

329  
330

## Supplement to “UMAMI: Unifying Masked Autoregressive Models and Deterministic Rendering for View Synthesis”

### 331 A Failure cases

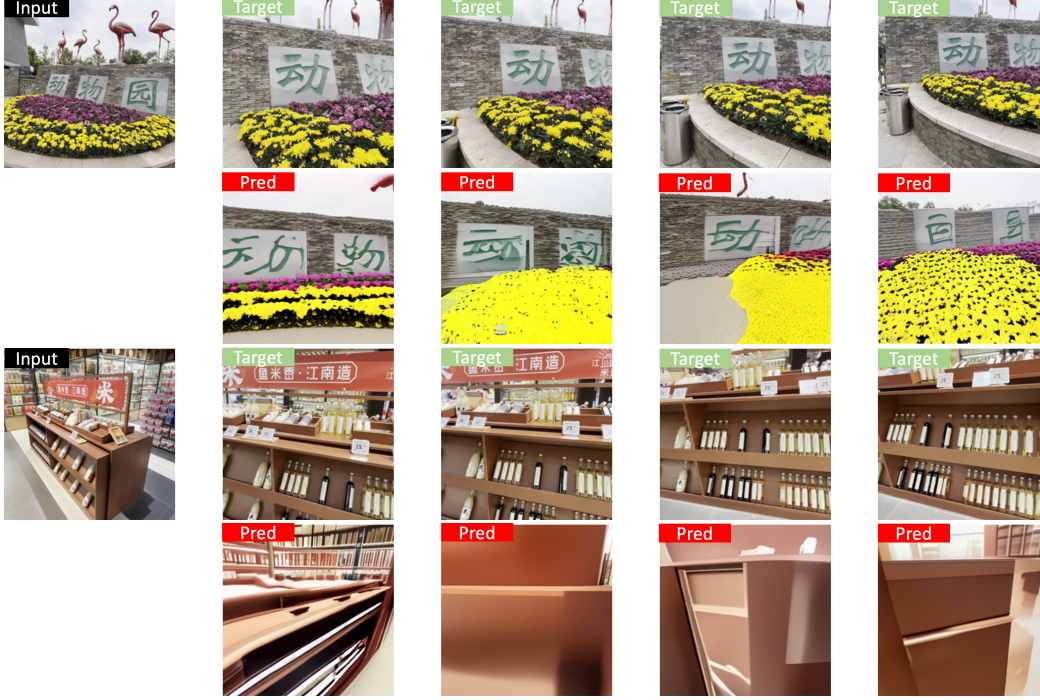


Figure 5: **Failure cases.** Our method may produce noticeable artifacts when target camera poses are too distant from the input view. Increasing the scale of training data and model parameters could improve the robustness of UMAMI.

### 332 B Related works

#### 333 B.1 Feed-forward deterministic NVS methods

334 Early generalizable methods for Novel View Synthesis (NVS) demonstrated the potential of neural  
 335 networks, trained across various scenes, to enable fast inference of novel views or underlying 3D rep-  
 336 resentations in a feed-forward manner. Prominent examples include PixelNeRF [97], MVSNeRF [10],  
 337 and IBRNet [82], which typically predict volumetric 3D representations by incorporating 3D-specific  
 338 priors like epipolar geometry or plane sweep cost volumes. Subsequent research has extended these  
 339 capabilities, improving performance particularly under challenging conditions such as sparse input  
 340 views [44, 31, 28, 29], and adapting these techniques for emerging representations like 3D Gaussian  
 341 Splatting (3DGS) [6, 72, 11, 73].

342 Recently, 3D Large Reconstruction Models (LRMs) have emerged [26, 37, 81, 92, 87, 89], leveraging  
 343 the power of scalable transformer architectures [76] trained on extensive datasets to learn generic 3D  
 344 priors. While these methods successfully avoid explicit architectural reliance on epipolar projection  
 345 or cost volumes, they still typically depend on pre-defined 3D representations such as tri-plane NeRFs,  
 346 meshes, or 3DGS, along with their corresponding rendering equations. This reliance can limit their  
 347 flexibility and overall potential.

