Zero-to-Hero: Enhancing Zero-Shot Novel View Synthesis via Attention Map Filtering

Ido Sobol[1] **Chenfeng Xu**[2] **Or Litany**[1,3]

[1]**Technion** [2]**UC Berkeley** [3]**NVIDIA**

https://zero2hero-nvs.github.io/

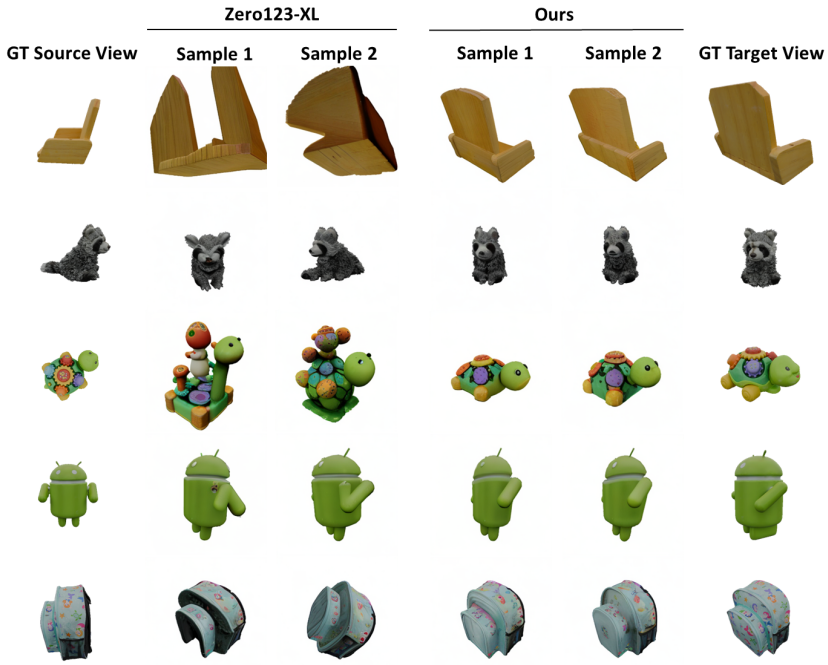

Figure 1: Novel views generated from a single source image (far left column) at a specific target view angle, compared between Zero123-XL [14] and our Zero-to-Hero method. Operating during inference, our method achieves significantly higher fidelity and maintains authenticity to the original image, all while ensuring realistic variation in the results (e.g. variations in chair backs in the top row). The ground-truth target view is displayed in the far right column.

Abstract

Generating realistic images from arbitrary views based on a single source image remains a significant challenge in computer vision, with broad applications ranging from e-commerce to immersive virtual experiences. Recent advancements in diffusion models, particularly the Zero-1-to-3 model, have been widely adopted for generating plausible views, videos, and 3D models. However, these models still struggle with inconsistencies and implausibility in new views generation, especially for challenging changes in viewpoint. In this work, we propose Zero-to-Hero, a novel test-time approach that enhances diffusion-based view synthesis by manipulating attention maps during the denoising process of Zero-1-to-3. By drawing an analogy between the denoising process and stochastic gradient descent (SGD), we implement a filtering mechanism that aggregates attention maps, enhancing generation reliability and authenticity. This process improves geometric

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

consistency without requiring retraining or significant computational resources. Additionally, we modify the self-attention mechanism to integrate information from the source view, reducing shape distortions. These processes are further supported by a specialized sampling schedule. Experimental results demonstrate substantial improvements in fidelity and consistency, validated on a diverse set of out-of-distribution objects. Additionally, we demonstrate the general applicability and effectiveness of Zero-to-Hero in multi-view, and image generation conditioned on semantic maps and pose.

