# Synthesizing Consistent Novel Views via 3D Epipolar Attention without Re-Training

Botao Ye[1,2] Sifei Liu[3] Xueting Li[3] Marc Pollefeys[1,4] Ming-Hsuan Yang[5]
[1]ETH Zurich [2]ETH AI Center [3]NVIDIA [4]Microsoft [5]UC Merced

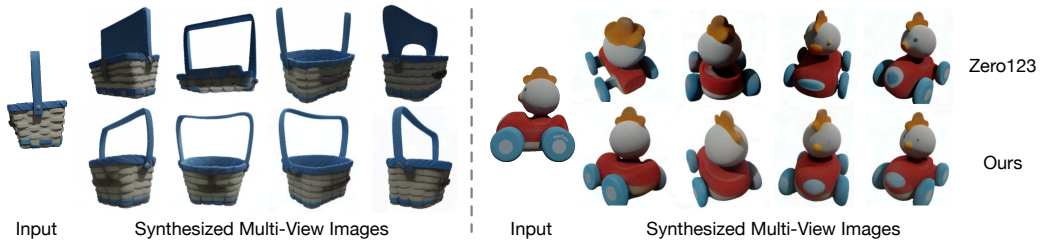


Figure 1. Given an input image and a sequence of relative camera pose transformations, our method synthesizes more consistent novel view images. Our method *does not need to re-train* the baseline model (Zero123) and *supports arbitrary relative camera poses*.

## Abstract

*Large diffusion models demonstrate remarkable zero-shot capabilities in novel view synthesis from a single image. However, these models often face challenges in maintaining consistency across novel and reference views. A crucial factor leading to this issue is the limited utilization of contextual information from reference views. Specifically, when there is an overlap in the viewing frustum between two views, it is essential to ensure that the corresponding regions maintain consistency in both geometry and appearance. This observation leads to a simple yet effective approach, where we propose to use epipolar geometry to locate and retrieve overlapping information from the input view. This information is then incorporated into the generation of target views, eliminating the need for training or fine-tuning, as the process requires no learnable parameters. Furthermore, to enhance the overall consistency of generated views, we extend the utilization of epipolar attention to a multi-view setting, allowing retrieval of overlapping information from the input view and other target views. Qualitative and quantitative experimental results demonstrate the effectiveness of our method in significantly improving the consistency of synthesized views without the need for any fine-tuning. Moreover, This enhancement also boosts the performance of downstream applications such as 3D reconstruction. The code is available at* https://github.com/botaoye/ConsisSyn.

## 1. Introduction

Synthesizing high-quality novel view images from a single input image is a long-standing and challenging problem. It requires inheriting the appearance of objects in the observed regions of the input image while also hallucinating unseen regions. Recent studies [13, 38] approach this problem as an image-to-image translation task and implement it using diffusion models [10, 27], drawing inspiration from their successful application in 2D image generation [21, 23]. While they exhibit remarkable zero-shot capabilities when trained with large-scale 2D and 3D datasets, they still face challenges in maintaining 3D consistency between the target view and the generated multi-view images, due to the probabilistic nature of diffusion models. This limitation adversely affects downstream applications such as 3D reconstruction [19, 34].

In this paper, we propose to improve the consistency of synthesized multi-view images by optimizing the utilization of reference image information. Notably, maintaining consistency between the generated image and the corresponding observed regions in the input view is a crucial requirement in the task of single-image conditioned novel view synthesis. However, existing methods often overlook this constraint by merely considering the input image as a condition or network input, which fails to guarantee such consistency. One straightforward method to fulfill this constraint is by warping the content from the input to the target view and subsequently conducting outpainting for the remaining regions [41, 44]. However, 3D warping relies on precise depth information, which is hard to obtain. Additionally, direct warping struggles with occlusion and illumination variations across different views.

We aim to utilize this constraint to improve the consistency in a more adaptable way. Despite the intricacies of obtaining depth, we can still reduce the search space for locating corresponding points by incorporating other 3D geometric priors. As depicted in Fig. 2, the corresponding

points in the reference views must be on the epipolar line. Therefore, we propose an epipolar attention module to locate and gather contextual information. For each point in the target view visible in the reference view, we can first constrain the corresponding point to its respective epipolar line in the reference image. Subsequently, we ascertain the corresponding location along the epipolar line by feature matching. The features at the localized positions are then retrieved and used to constrain the target view generation.

