4DIFF: 3D-Aware Diffusion Model for Third-to-First Viewpoint Translation

Feng Cheng[1,3*], Mi Luo[2*], Huiyu Wang[1], Alex Dimakis[2], Lorenzo Torresani[1], Gedas Bertasius[3†], and Kristen Grauman[1,2†]

[1] FAIR, Meta AI
[2] The University of Texas at Austin
[3] University of North Carolina at Chapel Hill
[*] Equal contribution, [†] Co-lead the project

Abstract. We present 4DIFF, a 3D-aware diffusion model addressing the exo-to-ego viewpoint translation task — generating first-person (egocentric) view images from the corresponding third-person (exocentric) images. Building on the diffusion model's ability to generate photorealistic images, we propose a transformer-based diffusion model that incorporates geometry priors through two mechanisms: (i) egocentric point cloud rasterization and (ii) 3D-aware rotary cross-attention. Egocentric point cloud rasterization converts the input exocentric image into an egocentric layout, which is subsequently used by a diffusion image transformer. As a component of the diffusion transformer's denoiser block, the 3D-aware rotary cross-attention further incorporates 3D information and semantic features from the source exocentric view. Our 4DIFF achieves state-of-the-art results on the challenging and diverse Ego-Exo4D multiview dataset and exhibits robust generalization to novel environments not encountered during training. Our code, processed data, and pretrained models are publicly available at `https://klauscc.github.io/4diff`.

Keywords: Egocentric Vision · View Synthesis

1 Introduction

From early developmental stages, humans adeptly observe external actions (exo) and seamlessly integrate them into their own repertoire (ego), forming the cornerstone of visual learning. This actor-observer translation mechanism not only shapes individual development but also holds profound implications for technological advancements. Imagine the ability to immerse yourself in the first-person perspective of renowned athletes like Messi or glean intricate piano techniques from online tutorials converted to a first-person viewpoint. Such experiences hinge on seamless translation from third-person to first-person perspectives, highlighting the pivotal role of cross-view translation in facilitating immersive and enriching experiences across diverse domains.

We leverage the recently released Ego-Exo4D dataset [18] to explore the third-person (exocentric) to first-person (egocentric) viewpoint translation task.

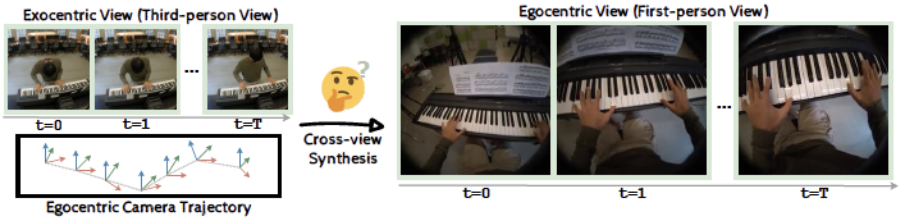

Fig. 1: Given exocentric images of an egocentric camera wearer engaged in daily activities and the corresponding camera trajectories, we aim to synthesize the corresponding egocentric view that captures the scene from the wearer's first-person perspective.

As illustrated in Figure 1, our focus is on transforming the exocentrically observed images containing a designated individual into images depicting the same scene from the individual's first-person perspective. Our task is a specific instance of the Novel View Synthesis (NVS) task, which aims to generate new views conditioned on a few given views of a scene. However, the Ego-Exo4D dataset presents a formidable challenge compared to traditional novel view synthesis datasets [8, 9, 17, 22, 57, 72] and multiview datasets [1, 26, 55, 70]. As illustrated in Figure 2, the scenes in the Ego-Exo4D dataset are characterized by numerous objects and dynamic actions performed by the participants. The dataset encompasses diverse scenes, ranging from indoor to outdoor activities such as cooking and basketball. Furthermore, the visual differences between exocentric and egocentric images are pronounced due to sharp viewpoint changes. Besides, unlike numerous NVS datasets that use 3D data for arbitrary viewpoint sampling during training, Ego-Exo4D dataset only provides several views (e.g., four exo and one ego view) for each dynamic scene, which presents a challenge for convergence of prior geometry-based methods that regress the entire scene.

Due to the challenges mentioned above, existing methods exhibit unsatisfactory performance in the exo-to-ego view translation task. Geometry-free generative models, including GAN-based [6,21] and diffusion-based [30,38,66] methods, face challenges in generating geometrically-correct images due to high complexity of the scenes. In contrast, geometry-based approaches, exemplified by NeRF-based methods [2,3,34,37,41,69], encounter limitations in achieving photorealistic images. Recent attempts [7, 10] aim to reconcile this dilemma by integrating a strong geometry-based method (e.g. NeRF-based) into diffusion models. However, these models are typically difficult to optimize on the extremely diverse scenes in the Ego-Exo4D benchmark, as we show in Sec. 4.2. Thus, they often fail to provide constructive geometry priors to the subsequent diffusion model.