1 Introduction

The pursuit of realistic image synthesis at arbitrary views, given only a single source image, has long been a cornerstone challenge in computer vision and graphics. This technology can cater to countless applications, such as interactive product inspection, robot-scene interaction, and immersive virtual experiences. In this work, we aim to advance this important line of research by improving the generation of novel views that are plausible and faithful to the input image. A recent, promising approach, Zero-1-to-3 [14] has developed a foundation model to synthesize novel views based on a single source image and a target view angle. By leveraging a pre-trained, image-conditioned stable diffusion model backbone [3], fine-tuned with target camera poses, and trained on paired source and target views from a vast collection of 3D models [8, 7], Zero-1-to-3 can generalize beyond its training set and generate plausible novel views. As a result, this model has quickly gained popularity, inspiring subsequent work in 3D and 4D scene generation [6, 11, 20, 15, 13, 12, 25, 17, 22, 19, 10].

While Zero-1-to-3 [14] has achieved substantial progress in novel view synthesis, several common issues limit its practical application. Firstly, the generated images might not fit real-world distributions, resulting in implausible and unrealistic outputs (e.g., first row in Fig.1). Secondly, the target image may appear plausible but be inconsistent with the input image in terms of shape or appearance (e.g., fifth row in Fig.1). Previous works have tried to mitigate these issues by retraining diffusion models with more data [7] or by generating multiple views [20, 12, 13, 11, 6, 15]. Despite substantial improvement, both approaches are resource-intensive due to the required re-training on large-scale 3D datasets.

In this work, we propose Zero-to-Hero, a novel test-time technique that addresses view synthesis artifacts through attention map manipulation. By drawing an analogy between the denoising process in diffusion models and stochastic gradient descent (SGD), we implement a filtering mechanism that aggregates attention maps, thereby enhancing generation reliability and authenticity. This process improves geometric consistency without requiring retraining or significant computational resources. Additionally, we modify the self-attention mechanism to integrate information from the source view, reducing shape distortions.

Our main contributions are as follows:

- To address the main limitations of the Zero-1-to-3 model, we perform an in-depth analysis and identify self-attention maps as the main candidate for correcting generation artifacts.
- We establish a conceptual analogy between model weights in stochastic gradient descent-based network training and the role of attentions map updates during generation of a denoising diffusion model. Based on this, we propose a simple yet powerful attention map filtering process resulting in enhanced target shape generation. We supplement our filtering technique with identity view information injection.
- Our method requires no additional training, and it avoids the overhead of external models or generating multiple views.

2 Method

We analyze the roles of self- and cross-attention layers in the model, and their contributions to the generated views.

Global pose conditioning through cross-attention. We first investigate cross-attention layers, as they are the only components in the model through which the target pose is injected. In the original text-to-image Stable Diffusion, the generation is conditioned on a prompt with multiple tokens. As

a part of stable diffusion cross-attention mechanism, Softmax is applied to the similarity scores between the latent elements and the text tokens, for normalization.

However, in Zero-1-to-3 the condition is a single embedding: a CLIP [18] embedding of the input image is concatenated with the relative transformation $[\mathcal{R}|\mathcal{T}]$ and mapped to the original CLIP dimension to form a single pose-CLIP embedding. Consequently, the Softmax operation is leading the post-softmax attention maps to be a degenerated constant all-ones matrix, as the summation of the Softmax is always 1 and there is a single condition. A visual demonstration is presented in Fig. 2. The post-softmax attention map is used to compute a weighted sum over the values matrix, obtained by a transformation of the condition. Since the attention matrix is an all-ones matrix, *we conclude that the cross-attention operation in Zero-1-to-3 degenerates into a global bias term, lacking spatially aware operations*. While in principle it is possible to improve the global bias term by additional optimization objectives and extra training overhead, we focus on the self-attention layers to enhance the results and mitigate the consistency issues while avoiding retraining the model.