348 An alternative line of work attempts to directly learn a geometry-free rendering function [69, 62,  
 349 67, 59, 36]. However, these approaches often face limitations in model capacity and scalability,  
 350 which can hinder their ability to capture high-frequency details. Notably, Scene Representation  
 351 Transformers (SRT) [62] aimed to avoid explicit, handcrafted 3D representations by learning a latent

scene representation via a transformer, an objective shared by our encoder-decoder architecture. Despite this similarity, certain design choices in SRT, such as its CNN-based token extractor and the use of cross-attention in the decoder, have been shown to lead to less effective performance. To address the issue, LVSM [30] proposes a method that is fully transformer-based, leveraging bidirectional self-attention for enhanced representational power. Furthermore, they introduce a novel and more scalable decoder-only architecture that directly learns the NVS function with minimal 3D inductive bias and without relying on an intermediate latent representation.

Our proposed method adopts the versatile and scalable decoder-only transformer backbone from LVSM, which has demonstrated its efficacy in NVS tasks by leveraging a data-driven approach with minimal handcrafted 3D inductive bias. However, a crucial distinction lies in the nature of our approach: unlike the deterministic LVSM, our method is generative. We aim to address the inherent limitations of deterministic methods by harnessing the generative capabilities of masked autoregressive diffusion models in an efficient manner.

## B.2 Generative-based NVS methods

The pursuit of generative-based (NVS) has recently seen significant advancements through the integration of diffusion models, drawing inspiration from successes in broader NVS [67, 62] and generative image-to-image tasks [60, 55, 61].

An early exploration in this domain was 3DiM [85], which trained image-to-image diffusion models for object-level multi-view rendering without explicit 3D representations. However, by training from scratch on limited 3D data, 3DiM’s applicability was restricted to category-specific scenarios and lacked zero-shot generalization capabilities. Building on this, Zero-1-to-3 [42] adopted a similar geometry-free pipeline but significantly improved generalization and output quality by fine-tuning a pretrained 2D diffusion model on a larger 3D object dataset [15]. Despite these improvements, a key challenge for Zero-1-to-3 and other early image-based diffusion models for NVS (e.g., for distant viewpoints [64]) was multi-view inconsistency, as they typically generated each target view independently and probabilistically, leading to jitter or inconsistencies when rendering a camera trajectory.

To address this multi-view inconsistency, subsequent research diverged into several directions. One line of work focused on integrating explicit 3D inductive biases—such as 3D representations or epipolar attention—into the diffusion denoising process. Examples include SyncDreamer [43], ConsistNet [94], Consistent-1-to-N [96], and MegaScenes [74], though these often came at the cost of increased computation. Another set of approaches, including Instant3D [37], MVDream [65], and Wonder3D [45], aimed to predict a single grid of multiple, specific views simultaneously. While this improved consistency across those fixed views, it sacrificed the ability for fine-grained camera control. Works like MVDream [65], SyncDreamer [43], and more recently HexGen3D [47], generate multiple fixed views from a conditional image but do not support arbitrary viewpoint selection. To achieve consistent 3D object geometry from these image-based models, further steps like NeRF distillation, using techniques such as Score Distillation Sampling (SDS) [54, 63] or direct optimization on sampled images [88, 20], are often necessary. However, distillation techniques such as SDS can introduce substantial computational overhead due to test-time optimization.

More recently, a promising trend has emerged with models that jointly predict multiple target views while maintaining accurate camera control and ensuring view consistency, often through mechanisms like cross-view attention. This category includes methods such as Free3D [100], EscherNet [35], CAT3D [20], and SV3D [78]. Several video model-based approaches [84, 22, 98, 93, 101] also fall into this paradigm, increasing NVS performance. Despite these advancements, achieving high-quality generation with these recent models often necessitates substantial computational resources and extensive training data. Furthermore, their reliance on full-image iterative sampling typically results in slow inference times, limiting practical applicability. Our proposed method, UMAMI, addresses this critical issue by enabling photorealistic novel view rendering while maintaining efficient inference times.