Motivated by these observations, we propose 4DIFF, a **3D**-Aware **Diff**usion model for exocentric to egocentric viewpoint translation. We propose two mechanisms to incorporate 3D geometry into the diffusion model: (i) egocentric point cloud rasterization, and (ii) 3D-aware rotary cross-attention layers. Rather than relying on a complex geometry model like NeRF, we render an egocentric prior image using a lightweight rasterization technique [5, 67]. As a result, our approach is both easy to train and adaptable, allowing it to incorporate existing

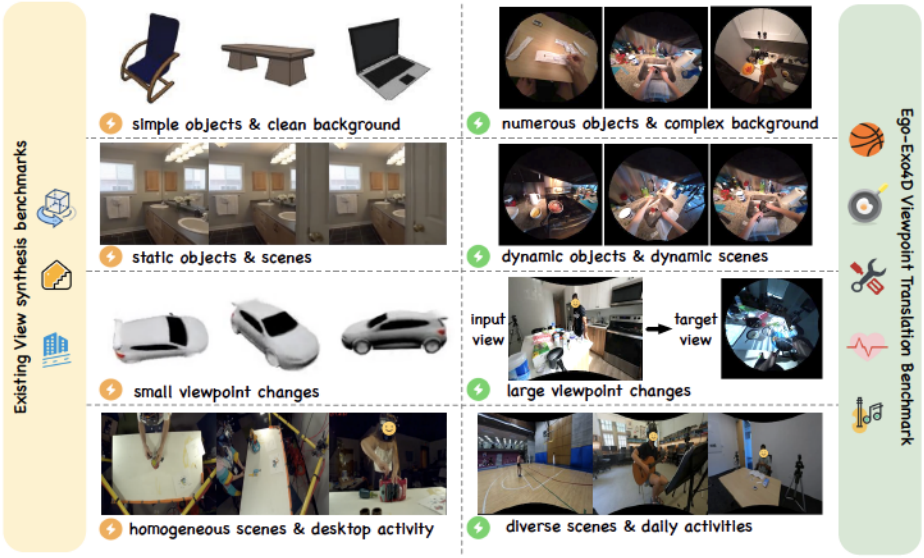

Fig. 2: Comparison of the Ego-Exo4D viewpoint translation (Ego-Exo4D-VT) benchmark, which we build on the Ego-Exo4D dataset [18], with existing novel view synthesis and cross-view translation benchmarks. Ego-Exo4D-VT presents numerous challenges that require fundamental advances in generative modeling to address.

open-source pretrained depth estimators. These estimators have demonstrated effectiveness in processing images from previously unseen environments [4, 68]. Solely rendering the egocentric prior feature map through point cloud rasterization can be problematic, as the source exo view often contains occluded and unobserved regions. To address this, we seamlessly integrate rasterization into the diffusion model, leveraging its substantial capacity for extrapolation and generating high-quality images. We further enhance the expressivity of our diffusion model by introducing 3D-aware rotary cross-attention, which is integrated into each denoising block of the model. This functionality aims to improve feature similarities and 3D spatial similarities between ego and exo views, allowing the diffusion feature maps to incorporate information from the semantic features encoded in the exocentric image more effectively.

Our method 4DIFF surpasses prior state-of-the-art techniques on the challenging Ego-Exo4D viewpoint translation benchmark, achieving a **3.6%** improvement in LPIPS. Furthermore, leveraging the extensive scale of Ego-Exo4D data, our approach demonstrates robust generalization to novel environments not encountered during training.

2 Related Work

Exo-to-Ego Viewpoint Translation. Prior methods [28, 44, 62] tackled this problem predominantly via GAN-based models [11]. Specifically, [43] proposed the X-Fork and X-Seq GAN-based architecture using an additional semantic map

for enhanced generation. [29] introduced STA-GAN, which focuses on learning spatial and temporal information to generate egocentric videos from exocentric views. [32] focuses on hand-object interactions, proposing to decouple hand layout generation and ego frame generation with a diffusion model. None of these methods develop an explicit geometry-aware generative framework. In contrast, our work introduces two effective mechanisms to incorporate 3D geometric priors into the diffusion model, specifically tailored to address the challenges posed by the Ego-Exo4D-VT benchmark.

Novel View Synthesis (NVS). Our exo-to-ego viewpoint translation task represents a distinct facet of the NVS task, which aims to generate a target image with an arbitrary target camera pose from given source images and their camera poses. Previous works in NVS can be categorized into geometry-based [15,16,31,46,47,56,64,72], regression-based methods [25,35,54,63–65,69,72] and generative models [24,45,48,50,66,67]. Recently, several geometry-aware generative models [7,10] have explored ways to integrate NeRF with diffusion models. For instance, GeNVS [7] incorporates geometry priors into their diffusion model using a variant of pixelNeRF [69], which renders a target feature map from a 3D feature field. SSDNeRF [10] proposes a unified approach that employs an expressive diffusion model to learn a generalizable prior of neural radiance field (NeRF). However, these geometry-based models, typically implemented as NeRFs, often struggle to provide meaningful geometry priors to the diffusion model, especially in the challenging Ego-Exo4D-VT benchmark. This is because complex geometry methods require strong supervision (e.g., many densely sampled views of the same scene), which Ego-Exo4D does not provide. In contrast, our method uses simple point-cloud rasterization that relies solely on accurate depth estimation, avoiding the modeling of occluded and unobserved areas in the exocentric view. This approach shows better generalization and benefits from existing large-scale pretrained depth estimators.

Diffusion Models [12,19,49] have made significant strides in producing photorealistic images and videos. They excel in modeling conditional distributions, including scenarios where conditioning is based on text [49,52] or another image [20,53]. Prior work has demonstrated a wide range of successful applications of diffusion models, including human pose generation [27] and depth estimation [14]. In our work, we employ a transformer-based diffusion model [39] to model the distribution of egocentric images conditioned on exocentric images.