More specifically, we first perform DDIM inversion on the input view and reconstruct the input image using the initial noise provided by the DDIM inversion. This process yields intermediate features of the input view, which can then be employed to constrain the generation of target views. Then, in the epipolar attention module, we traverse the corresponding epipolar line in the input view for every point within the target view. During this process, we compute the similarity between the features of the target point and those sampled from the input view. This similarity score is then used to aggregate the corresponding features from the input view. This soft operation is more adept at handling complex scenarios, such as occlusion (detailed analysis can be found in the Supplementary Material). Additionally, to avoid any parameter training or fine-tuning, we employ a simple parameter duplication strategy, *i.e.*, we copy all parameters directly from the self-attention layer to obtain the epipolar attention parameters. To further improve the consistency between different target views, we expand the application of epipolar attention to a multi-view context. Specifically, we generate multiple target views in an autoregressive manner. When generating a specific novel view, we consider the input view and previously generated target views close to the current viewpoint as context views. We employ epipolar attention to aggregate overlapping information from all context views, rather than solely from the input view, thereby improving consistency among all generated views. It is worth mentioning that our epipolar attention reduces the search space compared to locating corresponding points in the full image. Therefore, it requires much less memory when retrieving information from multiple views, making it more friendly to GPUs with small memory capacity.

We conduct experiences on the Google Scanned Objects [6] dataset to verify the zero-shot novel view synthesis capability and evaluate our method on both generated image quality and the view consistency [38]. Additionally, we apply our method to the downstream 3D reconstruction task [34] and compare it against the mesh constructed by our baseline model.

The main contributions of this work are:

- We propose a novel epipolar attention method to locate and retrieve the corresponding information in the reference view, which is then inserted into the generation process of the target view to enhance the consistency between multi-view images.
- Experimental results show that our method effectively improves the consistency of the synthesized multi-view images without any training or fine-tuning while maintaining the quality of the generated images.
- We apply the synthesized multi-view images to a downstream 3D reconstruction task, and the results show that the more consistent images further improve the 3D reconstruction results.

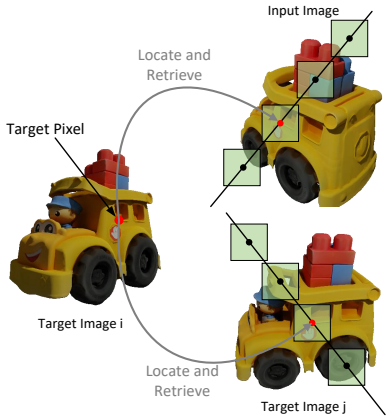


Figure 2. When the camera viewing frustum of two views overlaps, for a point on one of the images, we can find its correspondence on the epipolar line of the other view.

| | Zero123 | Fixed Camera | | | Additional Net | | | Ours |
|---|---|---|---|---|---|---|---|---|
| | Zero123 [13] | SyncDreamer [14] | MVDream [25] | Zero123++ [24] | SparseFusion [46] | PGD [30] | Consistent123 [42] | Ours |
| 1) S.V. Condition | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ |
| 2) Generalizability | ✓ | ✓ | ✓ | ✓ | × | × | ✓ | ✓ |
| 3) No Extra Training | × | × | × | × | × | × | × | ✓ |
| 4) Multi-view Consis | × | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 5) Free Trajectory | ✓ | × | × | × | ✓ | ✓ | ✓ | ✓ |

Table 1. **Comparison with related methods.** Each row represents the ability to: 2) generalize well to arbitrary objects, 3) work without requiring extra retraining, 4) generate multi-view consistent images, and 5) generate images in arbitrary camera poses. See Sec. 4.4 for a detailed comparison.

## 2. Related Work

**Diffusion Models for Novel View Synthesis.** Diffusion models show impressive results on the text-to-image task [20, 21, 23]. Therefore, a line of work aims to extend it to the novel view synthesis (NVS) task, where they generate novel view images based on reference images and desired relative camera poses. Such synthesized multi-view images find utility in various applications such as distillation purposes [3, 12, 19, 33, 37], or for directly training NeRF-like 3D assets [15, 16, 34]. 3DiM [38] implements this idea by training a diffusion model conditioned on reference images and relative camera poses. SparseFu-

sion [46] and GeNVS [2] first generate course latent feature of the target view as additional input to the diffusion model. However, these methods are trained on objects from specific classes or relatively small datasets, making it challenging to generalize to arbitrary objects. Zero123 [13] obtains impressive zero-shot generalizability by fine-tuning a 2D diffusion model, *i.e.*, Stable Diffusion [21], on a large-scale 3D rendered dataset [4]. However, novel view images generated by Zero123 can suffer from consistency problems, especially when relatively large pose transformations are present. To address this issue, some very recent studies [14, 24, 25, 39, 42] try to add additional modules and fine-tune the Zero123 or LDM model to obtain better consistency, which requires significant computational resources. In contrast to these approaches, we focus on enhancing the consistency of pre-trained models without the need for any fine-tuning. Tab. 1 provides an overview comparison, while Sec. 4.4 offers a detailed comparison.