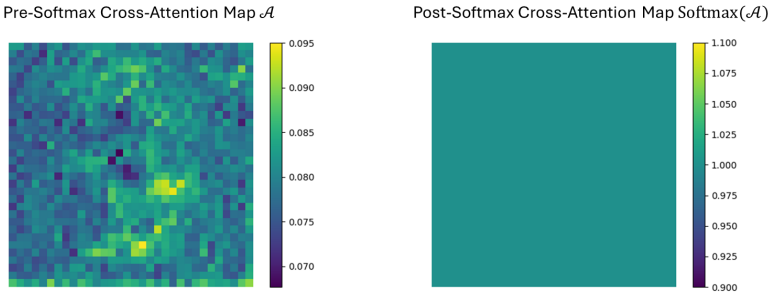

Figure 2: **Cross-Attention in Zero-1-to-3.** (Left) The cross-attention map before applying softmax. (Right) The degenerated all-ones attention map, produced by applying softmax on the left map. This behaviour holds across all UNet layers and timesteps.

Key observation: Spatial information flow through self-attention. Given the insight about the spatial-degeneracy in the cross-attention layers, we hypothesize that the self-attention layers preserve the information about the structure and geometry of the generated image, through the similarity scores between different elements in the latent vector. By monitoring the self-attention layers during the generation process, we observe that random noise introduced to the latent representation also introduces randomness to the attention maps. This randomness, while promoting generation diversity, can often lead to undesired strong correlations, that are misaligned with the true target. These strong correlations may persist through the denoising process, resulting in accumulated errors and visual artifacts.

From SGD to Diffusion Models: Attention Map Filtering as Weight-Space Manipulation. Recognizing attention maps as crucial for latent predictions, we hypothesize that enhancing robustness in attention maps predictions can significantly reduce generation misalignment. To achieve this, we draw an analogy between the denoising process in diffusion models and stochastic gradient descent (SGD) optimization of neural networks. Leveraging this analogy, we adapt techniques from SGD to enhance prediction consistency in diffusion models.

SGD is a fundamental tool in network training [4], designed to navigate the weight (network parameter) space towards local minima. For a neural network $F(x;\theta)$ with parameters θ, SGD samples training data points x_i and their corresponding labels y_i, and computes the gradient of the loss function $L(F(x_i;\theta), y_i)$ with respect to θ to update the parameters. In practice, aggregation of *gradients* and network *weights* during training is often performed to reduce variance and improve convergence. Gradient aggregation typically involves averaging gradient values over a batch, while weight aggregation accounts for the history of the weights in each update.

In this work, we view *the generation (denoising) process as an unrolled optimization, with attention maps as parameters of a score-prediction model*. Inspired by gradient aggregation and weight-averaging techniques that improve prediction robustness, we propose a filtering mechanism to enhance attention map reliability. We relate network weights that predict local gradients at each

optimization step, based on sampled training examples and labels, to the denoising network's attention maps that predict latent representations from sampled noise at each denoising step.

Robust View Generation via Attention Map Filtering. Our attention map filtering comprises of three parts:

- Minibatch via Resampling: Inspired by previous studies [16, 2], we implement per-step resampling throughout the image generation process. Through resampling we progressively generate R attention maps with different noise patterns. We propose to leverage these intermediate maps to further boost performance through in- and cross-step attention map manipulations.
- In-step update: Aggregating attention maps within the same denoising steps, generated through resampling. Specifically, a self-attention map is refined based on previous maps created at the same timestep. We find element-wise min pooling operation to work best for in-step updates.
- Cross-step update: averaging attention map across denoising steps. We pass the refined self-attention map at time t, to the next step in the denoising process, and aggregate the maps using exponential moving average (EMA).

The result is more reliable maps, particularly during the early denoising stages when coarse output shapes are formed, leading to more plausible and realistic views. A visual example of our attention map filtering mechanism in action is presented in Fig.3.

The attention map filtering described above significantly boosts plausibility and realism in generated views. Yet, we observe it is sometimes insufficient to enforce consistency with the input shape and appearance. To further promote consistency with the input, we propose to utilize mutual self-attention to propagate information from the input to the generated view [23, 5, 1, 25]. We modify the self-attention operation by running a parallel generation branch using the identity pose as target (in order to generate the input view), incorporating its keys and values into the attention layer of the target view. Unlike previous applications of this technique [5, 16, 1]), we find it beneficial in view synthesis to limit its use to the early denoising stages, preventing shape distortions.