3 Methodology

3.1 Problem Setup

Given an exocentric image $x \in \mathbb{R}^{h \times w \times 3}$ and the relative camera pose $P \in \mathbb{R}^{4 \times 4}$ from exo camera to the ego camera of the person of interest, our goal is to synthesize an egocentric image $y \in \mathbb{R}^{h \times w \times 3}$ from the conditional distribution:

$$p(y|x, P) \tag{1}$$

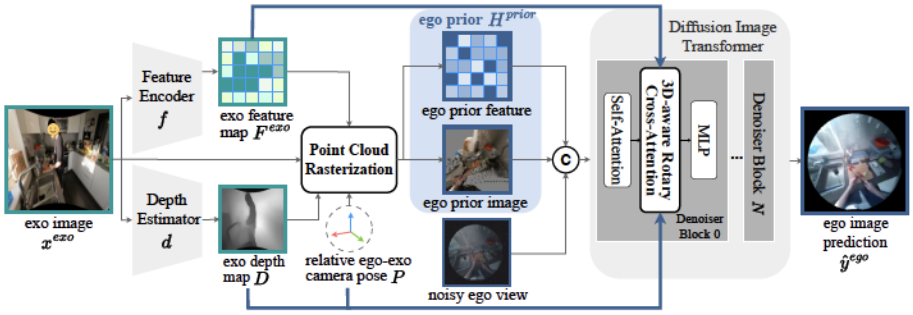

Fig. 3: We propose 4DIFF, a 3D-Aware Diffusion model for exocentric to egocentric viewpoint translation. Our framework uses a point cloud rasterization scheme first to compute an egocentric prior, which captures egocentric layout cues. Afterward, the egocentric prior is fed into the diffusion model augmented with the proposed 3D-aware rotary cross-attention for egocentric image generation. The proposed 3D-aware rotary cross-attention guides the attention to consider geometric relationships between the egocentric and exocentric diffusion feature maps.

We assume the relative camera pose (P) is known, similar to the standard NVS tasks [40, 61, 69].

Relation to the Official Ego-Exo4D Translation Benchmark. Ego-Exo4D [18] introduced an exo-to-ego translation benchmark, with the primary emphasis on object-level synthesis, i.e., generating an object at the correct location in the ego view based on an exo image and an exo segmentation mask of the object of interest. This approach is particularly valuable for precise object placement and detailed object-level interactions. In contrast, we focus on full-image synthesis — allowing for the generation of entire scenes, and enhancing the richness and diversity of generated viewpoints. Both are complementary; while Ego-Exo4D excels in object-specific scenarios, our method expands the scope to full-scene synthesis and can be seen as a new specialized NVS task.

3.2 Our Framework

Due to the inherent complexity and dynamism present in diverse scenes, we use an expressive transformer-based diffusion model to model the conditional distribution in Equation 1. However, due to the inability to explicitly model 3D cues, the standard diffusion model may struggle to generate geometry-consistent images. Thus, we propose two techniques to incorporate geometry into our diffusion model: (i) egocentric point cloud rasterization and (ii) 3D-aware rotary cross-attention. As shown in Figure 3, the point cloud rasterization first renders an egocentric prior from the input exocentric view, which is then fed into the diffusion model. Afterward, the conditioned diffusion model is augmented with the proposed 3D-aware rotary cross-attention to generate the target egocentric image. We now describe each module in more detail.

3.3 Egocentric Point Cloud Rasterization

As a first step in our framework, we render an egocentric prior via the point cloud rasterization from an exocentric view. Specifically, we first use a depth estimator to convert the exocentric 2D image x and a feature map F^{exo} into a feature point cloud. Then, a differential renderer [67] projects this point cloud into an egocentric prior H^{prior}:

$$H^{\text{prior}} = [x^{\text{prior}}, F^{\text{prior}}] = \text{render}([x, F^{\text{exo}}], D, P) \tag{2}$$

Here, F^{exo} is the semantic features of the exocentric image encoded by a feature encoder f, x^{prior} and F^{prior} are the egocentric prior image and a feature map, rendered from the exocentric image x and a feature map F^{exo} respectively. D denotes the depth map predicted by a depth estimator, and P represents the relative camera pose.

Depth Estimator. We construct the depth estimator based on the pretrained MiDaS [4]. Since MiDaS predicts relative disparity (the inverse of depth), we introduce two learnable scalars s and t for dataset-specific calibration. The depth map D is predicted using the formula:

$$D = 1/(s \cdot \text{MiDaS}(x^{\text{exo}}) + t). \tag{3}$$

Rasterization. We employ the differentiable renderer [67] for our rasterization. This renderer splats 3D points onto the image plane and calculates pixel values by blending point features. In contrast to more intricate rendering techniques like NeRF [34, 69] or Gaussian Splatting [23, 60], our renderer is simpler to converge. It relies solely on depth estimation from 2D images, leveraging large-scale pretrained depth estimators. This design choice ensures robust generalization across diverse scenarios.

3.4 3D-Aware Diffusion Image Transformer

Our diffusion model uses a denoiser network to predict added noise ϵ_t from the noisy target egocentric image $y_t = \sqrt{\bar{\alpha}_t}y + \sqrt{1-\bar{\alpha}_t}\epsilon_t$, conditioned on the previously obtained egocentric prior H^{prior} and the exocentric semantic features F^{exo}:

$$\hat{\epsilon}_t = \epsilon_\theta([y_t, H^{\text{prior}}], F^{\text{exo}}). \tag{4}$$

During inference, the target egocentric image y_0 is generated from a standard Gaussian noise y_T by applying the denoiser network ϵ_θ iteratively with a sampling strategy (e.g. DDIM [58]), i.e. $y_T \rightarrow y_{T-\delta} \rightarrow ... \rightarrow y_0$.

