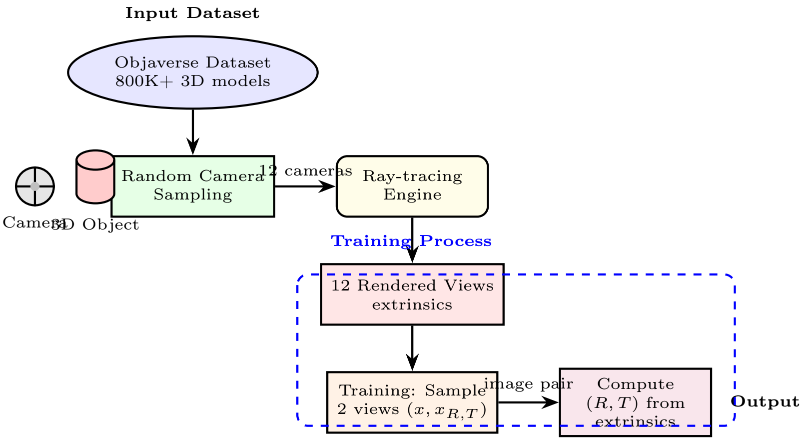


- NeRF representation
- Conditioning inputs
- Key technique
- Loss/Gradient flow
- Process steps
- Loss functions
- Main data flow
- Optimization loop

Mesh Extraction from VoxelRF



Data Preparation Pipeline from Objaverse



- (R, T) are practically elevation θ, azimuth ϕ and radius r
- Relative camera pose vector is represented as $[\theta_1 - \theta_2, \sin(\phi_1 - \phi_2), \cos(\phi_1 - \phi_2), r_1 - r_2]$

Evaluation Setup

Datasets and Tasks

- **Google Scanned Objects (GSO)** [9]: High-quality scanned household items.
- **RTMV** [10]: Complex scenes composed of 20 random objects. OOD.

PSNR (Peak Signal-to-Noise Ratio)

Measures pixel-level reconstruction fidelity between predicted and ground truth images:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right)$$

Higher values indicate better reconstruction fidelity.

SSIM (Structural Similarity Index)

Measures perceptual structural similarity between predicted and ground truth images:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Higher values indicate more perceptually similar images.

Metrics for Novel View Synthesis

Image Similarity and Quality Assessment

LPIPS (Learned Perceptual Image Patch Similarity)

Measures deep-feature perceptual similarity using a pretrained network:

$$\mathrm{LPIPS}(x, y) = \sum_{l} w_l \frac{1}{H_l W_l} \sum_{h,w} \left\| \hat{f}_l(x)_{hw} - \hat{f}_l(y)_{hw} \right\|_2^2$$

Lower values indicate greater perceptual similarity.

FID (Fréchet Inception Distance)

Measures similarity between real and generated image distributions in feature space:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\Big)$$

Lower values indicate higher generation quality.

Metrics for 3D Reconstruction

Geometric Accuracy Assessment

Chamfer Distance

Measures average closest-point distance between two point clouds:

$$d_{\mathrm{CD}}(S_1, S_2) = \frac{1}{|S_1|} \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|^2 + \frac{1}{|S_2|} \sum_{y \in S_2} \min_{x \in S_1} \|y - x\|^2$$

Lower values indicate better geometric accuracy.

Volumetric IoU (Intersection over Union)

Measures overlap between predicted and ground truth volumes:

$$\mathrm{IoU} = \frac{|V_{\mathrm{pred}} \cap V_{\mathrm{gt}}|}{|V_{\mathrm{pred}} \cup V_{\mathrm{gt}}|}$$

Higher values indicate better volumetric reconstruction.

Baselines

- DietNeRF [11]: NeRF+CLIP based method that performs few-shot 3D reconstruction.
- Image Variations (IV) [12]: Image-conditioned Stable Diffusion that generates many variants of image.
- SJC-I: Image-conditioned SJC diffusion model.
- MCC: Reconstructs 3D geometry from multiple RGB-D views.
- Point-E [13]: Text-to-image diffusion model + (two) point-cloud diffusion model that produces point cloud reconstruction.