**Image-to-Image Translation.** Image-to-image translation (I2I) involves learning a mapping from an input image to an output image while preserving specific properties like the scene layout or object structure. Our paper's primary focus can be viewed as an I2I task, where the condition is the pose, aiming to transform the input image to the desired pose. One of the main challenges in the pose-guided novel view generation task is maintaining consistency between the target images and the input image. This challenge shares similarities with the issues encountered in text-guided image-to-image translation tasks [1, 9, 11, 17, 31]. For instance, works such as [1, 9, 31] manipulate self-attention, cross-attention, or spatial features within the U-Net [22] structure to preserve the desired concept in the input image. However, these methods primarily target 2D image translation or editing tasks, lacking 3D structural information and struggling to discern what to preserve or discard in the context of the NVS task. In contrast, our method incorporates 3D geometry information into the translation process to better preserve the desired information in the input view.

**Epipolar Geometry in DNN.** Epipolar geometry is used in many previous works [8, 28, 30, 35, 40]. They often integrate epipolar geometry into network modules and employ it for network training. In contrast, we use the epipolar geometry to generate images without training or fine-tuning to localize better and retrieve the corresponding information using the features from a trained diffusion model.

## 3. Preliminaries

In this section, we revisit the pose-conditional diffusion model used in our approach (Sec. 3.1), and the DDIM inversion technique used to invert the reference image back to the initial Gaussian noise (Sec. 3.2).

### 3.1. Pose-Conditioned Diffusion Model

Diffusion models [5, 10, 26, 27] are probabilistic generative models, which transform an initial Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$ into an arbitrary meaningful data distribution. During training, the diffusion *forward* process is applied, in which Gaussian noise is added to the clean data $\boldsymbol{x}_0$ (image in our case):

$$\boldsymbol{x}_t = \sqrt{\alpha_t} \cdot \boldsymbol{x}_0 + \sqrt{1-\alpha_t} \cdot \boldsymbol{z}, \tag{1}$$

where $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$ is the random noise and $\{\alpha_t\}, t \in [0, T]$ is the noise schedule indexed by time step $t$. During inference, the *backward* diffusion process is utilized to progressively denoise $\boldsymbol{x}_T$ to obtain the clean data. This denoising process is facilitated by a neural network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$, which predicts noise at each step.

We focus on employing the diffusion model for synthesizing novel views from a single input view. This can be seen as an image-to-image translation process that transforms the original image into a novel view image based on their relative camera pose transformation. Formally, given the reference view image $\boldsymbol{x}_r$ and the relative camera pose transformation $\Delta\boldsymbol{p} = (\mathbf{R}, \mathbf{T})$ between the reference view and the target view, the denoising network predicts noise conditioned on both $\boldsymbol{x}_r$ and $\Delta\boldsymbol{p}$, denoted as:

$$\boldsymbol{z}_t = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t \mid \boldsymbol{x}_r, \Delta\boldsymbol{p}). \tag{2}$$

In this work, we leverage a pre-trained pose-conditioned diffusion model (Zero123 [13]), which in turn is fine-turned from a Latent Diffusion Model [21]. The network is implemented using a U-Net [22] structure, consisting of several residual blocks [7], self-attention blocks, and cross-attention blocks [32]. At each time step $t$, the feature maps from the previous layer $l-1$ are first feeded in the residual block to obtain feature $\boldsymbol{F}_t^l$. Subsequently, projection layers are employed to generate distinct query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$ feature maps (for simplicity, excluding time step $t$ and layer index $l$). The output feature of the self-attention block, denoted as $\hat{\boldsymbol{F}}$, is computed using the operation $\hat{\boldsymbol{F}} = \boldsymbol{A} \cdot \boldsymbol{V}$, where the attention matrix $\boldsymbol{A}$ is determined as follows:

$$\boldsymbol{A} = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right), \tag{3}$$

where $d$ is the feature dimension.