Denoiser Network ϵ_θ. Our proposed 3D-aware Diffusion image Transformer serves as the denoiser network. As shown in Figure 3 and Equation 4, our Transformer network takes as input the concatenation of the egocentric prior H^{prior} and the noisy target egocentric image y_t encoded via an off-the-shelf autoencoder from [49]. Following [39], the architecture of DiT is the same as ViT, consisting

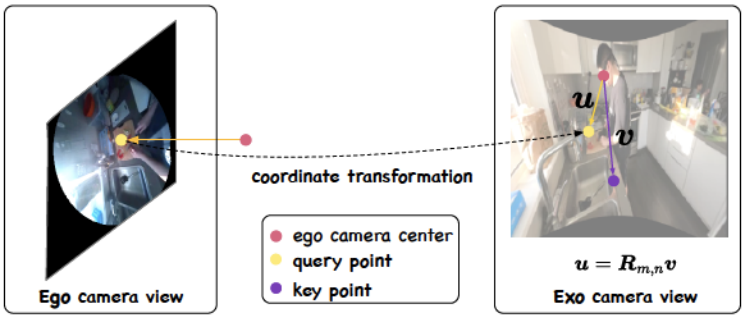

Fig. 4: An illustration of the calculation of the rotation matrix $R_{m,n}$ in our 3D-aware rotary cross attention.

of N transformer layers, each with a self-attention layer and a feedforward network. To further enhance the expressivity of our model and incorporate more geometric cues, we propose 3D-aware rotary cross-attention layers, which we describe next.

3D-aware Rotary Cross-Attention. When conditioning the diffusion model on the exocentric feature map, we should consider similarities in the semantic feature and spatial 3D space. Exocentric features similar in appearance (i.e., semantic feature space) and 3D location with respect to the query features should have higher attention values in the diffusion model. Motivated by RoPE [59], we achieve this by incorporating rotations during attention weight calculations. The degree of rotation between a query and a key is determined by the angle between their 3D coordinates, with the ego camera as the center. Consequently, the cosine similarity between the query and key features can incorporate their 3D spatial angle, effectively capturing the 3D relationships between corresponding points in the egocentric and exocentric views.

Specifically, given a feature map $Z \in \mathbb{R}^{l \times c}$ in the diffusion model and the exocentric semantic feature map $F^{\text{exo}} \in \mathbb{R}^{l \times c}$, the 3D-aware rotary cross-attention calculates the output $O \in \mathbb{R}^{l \times c}$ as:

$$a_{m,n} = \frac{\exp(\frac{q_m^T R_{m,n} k_n}{\sqrt{c}})}{\sum_{j=1}^{l} \exp(\frac{q_m^T R_{m,j} k_j}{\sqrt{c}})} \tag{5}$$

$$O_m = \sum_{n=1}^{l} a_{m,n} v_n \tag{6}$$

Here, $q_m = Z_m W_q$ is the m-th query token, $k_n = F_n^{\text{exo}} W_k$ is the n-th key token and $v_n = F_n^{\text{exo}} W_v$ is the n-th value token. W_q,W_k, W_v are learnable project matrices. $R_{m,n}$ is the rotation matrix that rotates the key token to align with the value token in 3D space, where the egocentric camera is used as the center. Since the query token is in the egocentric view, we map its coordinates to the exocentric view using the relative camera pose. The rotation matrix is computed in the exocentric view using the algorithm from [33]. When $R_{m,n}$ is an identity

matrix, our 3D-aware rotary cross-attention defaults to standard cross-attention. Figure 4 shows an illustration of this process. We insert such 3D-aware cross-attention layers after each self-attention layer in DiT.

3.5 Training and Inference

Loss Function. Our model is trained with the diffusion denoising loss, which is the L2 loss between the predicted noise and the ground-truth added noise.

Implementation Details We employ DINOv2 [36] pretrained ViT-L/14 as our feature encoder f and MiDaS [4] with DPT-L as our depth estimator. Our denoiser network is built on DiT-B/2 [38] augmented with the proposed 3D-aware rotary cross-attention layers. The image sizes are 256×256 for both egocentric and exocentric images. We freeze the feature encoder, as it is already well pretrained. The model is trained with the Adam optimizer, using a learning rate of $1e-5$ for the depth estimator and $1e-4$ for the other components. We employ a batch size of 4 per GPU and train the model across 32 V100 GPUs for 100 epochs, requiring approximately 48 hours. We set the diffusion steps T to 1000 during training and sample 20 steps during inference using DDIM [58].

4 Experiments

4.1 Experimental Setup

Ego-Exo4D-VT Benchmark. Our benchmark is constructed based on the Ego-Exo4D dataset [18]. Adhering to the official splits, we use 2680/708/900 takes for training, validation, and testing, respectively. Each take is approximately 30 seconds to 5 minutes long and depicts a person performing a skilled activity, such as cooking a dish, with footage from 4 exocentric cameras and 1 egocentric camera. This benchmark encompasses five diverse, skilled human activities: basketball, bike repair, cooking, health, and music.

The benchmark features 131 unique scenes, each characterized by complex backgrounds and numerous objects, demonstrating significant scale variation from 1 meter (e.g., a small kitchen) to 10 meters (e.g., a basketball court). These scenes are dynamic and depict subjects performing actions that involve interactions with objects. Additionally, the considerable viewpoint shift from exocentric to egocentric view causes objects to appear relatively small in the exocentric view compared to the egocentric view.

Baselines. Since this is a new benchmark, we re-purpose a few state-of-the-art methods for image generation: (a) pix2pix [21], a GAN-based method, (b) GNT [61], a NeRF-based method, (c) diffusion model DiT [39] and 3DiM [66]. To tailor DiT for our task, we eliminate its original class label conditioning and condition it on the exocentric image through concatenation. Additionally, we implement 3DiM based on DiT since the code for 3DiM is unavailable.