An overview of our method is presented in Fig.4.

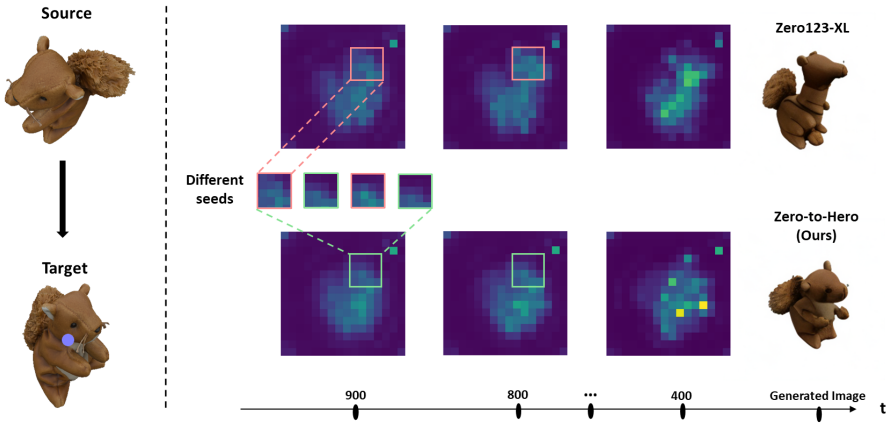

Figure 3: **Attention map filtering in action.** We compare the attention scores of Zero123-XL (top) and Zero-to-Hero (bottom) wrt the region marked with a purple circle at different denoising steps. Both methods are initialized with the same seed. We observe that the strong correlation values in the upper right corner lead to exaggerated content creation (note the unrealistically elongated neck). Conversely, through filtering, Zero-to-Hero mitigates these artifacts, leading to robust view synthesis.

3 Results

We evaluate Zero-to-Hero against Zero-1-to-3 and on Zero123-XL. In Tab. 1 we report various metrics for the original models using 25, 50 and 100 DDIM steps, and for our method applied to both models. We include the number of sampled timesteps T and the total number of network evaluation NFE (accounting for resampling). For the evaluation, we render random views from a challenging subset of Google Scanned Objects [9]. We report the standard image quality metrics PSNR, SSIM [24] and

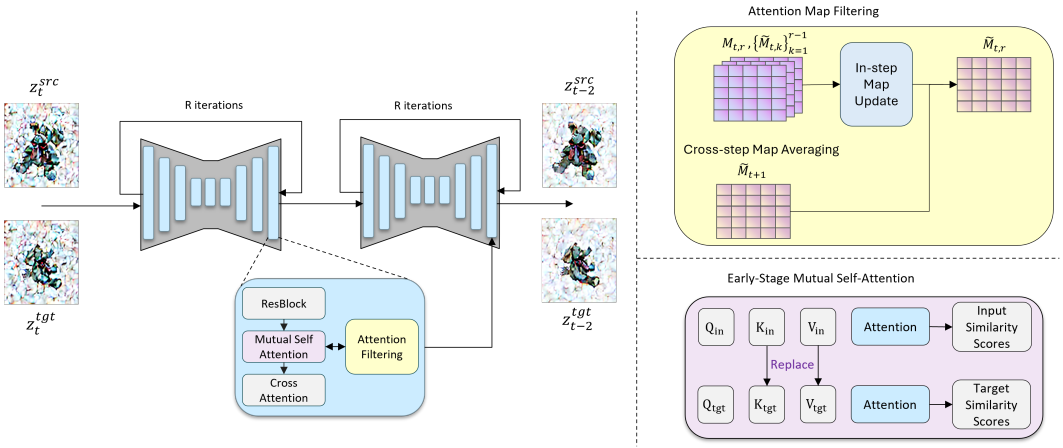

Figure 4: **Zero-to-Hero main modules.** (Left) Two denoising steps of the generation process of both the source (top) and target views (bottom). Each denoising step is iterated R times ("resampling"). (Right-top) **Attention map filtering**: Robustifying attention maps via an aggregation of same step and previous steps attention maps. (Right-bottom) **Mutual self-attention**: Guiding target shape through the keys and values of the source generation branch.