Table 5: Hyperparameters for training UMAMI. We use the same set of hyperparameters for both RealEstate10K and DL3DV experiments.

Component	Parameter	Value
Image Tokenizer	Image size	256
	Patch size	8
	Channels	9 (3 RGB + 6 for Plücker)
Transformer	Layers	24
	Hidden dim	768
	Head dim	64
	QK Norm	True
Training	Batch size / GPU	4
	Num GPUS	8
	Learning rate	0.0002
	Optimizer ( $\beta_1, \beta_2$ )	(0.9, 0.95)
	Grad clip norm	3.0
	Mixed precision	True
	Weight decay	0.02
	Train steps	100k
	Warmup steps	1000
Data Setup	Input / Target views	1 to 2 / 1 to 3
	Center Crop	True
Loss Weights	L2 loss	1.0
	LPIPS loss	0.0
	Perceptual loss	0.5
	Diffusion loss	10
	Confidence loss	1

## C Implementation details

### C.1 Hyperparameters

We report the hyperparameters used in Table 5.

### C.2 Algorithm

We describe the sampling process of UMAMI in Algorithm 1.

---

#### Algorithm 1 Hybrid Inference in UMAMI

---

**Require:** Trained model, context views  $\{(I_{\text{ctx}}, \pi_{\text{ctx}})\}$ , target pose  $\pi_{\text{tgt}}$ , threshold  $\tau$ , max unmasking steps  $T_{\text{max}}$

- 1: Tokenize context views into  $\mathbf{c}$ , initialize target tokens  $\mathbf{x}$  with masked tokens
  - 2: Encode  $(\mathbf{c}, \mathbf{x})$  with Transformer to obtain latent  $\mathbf{z}$
  - 3: Predict confidence map  $s_p$  and patch-level scores  $\mathbf{s}$
  - 4: Partition target tokens:
    - Deterministic tokens:  $\mathbf{x}_D \leftarrow \{x_i \mid s_i \geq \tau\}$
    - Stochastic tokens:  $\mathbf{x}_S \leftarrow \{x_i \mid s_i < \tau\}$
  - 5: Predict  $\mathbf{x}_D$  in one pass using deterministic head:  $\hat{\mathbf{x}}_D = \phi(\mathbf{z}_D)$
  - 6: Compute sampling steps:  $T_S = \lceil |x_S|/|x| \cdot T_{\text{max}} \rceil$
  - 7: **for**  $t = T_S$  **to** 1 **do**
  - 8:   Sample random unmasked set  $\mathbf{x}_t \subset \mathbf{x}_S$  following a cosine scheduler.
  - 9:   Update  $\mathbf{x}_t$  by DDPM sampling using  $\varphi$  head.
  - 10: **end for**
  - 11: Merge  $\hat{\mathbf{x}}_D$  and  $\hat{\mathbf{x}}_S$  into full target image  $\hat{I}_{\text{tgt}}$
-

## 407 D Additional quantitative results

### 408 D.1 Multiple images generation

Table 6: Multi-view generation results on RealEstate10K.

Dataset	# gen views	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$
Re10K-2views-Extra	1	28.95	0.107	0.897
	3	28.65	0.109	0.892
Re10K-2views-Interp	1	28.85	0.101	0.899
	3	28.52	0.105	0.894

409 As shown in Table 5, our model is trained to predict up to three target views simultaneously. This  
 410 joint prediction encourages consistency across generated images. In Table 6, we report results for  
 411 generating one and three views. The generation quality is comparable across both settings. Notably,  
 412 we use a fixed number of unmasking steps ( $T_{\max} = 32$ ) for all cases, which means generating  
 413 multiple views in parallel can improve inference efficiency without sacrificing quality.