Results

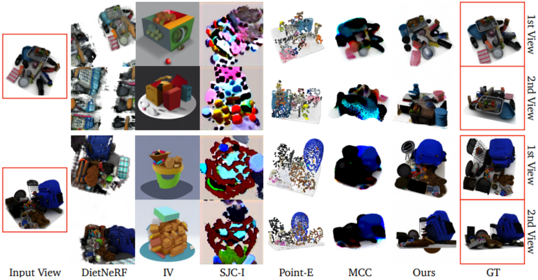
Novel View Synthesis

	DietNeRF	IV	SJC-I	Zero-1-to-3
PSNR ↑	8.933	5.914	6.573	**18.378**
SSIM ↑	0.645	0.540	0.552	**0.877**
LPIPS ↓	0.412	0.545	0.484	**0.088**
FID ↓	12.919	22.533	19.783	**0.027**

*GSO

	DietNeRF	IV	SJC-I	Zero-1-to-3
PSNR ↑	7.130	6.561	7.953	**10.405**
SSIM ↑	0.406	0.442	0.456	**0.606**
LPIPS ↓	0.507	0.564	0.545	**0.323**
FID ↓	5.143	10.218	10.202	**0.319**

*RTMV



- High fidelity even under big camera viewpoint changes
- Rich textual and geometric details

Results

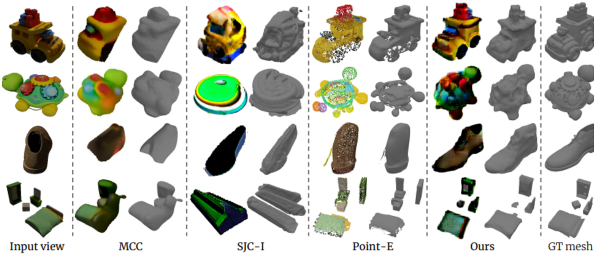
3D Reconstruction

	MCC	SJC-I	Point-E	Zero-1-to-3
CD ↓	0.1230	0.2245	0.0804	**0.0717**
IoU ↑	0.2343	0.1332	0.2944	**0.5052**

*GSO

	MCC	SJC-I	Point-E	Zero-1-to-3
CD ↓	0.1578	0.1554	0.1565	**0.1352**
IoU ↑	0.1550	0.1380	0.0784	**0.2196**

*RTMV

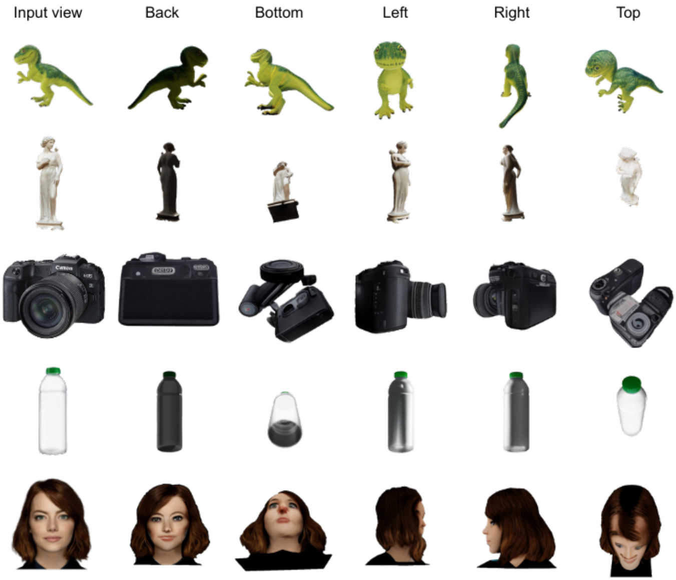


- SJC-I fails to produce 3D geometry; MCC struggles to generate the back side of objects; Point-E produces sparse output
- Zero-1-to-3 produces high fidelity results