LPIPS [27]. As these metrics are sensitive to slight color variations, we segment the generated targets and their corresponding real images, and report the Intersection Over Union (IoU) score.

Through comprehensive experiments on out-of-distribution objects, we demonstrate that our technique robustifies Zero-1-to-3 and its extended version, Zero123-XL, leading to views that are more faithful to both the input image and desired camera transformation. Our results show significant and consistent improvements across both appearance and shape evaluation metrics.

Table 1: **Quantitative evaluation on a challenging subset of Google Scanned Objects [9].** Zero-to-Hero consistently improves performance upon baselines.

Name	T	NFE	PSNR ↑	SSIM ↑	LPIPS ↓	IOU ↑
Zero-1-to-3	25	25	17.27	0.851	0.173	73.5%
Zero-1-to-3	50	50	17.24	0.850	0.173	73.5%
Zero-1-to-3	100	100	17.21	0.850	0.173	73.4%
Ours (Zero-1-to-3)	26	66	**17.67**	**0.859**	**0.163**	**75.2%**
Zero123-XL	25	25	17.72	0.854	0.163	76.4%
Zero123-XL	50	50	17.71	0.854	0.162	76.4%
Zero123-XL	100	100	17.68	0.854	0.163	76.4%
Ours (Zero123-XL)	26	66	**18.35**	**0.864**	**0.153**	**78.3%**

4 Attention Map Filtering Beyond Novel View Synthesis

Although our work addresses the core limitations of single view synthesis models, the condition enforcing effect of our Attention Map Filtering (AMF) is more general. We have conducted several preliminary experiments which demonstrate promising results.

Conditional image generation. A brief study of ControlNet models [26] demonstrated that they suffer from similar limitations as Zero-1-to-3 and its follow ups. Namely, lack of condition enforcement and frequent appearance of visual artifacts. We implemented our proposed AMF module for two pre-trained ControlNet models (for Pose- and Segmentation-conditioned image generation) and found that it robustly mitigates artifacts across various prompts and seeds, as shown in Fig. 5

Multi-view synthesis. We integrate AMF into MVDream [21], a text-to-multiview model, and find that it helps to mitigate the same issues as in the single view case. In Fig. 6, we provide qualitative results.