Metrics. Following NVS methods [10, 69], we employ perceptual metrics, including LPIPS [71], DISTS [13] and CLIP score [42], to measure the structural

Table 1: Quantitative comparison on the test set of Ego-Exo4D-VT benchmark. † we reimplement 3DiM based on DiT as their code is not publicly available. Our 4DIFF achieves the best results on all the metrics, outperforming the second best method 3DiM by **3.6%** in LPIPS and **1.9%** in DISTS.

Method	LPIPS ↓	DISTS ↓	CLIP ↑	PSNR ↑	SSIM ↑
pix2pix [28]	0.372	0.262	68.85	15.80	0.515
GNT [61]	0.482	0.392	63.75	14.61	0.538
DiT [39]	0.412	0.231	77.98	15.47	0.564
3DiM† [66]	0.385	0.226	78.22	15.91	0.575
4DIFF (ours)	**0.349**	**0.207**	**79.72**	**16.65**	**0.592**

Table 2: Comparison on the seen and unseen test sets of Ego-Exo4D-VT benchmark. † we reimplement 3DiM based on DiT as their code is not publicly available.

Split Setting	Method	LPIPS ↓	DISTS ↓	CLIP ↑	PSNR ↑	SSIM ↑
Seen Scenes	pix2pix [21]	0.371	0.260	68.68	15.90	0.519
	GNT [61]	0.479	0.390	63.44	14.71	0.542
	DiT [39]	0.406	0.226	78.74	15.64	0.570
	3DiM† [66]	0.365	0.217	78.30	15.98	0.583
	4DIFF (ours)	**0.316**	**0.184**	**82.79**	**17.09**	**0.600**
Unseen Scenes	pix2pix [21]	**0.376**	0.272	69.87	**15.23**	0.491
	GNT [61]	0.497	0.405	65.60	13.97	0.513
	DiT [39]	0.440	0.256	73.67	14.86	0.528
	3DiM† [66]	0.436	0.269	73.26	14.90	**0.542**
	4DIFF (ours)	0.427	**0.246**	**76.54**	14.45	0.508

and texture similarity between the synthesized egocentric image and the ground-truth image. Additionally, we include PSNR and SSIM for completeness, even though numerous existing works [7,51,53] have demonstrated that these metrics are suboptimal for evaluating image and video generation models, as they tend to favor conservative and blurry estimates.

4.2 Comparison with State-of-the-art Methods

In Table 1, we present the comparison of our method to various baselines. Notably, diffusion-based models—DiT [39], 3DiM [66], and our 4DIFF—outperform other approaches across all metrics by large margins, including the GAN-based pix2pix and NeRF-based GNT. The poor performance of the NeRF-based method GNT on our benchmark can be attributed to itslimited capacity for modeling hundreds of different scenes.

In Table 2, we present the results on seen scenes and unseen scenes respectively and show that our method achieves the best performance. Overall, our method surpasses the second-best performing diffusion-based 3DiM by **3.6%**