## 414 E Additional qualitative results

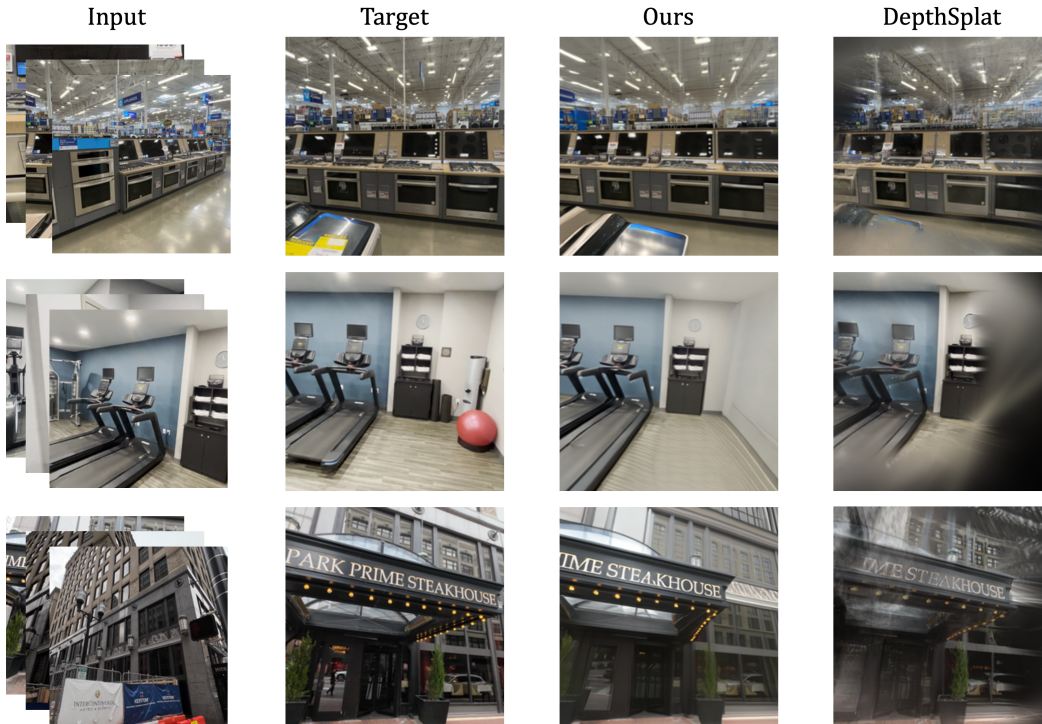


Figure 6: Additional qualitative comparisons on DL3DV dataset.

415 Additional qualitative evaluations are presented on the DL3DV dataset [40], where our method is  
 416 compared against DepthSplat [90] under a three-view input configuration (Figure 6). As depicted  
 417 UMAMI demonstrates notably sharp rendering, particularly in unobserved regions. This is achieved  
 418 by leveraging its generative capabilities to synthesize plausible details unobserved region of input  
 419 images.

420 Furthermore, to investigate the impact of the diffusion threshold hyperparameter,  $\tau$ , on UMAMI’s  
 421 performance, its value was systematically varied, with findings illustrated in Figure 7. An initial  
 422 setting of  $\tau = 0$ , corresponding to a fully deterministic operation of UMAMI, achieved rapid inference.