### 3.2. DDIM Inversion

During *backward* diffusion, deterministic DDIM sampling [27] is commonly used to convert noise $\boldsymbol{x}_T$ into clean data $\boldsymbol{x}_0$. In contrast, DDIM inversion [5, 27] converts the original clean image data $\boldsymbol{x}_0$ back to Gaussian noise $\boldsymbol{x}_T$ by incrementally adding the noise predicted by the network $\boldsymbol{\epsilon}_\theta$. We also employ DDIM inversion to convert the input image to its initial noise $\boldsymbol{x}^R$. Throughout this conversion, we utilize the input image feature and a fixed relative pose transformation of $[0, 0, 0]$ as the network condition.

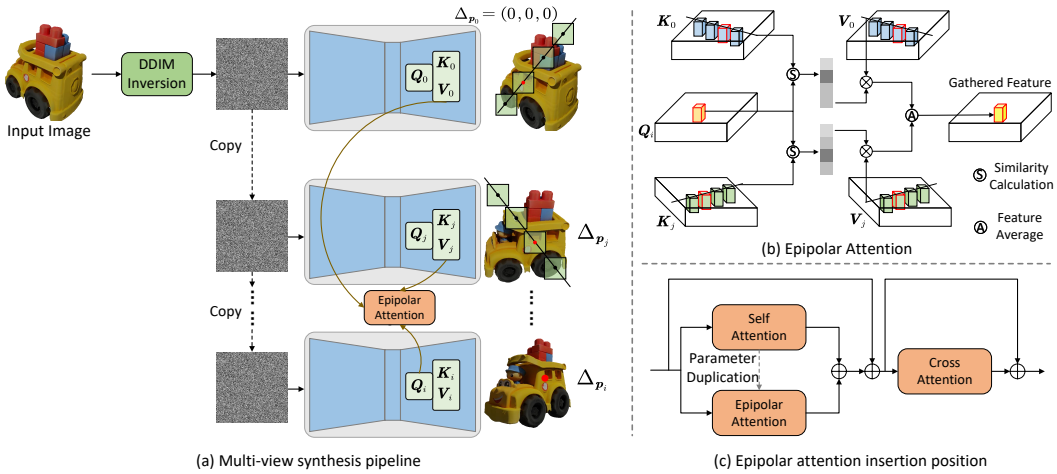


Figure 3. **Overview of our method.** (a) We first perform DDIM inversion on the input image to obtain the initial noise, which is shared during the multi-view image generation process. Throughout the generation of each view, our epipolar attention block efficiently locates and retrieves corresponding information from both the input image and other target views. (b) The architecture of our 3D epipolar attention module. (c) Location of our inserted epipolar attention block.

# 4. Approach

As mentioned above, our goal is to improve the consistency of the synthetic multi-view image by locating and retrieving the corresponding information (features) in the reference view that overlap with the target view and then using the retrieved features to constrain the target view generation process. Therefore, we first explicate the methodology for computing the epipolar line and sampling points along it to effectively reduce the searching space concerning the corresponding locations (Sec. 4.1). Then, we will describe how to locate the corresponding locations along the epipolar line (Sec. 4.2). The general idea is to find the correspondence between a point in the target view and sampled points on the epipolar line through feature similarity. To better obtain the similarity information, we analyze the attributes of different features in the U-Net block and then find the appropriate features used to compute the similarity. Next, we introduce the parameter duplication strategy that facilitates the training-free module, and how to inject the retrieved reference features to constrain the generation process of the target views (Sec. 4.3). Finally, we will provide a detailed analysis of our epipolar attention (Sec. 4.4). Fig. 3 shows the overall framework of our tuning-free multi-view epipolar attention that enables consistent novel view synthesis.

## 4.1. Point Sampling from Epipolar Lines

Ideally, with the target view's depth map, we can accurately find its correspondence in the reference view by unprojecting each point to 3D space and then re-projecting it into the reference view. However, obtaining an accurate depth value for arbitrary real-world objects remains challenging, if not infeasible. Alternatively, when considering a point $\boldsymbol{p}_i$ in the target view, its corresponding point $\boldsymbol{p}'_i$ in the reference view, if visible, must lie on the corresponding epipolar line $\mathbf{l}_i$. Therefore, we opt to find the corresponding feature in the reference view along the epipolar line, which significantly reduces the search space and the memory required for subsequent computations.