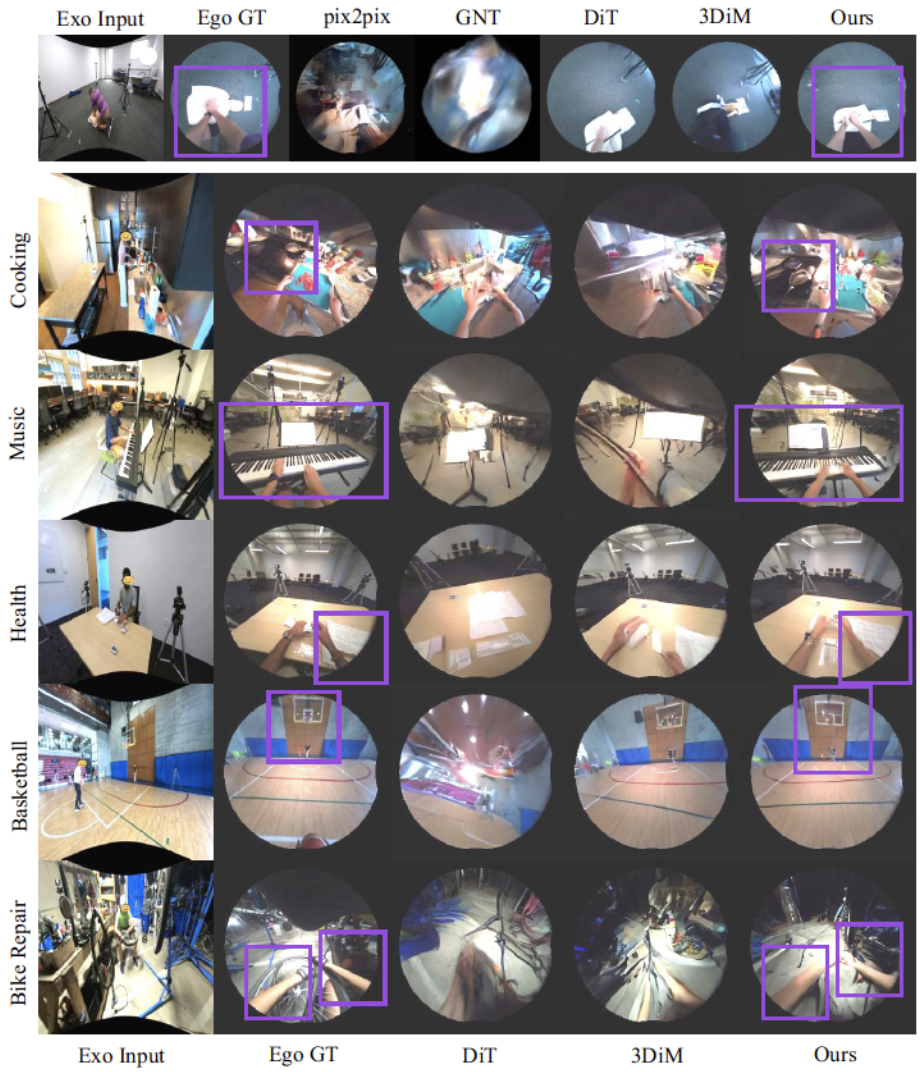

Fig. 5: Generated samples from five scenarios: cooking, music, health, basketball, and bike repair. Our 4DIFF demonstrates the best performance across all examples in terms of geometry correctness and object quality. We brighten the images and exclude pix2pix and GNT in the scenario breakdown for a better visual experience.

in LPIPS and **1.9%** in DISTS, underscoring the effectiveness of our proposed geometry-based approach.

Figure 5 presents qualitative comparisons with existing methods. GAN-based pix2pix [21] and NeRF-based GNT [61] exhibit challenges in producing photorealistic images, emphasizing the necessity of a robust generative model for the Ego-Exo4D-VT benchmark. Our 4DIFF demonstrates superior performance across various scenarios, excelling in both geometry correctness and object qual-

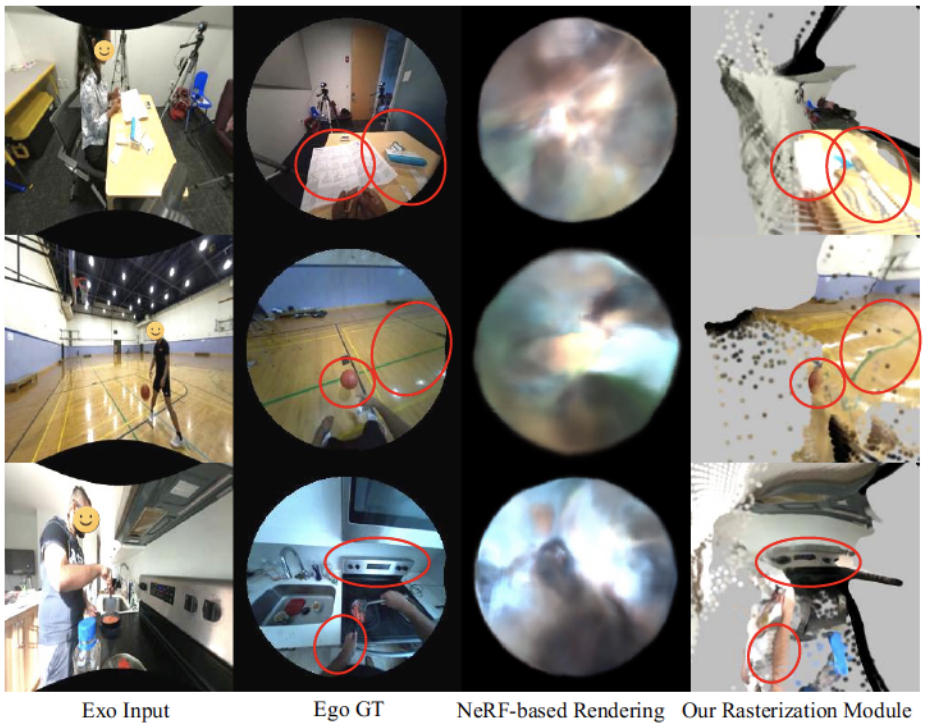

Fig. 6: We evaluate the effectiveness of our egocentric prior rendering module by visualizing the rendered prior image. Compared to NeRF-based rendering (GNT), our rendered prior image exhibits predominantly correct geometry, offering valuable egocentric cues to the diffusion model. Distortions and missing pixels arise from inaccurate depth estimation and occluded or unobserved regions in the exocentric view, which can be corrected by the diffusion model.

ity. Our 4DIFF is especially advantageous for view synthesis in complex scenes, such as the cooking scenario, where numerous objects exhibit intricate layouts. The qualitative results align well with our quantitative results in Table 1.

4.3 Qualitative Analysis

Investigating the visual results helps to gain a deeper insight into generative models. Thus, we perform a qualitative analysis below.

Is the egocentric prior useful? We address this question by visualizing the rendered egocentric prior RGB image. In Figure 6, the NeRF-based renderer GNT [61] generates blurry images for all scenes, possibly due to its limited capacity to model many diverse scenes with limited views for supervision. In contrast, our rendered egocentric images produced by point cloud rasterization are mostly correct, offering valuable egocentric cues to the diffusion model. Despite distortions and missing pixels, our diffusion model demonstrates sufficient capacity to rectify these issues effectively.

Fig. 7: Results on the unseen scenes. When synthesizing views from the scenes not encountered during training, our 4DIFF exhibits slight hallucinations but consistently outperforms existing methods, producing significantly improved results.

Generalization to unseen scenes. Figure 7 shows our generation results on the unseen scenes. We observe that our 4DIFF displays slight hallucinations, particularly noticeable in elements such as walls. Despite this, our method consistently outperforms existing methods. Such a robust performance can be attributed to the highly generalizable depth-based geometry priors used by our model.