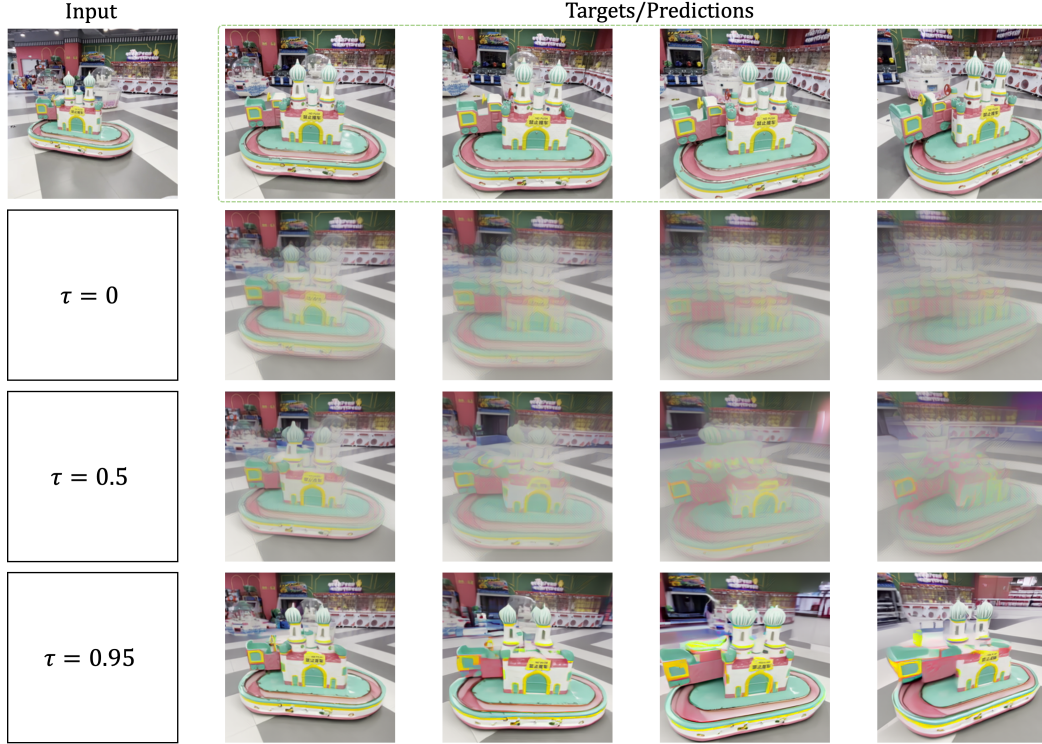


Figure 7: Impact of the diffusion threshold hyperparameter  $\tau$  on rendering outcomes. The top row shows the single input view alongside four corresponding target views. The subsequent rows (2-4) illustrate the results as  $\tau$  is incrementally increased. While a lower  $\tau$  promotes deterministic behavior and faster inference, higher values of  $\tau$  lead to notably sharper image rendering quality.

423 However, this configuration resulted in image blurring, an artifact attributable to unobserved regions  
 424 in the input view. Progressively increasing  $\tau$  to 0.5 and subsequently to 0.95 yielded a significant  
 425 enhancement in rendering quality. This improvement, however, was accompanied by an increase in  
 426 running time. Finally, to demonstrate the complete sampling dynamics of our method, the unmasking  
 427 processes for  $\tau = 0.95$  and for full unmasking diffusion process ( $\tau = 1$ ) are presented in the  
 428 supplementary video.

## 429 References

- 430 [1] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan.  
 431 Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings*  
 432 *of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021.
- 433 [2] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Zip-nerf: Anti-aliased  
 434 grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference*  
 435 *on Computer Vision*, pages 19697–19705, 2023.
- 436 [3] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi,  
 437 Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion  
 438 models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- 439 [4] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas,  
 440 J. Tremblay, S. Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In  
 441 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages  
 442 16123–16133, 2022.
- 443 [5] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. Maskgit: Masked generative image  
 444 transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*  
 445 *recognition*, pages 11315–11325, 2022.