We assume that the synthesized novel view images have the same camera intrinsic parameters $\mathbf{K}$ as the reference image, as being commonly set for the NVS task. Specifically, given the relative camera rotation $\mathbf{R}$ and translation $\boldsymbol{t}$ from the reference image to the target image, for each point $\boldsymbol{p}_i$ in the target image, the corresponding epipolar line $\mathbf{l}_i$ is:

$$\boldsymbol{l}_i = \mathbf{R}[\boldsymbol{t}]_\times \mathbf{K}^{-1}\boldsymbol{p}_i, \tag{4}$$

where $\mathbf{l}_i$ is the epipolar line of $\boldsymbol{p}_i$ in the reference image, and $[\boldsymbol{t}]_\times$ is the skew-symmetric matrix representation of $\boldsymbol{t}$. Despite the unknown exact camera focal length $f$, the computation of the epipolar lines $\mathbf{l}_i$ remains feasible, as the computation can be independent of $f$ (see Supplementary Material for proof).

Subsequently, we sample a set of points denoted as $\boldsymbol{p}' \in P'$ along the epipolar line, specifically along the direction of the image width, at intervals of each feature pixel. Note that some sample points may be outside the feature plane and will be masked during the similarity calculation and feature retrieval process.

## 4.2. Corresponding Point Searching

**Paired Feature Acquiration.** The epipolar sampling operation essentially reduces the search space of the corresponding points. However, how to more accurately locate the actual corresponding point in the epipolar line remains unsolved. Previous works [29, 43] show that the features extracted by diffusion models show good semantic correspondence between two input images. Thus, a plausible approach is to seek the corresponding position in the reference image for each pixel in the target view by assessing

the similarity between their respective features. However, previous feature matching methods [29, 43] require feeding two paired images into the diffusion model separately and extracting their features for matching, making them unsuitable for our scenario where the target image is pending generation. To address this, we employ DDIM inversion, as detailed in Sec. 3.2, to acquire noise from the reference image. This noise is then utilized to concurrently reconstruct the reference image alongside the denoising process of the target image, which we used to obtain the paired features. Specifically, we progressively denoise the DDIM inversed initial noise $\boldsymbol{x}^R$ of the reference image using DDIM sampling and set the relative camera pose transformation as $[0, 0, 0]$ so that the reference image can be recovered. Meanwhile, we use the same $\boldsymbol{x}^R$ as the initial noise to generate the target view. We can then obtain paired features by retrieving the features in the same denoising step and at the corresponding layer of both the input and target generation branches. Since the sampled point in the epipolar line is in the sub-pixel location, we employ bilinear interpolation to obtain the feature value of each point $\boldsymbol{p}'_i$ in the epipolar line. We then analyze the similarity of the corresponding intermediate feature of the target and reference branch.

**Computing Epipolar Attention.** We now have access to the paired features of the input and target images. However, the specific features within the U-Net structure to utilize, as well as the methodology for calculating their similarity, remain unclear. Previous feature matching methods [29, 43] calculate the similarity of output feature $\boldsymbol{F}$ of the attention block use cosine similarity $\mathrm{CosSim}(\boldsymbol{F}_{tgt}(\boldsymbol{q}), [\boldsymbol{F}_{ref}(\boldsymbol{q}')])$ ($[\cdot]$ is the bilinear interpolation operation), followed by a softmax operation. However, our experiments reveal that the similarity derived from these features does not align well with our intended application. The resulting similarity map exhibits a relatively uniform distribution, indicating insufficient localization of the desired corresponding location and inadequate corresponding feature aggregation (see Fig. C.1. in the supp. mat.). It is important to note that the query and key features employed within the multi-head self-attention block are intended for similarity calculation, so we opt to use the query $\boldsymbol{Q}$ from the target branch and the key $\boldsymbol{K}$ from the reference branch to compute the similarity according to Eq. 3. Such similarity scores can pinpoint the corresponding location (see Fig. C.1. in the supp. mat.).

## 4.3. Reference Feature Injection

After finding the location of the corresponding point, we introduce how to use such information to constrain the generation process of the target image. First, to neglect the necessity of further training or fine-tuning, we employ a simple parameter duplication strategy, as shown in Fig. 3(c), in which *we directly instantiated the epipolar attention block with the well-trained parameters of the self-attention block*.