What causes poor generation? We conduct an analysis to discern errors arising from the diffusion model or geometry priors. In Figure 8, we present two representative examples. The first showcases generation results in an unseen scene, where the egocentric prior image is reasonably good, but the diffusion model exhibits significant hallucinations, yielding an incorrectly generated image. We posit that this discrepancy arises because the diffusion model focuses

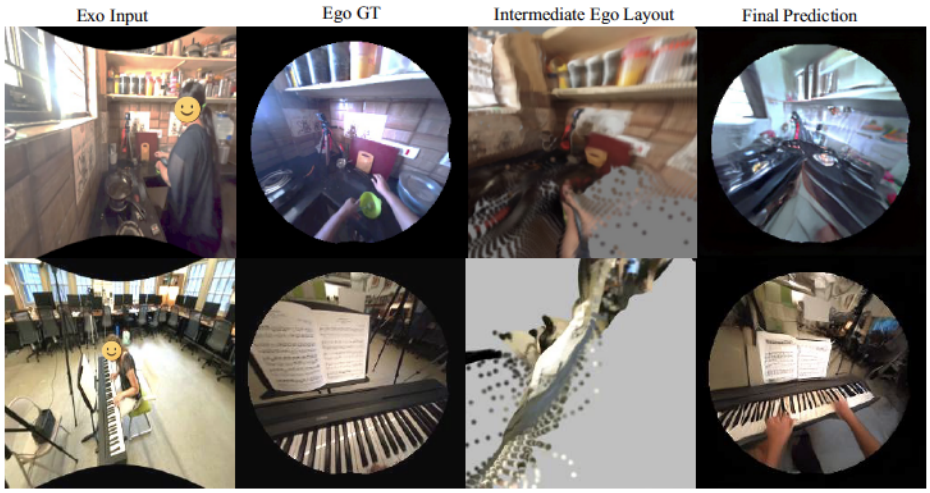

Fig. 8: Failure case examples of our method. Top: While the point cloud rasterization module performs effectively, the diffusion model produces errors when generating an egocentric view. Bottom: Although the diffusion model accurately predicts objects, the synthesized egocentric view appears more zoomed-out than the ground truth view. This can be attributed to suboptimal egocentric layout synthesis.

on modeling the conditional training distribution, limiting its generalization to substantially different scenes not present in the training data. This limitation can be mitigated by employing a large-scale pretrained diffusion model that has already acquired knowledge from diverse scenes and objects in 2D space.

In the second example, we show that despite the incorrectly rendered egocentric prior image, the diffusion model can generate a photorealistic image, which is more zoomed-out than the ground-truth egocentric image. This observation suggests that the diffusion model can robustly handle inaccurately generated egocentric geometry priors.

4.4 Ablation Studies

How important are our proposed modules? We study the importance of (i) 3D-aware rotary cross-attentions and (ii) egocentric point cloud rasterization by sequentially removing them from our framework. As shown in Tab. 3a, removing the 3D cross-attention worsens the LPIPS by **2.4%**. Additionally, removing the point cloud rasterization further degrades LPIPS by **3.9%**. Moreover, as shown in Figure 5, our 4DIFF with the proposed geometry priors consistently outperforms geometry-free diffusion models DiT and 3DiM in all scenarios. These results show the effectiveness of our proposed modules.

Can we pretrain the depth estimator from scratch? Tab. 3b shows that training our model without using a pretrained depth estimator results in a significant 4.3% degradation in LPIPS. This suggests that an inaccurate depth

Table 3: Ablation studies on various design choices. **(a)** We study the importance of each module by removing each module sequentially; **(b)** Using a pretrained depth estimator significantly improves the LPIPS by **4.3%**); **(c)** DINOV2 outperforms CLIP by **1.7%** in LPIPS.

(a) Module ablation.

Model	LPIPS ↓
4DIFF	**0.349**
− 3D Rotary CA	0.373
− ego rasterization	0.412

(b) Depth estimator.

Pretrained	LPIPS ↓
✓	**0.349**
✗	0.392

(c) Feature encoder.

Feat. Enc.	LPIPS ↓
DinoV2	**0.349**
CLIP	0.366

estimation may lead to most points from the exocentric view projected outside of the egocentric view. Consequently, these points will not receive sufficient gradient updates during training, leading to poor convergence. Thus, we conclude that a sufficiently accurate initial depth prediction is crucial for good performance.

Which feature encoder should we use? We evaluate two strong feature encoders for obtaining a semantic representation for an exocentric RGB image: DINOv2 [36], and CLIP [42], both employing a ViT-L/14 backbone. The DINOv2 variant outperforms the CLIP variant by **1.7%** LPIPS. We conjecture that compared to CLIP's vision-language pretraining, DINOv2's self-supervised pretraining leads to higher quality lower-level visual features which are important for exocentric to egocentric image translation problem.

5 Discussion and Conclusion

In this work, we proposed 4DIFF, a 3D-aware transformer-based diffusion model that significantly outperforms prior approaches on the challenging Ego-Exo4D-VT benchmark. Our method demonstrates robust generalization to novel environments not encountered during training. Despite our excellent results, we also acknowledge a few limitations. Firstly, our method assumes known camera poses during training and inference, limiting its applicability to real-world scenarios. Integrating camera pose estimation via a head pose estimator could address this limitation, while remains difficult to estimate automatically Secondly, our method focuses on image-to-image translation, leaving room for video generation by incorporating spatial-temporal cues. Thirdly, enhancing the quality of generated objects and improving generalization to unseen environments could be achieved by leveraging a more powerful pretrained diffusion model (e.g., Stable Diffusion [49]). Lastly, extending our framework from frame-level synthesis to object-level synthesis, considering the locations and appearances of objects such as hands and interacted objects, would bring it closer to real-world applications like AR/VR coaching. We plan to explore these research directions in our future work.