- 446 [6] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann. pixelsplat: 3d gaussian splats from  
447 image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF*  
448 *conference on computer vision and pattern recognition*, pages 19457–19467, 2024.
- 449 [7] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis. Depth synthesis and local  
450 warps for plausible image-based navigation. *ACM transactions on graphics (TOG)*, 32(3):1–12,  
451 2013.
- 452 [8] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su. Tensorf: Tensorial radiance fields. In *European*  
453 *conference on computer vision*, pages 333–350. Springer, 2022.
- 454 [9] A. Chen, Z. Xu, X. Wei, S. Tang, H. Su, and A. Geiger. Factor fields: A unified framework for  
455 neural fields and beyond. *arXiv preprint arXiv:2302.01226*, 2023.
- 456 [10] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su. Mvsnerf: Fast generaliz-  
457 able radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF*  
458 *international conference on computer vision*, pages 14124–14133, 2021.
- 459 [11] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai. Mvsplat:  
460 Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on*  
461 *Computer Vision*, pages 370–386. Springer, 2024.
- 462 [12] I. Choi, O. Gallo, A. Troccoli, M. H. Kim, and J. Kautz. Extreme view synthesis. In  
463 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7781–  
464 7790, 2019.
- 465 [13] A. Davis, M. Levoy, and F. Durand. Unstructured light fields. In *Computer Graphics Forum*,  
466 volume 31, pages 305–314. Wiley Online Library, 2012.
- 467 [14] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from pho-  
468 tographs: A hybrid geometry-and image-based approach. In *Seminal Graphics Papers:*  
469 *Pushing the Boundaries, Volume 2*, pages 465–474. 2023.
- 470 [15] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani,  
471 A. Kembhavi, and A. Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings*  
472 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages  
473 13142–13153, 2023.
- 474 [16] H. Deng, T. Pan, H. Diao, Z. Luo, Y. Cui, H. Lu, S. Shan, Y. Qi, and X. Wang. Autoregressive  
475 video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024.
- 476 [17] S. Elfving, E. Uchibe, and K. Doya. Sigmoid-weighted linear units for neural network function  
477 approximation in reinforcement learning. arxiv e-prints, art. *arXiv preprint arXiv:1702.03118*,  
478 2017.
- 479 [18] W. Feng, J. Li, H. Cai, X. Luo, and J. Zhang. Neural points: Point cloud representation  
480 with neural fields for arbitrary upsampling. In *Proceedings of the IEEE/CVF conference on*  
481 *computer vision and pattern recognition*, pages 18633–18642, 2022.
- 482 [19] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels:  
483 Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on*  
484 *computer vision and pattern recognition*, pages 5501–5510, 2022.
- 485 [20] R. Gao, A. Holynski, P. Henzler, A. Brussee, R. Martin-Brualla, P. Srinivasan, J. T. Barron,  
486 and B. Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in*  
487 *Neural Information Processing Systems*, 2024.
- 488 [21] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra. Deep autoregressive networks.  
489 In *International Conference on Machine Learning*, pages 1242–1250, 2014.
- 490 [22] H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang. Cameractrl: Enabling camera  
491 control for text-to-video generation, 2024.

- [23] P. Hedman, P. P. Srinivasan, B. Mildenhall, J. T. Barron, and P. Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5875–5884, 2021.
- [24] B. Heigl, R. Koch, M. Pollefeys, J. Denzler, and L. Van Gool. Plenoptic modeling and rendering from image sequences taken by a hand-held camera. In *Mustererkennung 1999: 21st DAGM Symposium*, pages 94–101, 1999.
- [25] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.
- [26] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. LRM: Large reconstruction model for single image to 3d, 2024.
- [27] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3121–3128, 2011.
- [28] H. Jiang, Z. Jiang, K. Grauman, and Y. Zhu. Few-view object reconstruction with unknown categories and camera poses. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 31–41, 2024.
- [29] H. Jiang, Z. Jiang, Y. Zhao, and Q. Huang. Leap: Liberate sparse-view 3d modeling from camera poses, 2023.
- [30] H. Jin, H. Jiang, H. Tan, K. Zhang, S. Bi, T. Zhang, F. Luan, N. Snavely, and Z. Xu. LVSM: A large view synthesis model with minimal 3D inductive bias, 2024.
- [31] M. M. Johari, Y. Lepoittevin, and F. Fleuret. Geonerf: Generalizing nerf with geometry priors, 2022.
- [32] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711, 2016.
- [33] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.
- [34] D. P. Kingma, M. Welling, et al. Auto-encoding variational bayes, 2013.
- [35] X. Kong, S. Liu, X. Lyu, M. Taher, X. Qi, and A. J. Davison. Eschernet: A generative model for scalable view synthesis, 2024.
- [36] J. Kulhánek, E. Derner, T. Sattler, and R. Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision (ECCV)*, 2022.
- [37] J. Li, H. Tan, K. Zhang, Z. Xu, F. Luan, Y. Xu, Y. Hong, K. Sunkavalli, G. Shakhnarovich, and S. Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model, 2023.
- [38] T. Li, H. Chang, S. Mishra, H. Zhang, D. Katabi, and D. Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152, 2023.
- [39] T. Li, Y. Tian, H. Li, M. Deng, and K. He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024.
- [40] L. Ling, Y. Sheng, Z. Tu, W. Zhao, C. Xin, K. Wan, L. Yu, Q. Guo, Z. Yu, Y. Lu, et al. D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.
- [41] L. Liu, J. Gu, K. Z. Lin, T. Chua, and C. Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15651–15663, 2020.