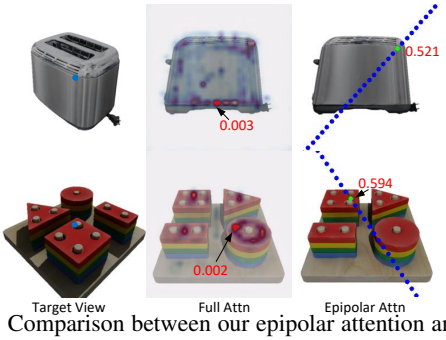


Figure 4. Comparison between our epipolar attention and the full attention. Our epipolar attention better locates and retrieves the corresponding information in the reference view.

Similar to the attention operation [32], we use the weighted sum to aggerate the corresponding information in the reference image as follows:

$$\hat{\boldsymbol{F}}_{src} = \sum_{p' \in \mathcal{P}'} \mathrm{sim}\left(\boldsymbol{Q}_{tgt}(p), \boldsymbol{K}_{ref}(p')\right) \cdot \boldsymbol{F}_{\mathrm{src}}\left(p'\right), \quad (5)$$

where $\mathrm{sim}(\cdot)$ is the similarity calculation operation as Eq. 3. Similar to the residual connection in the original U-Net block, we fuse the output feature from our epipolar attention block with the original self-attention block with a predefined weight parameter $\alpha$, which can be formulated as $\boldsymbol{F} = \alpha \hat{\boldsymbol{F}}_{src} + (1-\alpha)\hat{\boldsymbol{F}}$.

**Attending Multi-Views at Once.** While aggregating the overlapping feature from the input view improves the consistency between the output view and the input view, the consistency between different target views still is not well preserved as there are regions in the target images that are not visible to the input views. We further extend the epipolar attention to the multi-view setting to address this issue. Specifically, we generate multiple views $\Delta\boldsymbol{p}_i, i \in [1, N]$ in an auto-regressive manner. When synthesizing a specific novel view $\Delta\boldsymbol{p}_i$, we designate it as the target view. $M$ previous views, along with the input view, are considered context views, collectively containing specific information that overlaps with the target view. Subsequently, we apply epipolar attention to all context views and compute the average features derived from these views. Fig. 3 (a) shows an example synthesis process for view $\Delta\boldsymbol{p}_i$.

## 4.4. Discussion About the Epipolar Attention

**Comparison with Full Image Attention.** An alternative to our epipolar attention mechanism is directly using full attention to gather corresponding information in the reference view, which finds the corresponding points in the full image. In contrast, our epipolar attention significantly reduces the search space for the corresponding point searching by introducing additional geometric priors. Illustrated in Fig. 4, our method exhibits sharper similarity scores and more precise localization of corresponding positions, resulting in a more effective retrieval of desired corresponding features. Thus, as shown in Tab. 5, epipolar attention per-

forms better than full attention, especially when multiple reference views are employed. Additionally, by reducing the search space, our epipolar attention significantly decreases memory consumption during the feature retrieval process. The space and time complexity of the epipolar attention is $O(L^3)$, while that of the full attention is $O(L^4)$, where $L$ is the length of the feature map.

**Comparison with Recent Methods.** Some recent works, such as MVDream [25], SyncDreamer [14], and Zero123++ [24], also aim to improve the consistency of synthesized multi-view images. However, these methods require time-consuming re-training. Moreover, they constrain the camera pose during training, limiting their ability to synthesize images to a fixed set of camera poses. For example, MVDream [25] can only synthesize images with four fixed camera views. In contrast, our method can synthesize consistent multi-view images with arbitrary camera poses without re-training.

Previous work, *i.e.*, PGD [30] also utilizes epipolar attention in the generation task. However, it differs from our method mainly in two aspects. **1)** Our method aims to enhance baseline model consistency without tuning, while PGD treats epipolar attention as a network module requiring full network training, making it resource-intensive. These differences also lead to problem formulation in using epipolar constraints. PGD computes per-pixel distances to the epipolar line as an additional weight map multiplied by the original attention matrix, thereby altering the original distribution of attention weights. Consequently, this approach is not suitable for a non-training pipeline. In contrast, we aim to *locate* and *retrieve* corresponding information from the reference views using the epipolar constraint to roughly approximate the correspondence, followed by sampling and soft fine-locating. Thus, we avoid the need for time-consuming retraining. Furthermore, our method reduces GPU memory consumption compared to PGD, as PGD still utilizes full attention. Inserting PGD's epipolar module into our pipeline yields inferior results and has no significant improvements over full attention (see Tab.4 and Tab. 5). **2)** To make the whole pipeline work without any fine-tuning, we invest considerable effort in its design, which is not explored in PGD. For instance, we provide insights into how to generate input view features based on pre-trained Zero123, determine appropriate features for similarity computation, and how to extend epipolar attention to multi-view setting.