Acknowledgment We thank Hanwen Jiang, Yan-Bo Lin, Md Mohaiminul Islam, Ce Zhang, Yue Yang, and Soumitri Chattopadhyay for their helpful discussions. UT Austin is supported by NSF Grants AF 1901292, CNS 2148141, Tripods CCF 1934932, IFML CCF 2019844 and research gifts by Western Digital, Amazon, WNCG IAP, UT Austin Machine Learning Lab (MLL), Cisco, the Stanly P. Finch Centennial Professorship in Engineering. UNC is supported by Sony Faculty Innovation Award, Laboratory for Analytic Sciences via NC State University, ONR Award N00014-23-1-2356. K.G. is paid as a research scientist at Meta.

References

1. Ardeshir, S., Borji, A.: Ego2top: Matching viewers in egocentric and top-view videos. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 253–268. Springer (2016)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
4. Birkl, R., Wofk, D., Müller, M.: Midas v3.1 – a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
5. Cao, A., Rockwell, C., Johnson, J.: Fwd: Real-time novel view synthesis with forward warping and depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15713–15724 (2022)
6. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
7. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., De Mello, S., Karras, T., Wetzstein, G.: Generative novel view synthesis with 3d-aware diffusion models. arXiv preprint arXiv:2304.02602 (2023)
8. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
9. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
10. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. arXiv preprint arXiv:2304.06714 (2023)
11. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine **35**(1), 53–65 (2018)

12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
13. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
14. Duan, Y., Guo, X., Zhu, Z.: Diffusiondepth: Diffusion denoising approach for monocular depth estimation. arXiv preprint arXiv:2303.05021 (2023)
15. Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., Tucker, R.: Deepview: View synthesis with learned gradient descent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2367–2376 (2019)
16. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: Deepstereo: Learning to predict new views from the world's imagery. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5515–5524 (2016)
17. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: The kitti vision benchmark suite. URL http://www. cvlibs. net/datasets/kitti **2**(5) (2015)
18. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. arXiv preprint arXiv:2311.18259 (2023)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
20. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022)
21. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
22. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
24. Koh, J.Y., Lee, H., Yang, Y., Baldridge, J., Anderson, P.: Pathdreamer: A world model for indoor navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14738–14748 (2021)
25. Kulhánek, J., Derner, E., Sattler, T., Babuška, R.: Viewformer: Nerf-free neural rendering from few images using transformers. In: European Conference on Computer Vision. pp. 198–216. Springer (2022)
26. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10138–10148 (2021)
27. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023)
28. Liu, G., Tang, H., Latapie, H., Yan, Y.: Exocentric to egocentric image generation via parallel generative adversarial network. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1843–1847. IEEE (2020)

29. Liu, G., Tang, H., Latapie, H.M., Corso, J.J., Yan, Y.: Cross-view exocentric to egocentric video synthesis. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 974–982 (2021)
30. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9298–9309 (2023)
31. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019)
32. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In: ECCV (2024)
33. Mathews, J.: Coordinate-free rotation formalism. American Journal of Physics **44**(12), 1210–1210 (1976)
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
35. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. ACM Transactions on Graphics (ToG) **38**(6), 1–15 (2019)
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
37. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
38. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
39. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
40. Popov, S., Bauszat, P., Ferrari, V.: Corenet: Coherent 3d scene reconstruction from a single rgb image. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 366–383. Springer (2020)
41. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
43. Regmi, K., Borji, A.: Cross-view image synthesis using conditional gans. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3501–3510 (2018)
44. Ren, B., Tang, H., Sebe, N.: Cascaded cross mlp-mixer gans for cross-view image translation. arXiv preprint arXiv:2110.10183 (2021)
45. Ren, X., Wang, X.: Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3563–3573 (2022)
46. Riegler, G., Koltun, V.: Free view synthesis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16. pp. 623–640. Springer (2020)

47. Riegler, G., Koltun, V.: Stable view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12216–12225 (2021)
48. Rockwell, C., Fouhey, D.F., Johnson, J.: Pixelsynth: Generating a 3d-consistent experience from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14104–14113 (2021)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
50. Rombach, R., Esser, P., Ommer, B.: Geometry-free view synthesis: Transformers and no 3d priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14356–14366 (2021)
51. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
52. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
53. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
54. Sajjadi, M.S., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lučić, M., Duckworth, D., Dosovitskiy, A., 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. pp. 6229–6238 (2022)
55. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)
56. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2437–2446 (2019)
57. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in Neural Information Processing Systems **32** (2019)
58. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (October 2020), `https://arxiv.org/abs/2010.02502`
59. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
60. Szymanowicz, S., Rupprecht, C., Vedaldi, A.: Splatter image: Ultra-fast single-view 3d reconstruction. arXiv preprint arXiv:2312.13150 (2023)
61. T, M.V., Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z.: Is attention all that neRF needs? In: The Eleventh International Conference on Learning Representations (2023), `https://openreview.net/forum?id=xE-LtsE-xx`
62. Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J.J., Yan, Y.: Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2417–2426 (2019)

63. Trevithick, A., Yang, B.: Grf: Learning a general radiance field for 3d scene representation and rendering (2020)
64. Tucker, R., Snavely, N.: Single-view view synthesis with multiplane images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 551–560 (2020)
65. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)
66. Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel view synthesis with diffusion models. arXiv preprint arXiv:2210.04628 (2022)
67. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7467–7477 (2020)
68. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 (2024)
69. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
70. Zhai, M., Bessinger, Z., Workman, S., Jacobs, N.: Predicting ground-level scene layout from aerial imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 867–875 (2017)
71. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
72. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)