- [42] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023.
- [43] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang. Syncdreamer: Generating multiview-consistent images from a single-view image, 2023.
- [44] Y. Liu, S. Peng, L. Liu, Q. Wang, P. Wang, C. Theobalt, X. Zhou, and W. Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7824–7833, 2022.
- [45] X. Long, Y. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S. Zhang, M. Habermann, C. Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion, 2023.
- [46] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [47] A. Mercier, R. Nakhli, M. Reddy, R. Yasarla, H. Cai, F. Porikli, and G. Berger. HexaGen3D: StableDiffusion is just one step away from fast and diverse text-to-3D generation. *arXiv preprint arXiv:2401.07727*, 2024.
- [48] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421, 2020.
- [49] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. Sajjadi, A. Geiger, and N. Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5480–5490, 2022.
- [50] A. V. D. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756, 2016.
- [51] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [52] E. Penner and L. Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics*, 36(6):1–11, 2017.
- [53] J. Plücker. On a new geometry of space. *Philosophical Transactions of the Royal Society of London*, pages 725–791, 1865.
- [54] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3D using 2D diffusion. *arXiv*, 2022.
- [55] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents, 2022.
- [56] C. Reiser, S. Peng, Y. Liao, and A. Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14335–14345, 2021.
- [57] C. Reiser, R. Szeliski, D. Verbin, P. Srinivasan, B. Mildenhall, A. Geiger, J. Barron, and P. Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics*, 42(4):1–12, 2023.
- [58] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.
- [59] R. Rombach, P. Esser, and B. Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14356–14366, 2021.
- [60] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *Proceedings of SIGGRAPH*, 2022.