# 5. Experiments

## 5.1. Experimental Setups

**Dataset.** Following previous work [13, 14], we evaluate our work on the Google Scanned Object (GSO) [6] dataset to verify the zero-shot novel view image synthesis capability. We also provide results for additional datasets in the Supplementary Material. Specifically, we randomly select 30 objects from the GSO dataset with various object categories. Unlike recent approaches [14, 25] that aim to enhance the consistency of novel view synthesis models by generating multiple fixed-view images, our method can generate images from any camera pose and any number of views. Therefore, we conduct experiments under different camera pose settings to validate our approach: specifically, 1) *16-views with free camera pose*: for each object, we circularly render 16 views with the elevation angles ranging in $[-10°, 40°]$ and the azimuth angles are evenly distributed in $[0°, 360°]$. 2) *16-views with fixed camera pose*: We maintain a constant elevation angle of $30°$ and uniformly sample azimuth angles (same as SyncDreamer [14]). 3) *32-views with free camera pose*: Similar to the first setting, but we sample 32 views. It's important to note that our method does not require additional training or fine-tuning on any datasets.

**Metrics.** To validate the effectiveness of our method, we mainly evaluate it based on three criteria: 1) *Quality Score*. We evaluate the image quality of synthesized multi-view images by measuring their similarity with ground truth images. Following prior research [13, 46], we report the similarity between the synthesized images and the ground truth images with standard metrics: PSNR, SSIM [36], and LPIPS [45]. 2) *Multi-view Consistency Score*. As the primary goal of our work is to improve the consistency of generated images, we also employ the 3D consistency score [38] to verify the consistency among the synthesized images. Specifically, we train an Instant-NGP [18] with the input image and part of the synthesized novel view images of our model and evaluate the similarity between the remaining synthesized images and the rendered images of Instant-NGP. For the synthesized multi-view images of each object, we allocate $3/4$ for training and reserve the remaining $1/4$ for validation. Intuitively, if the consistency of synthesized images is improved, the NeRF-like model will train a better object representation, and the re-rendered images will agree more with the validation images. 3) *Input Consistency Score*. To assess the faithfulness of synthesized images in preserving the identity of the input condition image, we introduce the input consistency score. This score calculates the similarity of each synthesized image with the input condition image, utilizing the LPIPS metric.

In addition, we use synthesized multi-view images to train a neural 3D reconstruction model (NeuS [34]) and report commonly used Chamfer Distances (CD) and Volume IoUs between the trained 3D model and the ground truth.

**Baselines.** Given that our main goal is to improve the consistency of the trained baseline model without further fine-tuning, we mainly compare our approach with the used baseline model Zero123 [13]. Additionally, we compare our method to the SOTA approaches such as PGD [30] and

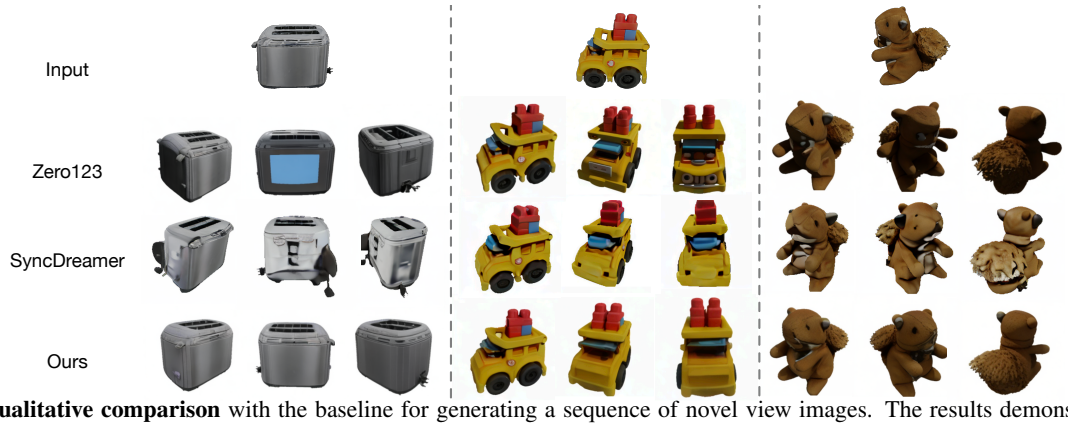


Figure 5. **Qualitative comparison** with the baseline for generating a sequence of novel view images. The results demonstrate that our method synthesizes more consistent multi-view images compared to our baseline model (Zero123). In addition, compared to SyncDreamer, our method visually maintains better similarity to the conditioned image and appears more natural.