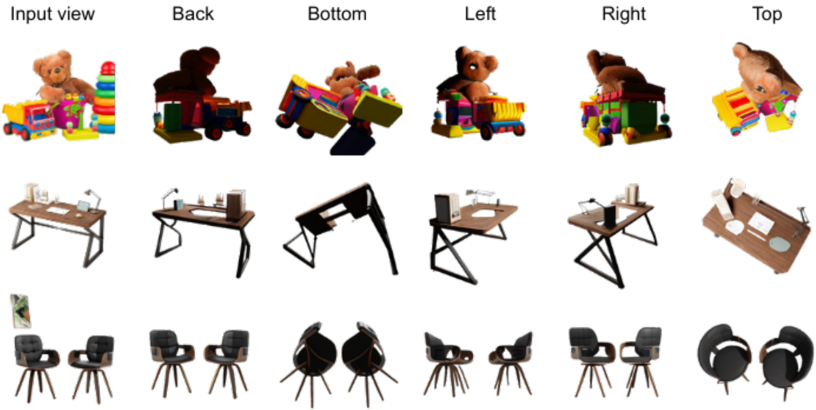
Our Results

- Code is available via link

Our Results



Our Results



Limitations

Figure 1: Hallucination under uncertainty

Figure 2: Single-image sparsity

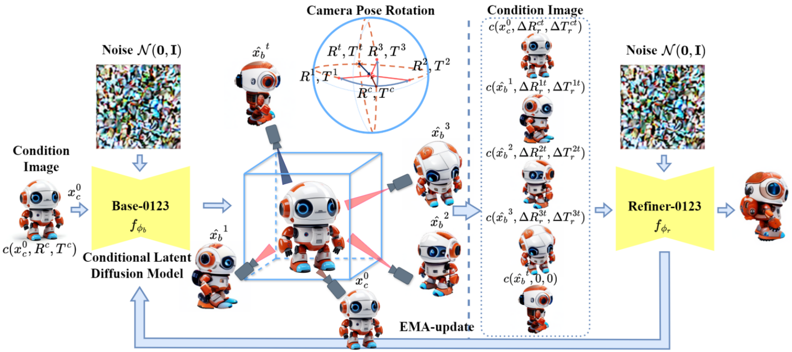
Figure 3: Viewpoint bias inherited from pretraining

Future work

Core Goal: Improve multi-view consistency and generalization beyond single-object scenes.

1. Direct Successor: Cascade-Zero123

Approach: Two-stage cascade (Base → Refiner) using self-generated nearby views.



Future work

2. Critical Correction: "Fixing the Perspective"
Problem: Flawed cross-attention reduces spatial reasoning; single-view input limits occluded regions.
Solution: Multi-view conditioning and revised embedding architecture.
Outcome: Meaningful cross-attention and improved 3D consistency.

Possible future applications

- Text (to image) to 3D
- From objects to scenes, from scenes to videos
- Using the same idea for scene relighting, material refinement, rendering...

Thank You!

References I

- Ruoshi Liu et al. "Zero-1-to-3: Zero-shot One Image to 3D Object". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. URL: https://arxiv.org/abs/2303.11328.
- Jeremy Reizenstein et al. "Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. DOI: 10.1109/ICCV48922.2021.01072. URL: https://arxiv.org/abs/2109.00512.
- Daniel Watson et al. "Novel View Synthesis with Diffusion Models". In: *arXiv preprint arXiv:2210.04628* (2022). URL: https://arxiv.org/abs/2210.04628.

References II

Markus Worchel et al. “Multi-view Mesh Reconstruction with Neural Deferred Shading”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Worchel_Multi-View_Mesh_Reconstruction_With_Neural_Deferred_Shading_CVPR_2022_paper.html.

Rohit Girdhar et al. “Learning a Predictable and Generative Vector Representation for Objects”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2016. URL: https://link.springer.com/chapter/10.1007/978-3-319-46448-0_36.

- Haoqiang Fan, Hao Su, and Leonidas J Guibas. "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 (CVPR)*. 2017. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Fan_A_Point_Set_CVPR_2017_paper.html.
- Chao-Yuan Wu et al. "Multiview compressive coding for 3D reconstruction". In: *arXiv preprint arXiv:2301.08247* (2023).

Haochen Wang et al. "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. URL: https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Score_Jacobian_Chaining_Lifting_Pretrained_2D_Diffusion_Models_for_3D_CVPR_2023_paper.html.

Laura Downs et al. "Google Scanned Objects: A high-quality dataset of 3D scanned household items". In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 2022.

References V

Jonathan Tremblay et al. "RTMV: A ray-traced multi-view synthetic dataset for novel view synthesis". In: *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*. 2022.

Ajay Jain, Matthew Tancik, and Pieter Abbeel. "Putting NeRF on a Diet: Semantically consistent few-shot view synthesis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

Lambda Labs. *Stable Diffusion Image Variations*. HuggingFace Spaces. `https://huggingface.co/spaces/lambdalabs/stable-diffusion-image-variations`. 2022.

Alex Nichol et al. "Point-E: A system for generating 3D point clouds from complex prompts". In: *arXiv preprint arXiv:2212.08751* (2022).