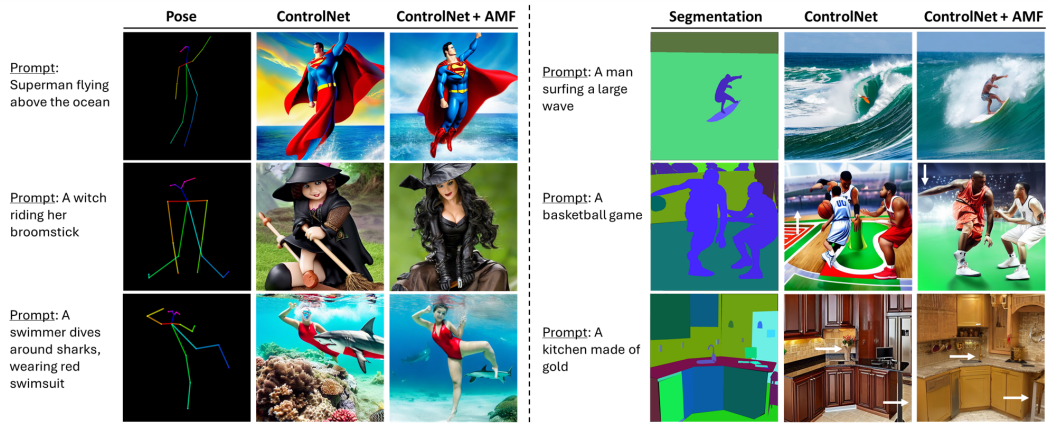

Figure 5: Qualitative results for ControlNet, without and with AMF. Both methods are initialized with the same seed. (Left) pose-conditioned ControlNet. (Right) Segmentation-conditioned ControlNet. In both cases, AMF leads to results that are more plausible and better aligns with the conditions. White arrows are used in some examples to highlight areas that don't align with the condition.

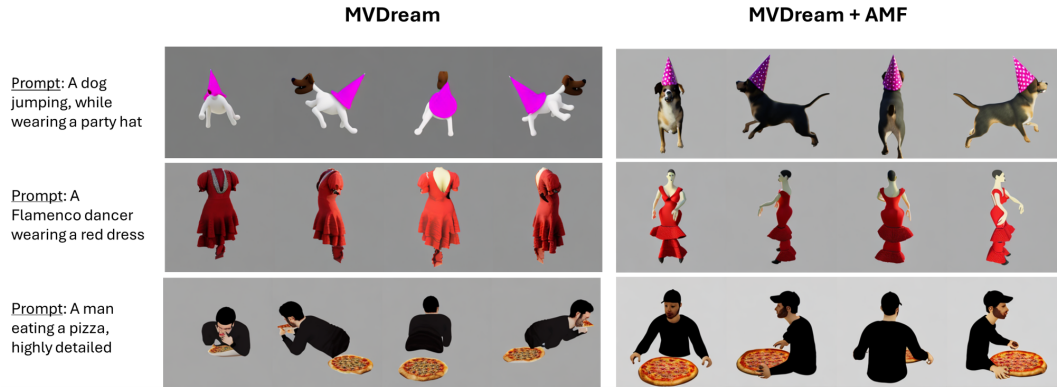

Figure 6: **Qualitative results for MVDream.** Qualitative results for MVDream, without and with AMF. Both methods are initialized with the same seed, and we provide results with two random seed per prompt. AMF leads to results that are more plausible and spatially consistent, while also better align with the conditions.

5 Conclusions

In this paper, we introduced Zero-to-Hero, a training-free method to boost the robustness of novel view synthesis. We enhanced the performance of a pre-trained Zero-1-to-3 diffusion model using two key innovations: a test-time attention map filtering mechanism that enhances output realism, and an effective use of source view information to improve input consistency.

Limitations. Our method, operating at test-time, is limited by the generative capabilities of the pre-trained model. If Zero-1-to-3 is not capable of correctly generating the target pose, our method may not enhance the output.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. arXiv preprint arXiv:2311.03335, 2023.

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 843–852, 2023.

[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.

[4] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers, pages 177–186. Springer, 2010.

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22560–22570, 2023.

[6] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. arXiv preprint arXiv:2312.04424, 2023.

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems, 36, 2024.

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13142–13153, 2023.

[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.

[10] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. arXiv preprint arXiv:2311.02848, 2023.

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

[12] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. arXiv preprint arXiv:2311.07885, 2023.

[13] 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.

[14] 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 ICCV, 2023.

[15] 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.

[16] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11461–11471, 2022.

[17] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843, 2023.

[18] 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.

[19] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. arXiv preprint arXiv:2312.17142, 2023.

[20] 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.

[21] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512, 2023.

[22] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653, 2023.

[23] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1921–1930, 2023.

[24] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004.

[25] Junwu Zhang, Zhenyu Tang, Yatian Pang, Xinhua Cheng, Peng Jin, Yida Wei, Wangbo Yu, Munan Ning, and Li Yuan. Repaint123: Fast and high-quality one image to 3d generation with progressive controllable 2d repainting. arXiv preprint arXiv:2312.13271, 2023.

[26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023.

[27] 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.