- [61] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.
- [62] M. S. Sajjadi, H. Meyer, E. Pot, U. Bergmann, K. Greff, N. Radwan, S. Vora, M. Lučić, D. Duckworth, A. Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6229–6238, 2022.
- [63] K. Sargent, Z. Li, T. Shah, C. Herrmann, H.-X. Yu, Y. Zhang, E. R. Chan, D. Lagun, L. Fei-Fei, D. Sun, et al. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. arXiv preprint arXiv:2310.17994, 2023.
- [64] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su. Zero123++: A single image to consistent multi-view diffusion base model, 2023.
- [65] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang. MVDream: Multi-view diffusion for 3d generation, 2023.
- [66] S. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1881–1888, 2009.
- [67] V. Sitzmann, S. Rezhikov, B. Freeman, J. Tenenbaum, and F. Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 19313–19325, 2021.
- [68] Stability AI. Stable diffusion 3 (technical preview). Press release, 2024.
- [69] M. Suhail, C. Esteves, L. Sigal, and A. Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision (ECCV)*, pages 156–174, 2022.
- [70] M. Suhail, C. Esteves, L. Sigal, and A. Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8269–8279, 2022.
- [71] C. Sun, M. Sun, and H. Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5459–5469, 2022.
- [72] S. Szymanowicz, E. Insafutdinov, C. Zheng, D. Campbell, J. Henriques, C. Rupprecht, and A. Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image, 2024.
- [73] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.
- [74] J. Tung, G. Chou, R. Cai, G. Yang, K. Zhang, G. Wetzstein, B. Hariharan, and N. Snavely. Megascenes: Scene-level view synthesis at scale. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.
- [75] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.
- [76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [77] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, and P. P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- [78] V. Voleti, C. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani. SV3D: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 439–457, 2025.
- [79] V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, 2024.
- [80] S. Wan, T.-Y. Wu, W. H. Wong, and C.-Y. Lee. Confnet: predict with confidence. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2921–2925. IEEE, 2018.
- [81] P. Wang, H. Tan, S. Bi, Y. Xu, F. Luan, K. Sunkavalli, W. Wang, Z. Xu, and K. Zhang. PF-LRM: Pose-free large reconstruction model for joint pose and shape prediction, 2023.
- [82] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021.
- [83] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu. Nerf--: Neural radiance fields without known camera parameters, 2021.
- [84] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.
- [85] D. Watson, W. Chan, R. Martin-Brualla, J. Ho, A. Tagliasacchi, and M. Norouzi. Novel view synthesis with diffusion models, 2022.
- [86] D. Watson, S. Saxena, L. Li, A. Tagliasacchi, and D. J. Fleet. Controlling space and time with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [87] X. Wei, K. Zhang, S. Bi, H. Tan, F. Luan, V. Deschaintre, K. Sunkavalli, H. Su, and Z. Xu. Meshlrn: Large reconstruction model for high-quality mesh, 2024.
- [88] R. Wu, B. Mildenhall, P. Henzler, K. Park, R. Gao, D. Watson, P. P. Srinivasan, D. Verbin, J. T. Barron, B. Poole, et al. Reconfusion: 3D reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024.
- [89] D. Xie, S. Bi, Z. Shu, K. Zhang, Z. Xu, Y. Zhou, S. Pirk, A. Kaufman, X. Sun, and H. Tan. Lrm-zero: Training large reconstruction models with synthesized data, 2024.
- [90] H. Xu, S. Peng, F. Wang, H. Blum, D. Barath, A. Geiger, and M. Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024.
- [91] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5438–5448, 2022.
- [92] Y. Xu, H. Tan, F. Luan, S. Bi, P. Wang, J. Li, Z. Shi, K. Sunkavalli, G. Wetzstein, Z. Xu, and K. Zhang. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model, 2023.
- [93] Y. Yan, Z. Xu, H. Lin, H. Jin, H. Guo, Y. Wang, K. Zhan, X. Lang, H. Bao, X. Zhou, and S. Peng. Streetcrafter: Street view synthesis with controllable video diffusion models, 2024.
- [94] J. Yang, Z. Cheng, Y. Duan, P. Ji, and H. Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion, 2023.
- [95] B. Ye, S. Liu, H. Xu, X. Li, M. Pollefeys, M.-H. Yang, and S. Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

- 681 [96] J. Ye, P. Wang, K. Li, Y. Shi, and H. Wang. Consistent-1-to-3: Consistent image to 3d view  
682 synthesis via geometry-aware diffusion models, 2023.
- 683 [97] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelNeRF: Neural radiance fields from one or  
684 few images. In *CVPR*, 2021.
- 685 [98] W. Yu, J. Xing, L. Yuan, W. Hu, X. Li, Z. Huang, X. Gao, T. Wong, Y. Shan, and Y. Tian.  
686 Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis, 2024.
- 687 [99] Q. Zhang, S. Baek, S. Rusinkiewicz, and F. Heide. Differentiable point-based radiance fields  
688 for efficient view synthesis. In *SIGGRAPH Asia Conference Papers*, pages 1–12, 2022.
- 689 [100] C. Zheng and A. Vedaldi. Free3d: Consistent novel view synthesis without 3d representation.  
690 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*  
691 (*CVPR*), pages 9720–9731, 2024.
- 692 [101] J. J. Zhou, H. Gao, V. Voleti, A. Vasishta, C.-H. Yao, M. Boss, P. Torr, C. Rupprecht, and  
693 V. Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv*  
694 *preprint arXiv:2503.14489*, 2025.
- 695 [102] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: Learning view  
696 synthesis using multiplane images. In *SIGGRAPH*, 2018.