Table 2. Comparison of multi-view consistency, image quality, and input consistency of synthesized multi-view images at the 16-view setting with free camera pose.

| | Multi-view Consistency | | | Quality Score | | | Input Consis. |
|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | LPIPS↓ |
| Zero123 | 15.225 | 0.645 | 0.408 | 14.255 | 0.747 | 0.208 | 0.303 |
| SyncDreamer | 14.830 | 0.626 | 0.434 | 12.650 | 0.713 | 0.254 | 0.317 |
| Ours | **18.300** | **0.734** | **0.355** | **14.947** | **0.763** | **0.191** | **0.282** |

Table 3. Comparison of multi-view consistency, image quality, and input consistency at the 16-view setting with fixed camera pose as SyncDreamer [14].

| | Multi-view Consistency | | | Quality Score | | | Input Consis. |
|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | LPIPS↓ |
| Zero123 | 16.556 | 0.682 | 0.378 | 14.592 | 0.750 | 0.207 | 0.305 |
| SyncDreamer | **22.424** | **0.812** | **0.268** | 15.269 | 0.749 | 0.196 | 0.300 |
| Ours | 21.151 | 0.780 | 0.302 | **15.293** | **0.764** | **0.184** | **0.287** |

Table 4. Comparison of multi-view consistency and image quality scores of synthesized multi-view images at the 32-view setting with free camera pose.

| | Multi-view Consistency | | | Quality Score | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| Zero123 | 16.515 | 0.694 | 0.378 | 15.142 | 0.733 | 0.211 |
| PGD [30] | 18.481 | 0.720 | 0.343 | 15.281 | 0.739 | 0.205 |
| Ours | **20.655** | **0.792** | **0.305** | **15.268** | **0.742** | **0.203** |

SyncDreamer [14] using the same Zero123 base model.

**Implementation Details.** We use the official checkpoint provided by Zero123 [13], which is trained on objaverse [4] for 165,000 steps. We inject our epipolar attention layer after step $T = 4$ and layer $L = 10$ by default. We find that feature fusion weight $\alpha = 0.5$, and the number of context views $M = 2$ work better.

## 5.2. Comparison With Baseline Models

The quantitative comparison on three settings are shown in Tab. 2, Tab. 3, and Tab. 4. The qualitative comparison is shown in Fig. 5.

**Multi-view Consistency.** Tab. 3 presents the 3D consistency scores compared to our baseline model (Zero123) and SyncDreamer. The results indicate a significant improvement across all three metrics achieved by our method when compared with Zero123. While our method exhibits a marginally lower numerical consistency score compared to SyncDreamer, it enables the synthesis of images with arbitrary camera poses. This capability is illustrated in Tab. 2, where our method consistently enhances consistency with changes in camera pose settings, whereas SyncDreamer fails to do so and exhibits inferior results compared to Zero123. Furthermore, our method facilitates the synthesis of multi-view images with any number of camera views. This versatility is demonstrated in Tab. 4, where our method continues to achieve significant improvements in consistency scores, while SyncDreamer is unable to operate under such conditions.

Table 5. Ablation Study on consistency score and quality score. Each of the different design choices is added to the baseline model.

| | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| Baseline (Zero123) | 16.515 | 0.694 | 0.378 |
| + Full Attention (Single) | 18.208 | 0.749 | 0.346 |
| + Epipolar Attention (Single) | 18.514 | 0.761 | 0.342 |
| + Full Attention (Multi) | 19.511 | 0.784 | 0.312 |
| + Epipolar Attention (Multi) | **20.655** | **0.792** | **0.305** |

Meanwhile, Fig. 5 provides a qualitative comparison with the baseline. While both our method and SyncDreamer enhance consistency, our method visually preserves better similarity to the input image, including color and texture details. The input consistency score further corroborates this.

**Image Quality.** While our primary goal centers around enhancing the consistency of synthesized multi-view images, we also evaluate the image quality by comparing the similarity with the ground truth images. The results shown in Tab. 2, Tab. 3, and Tab. 4 indicate that our method also enhances the image quality under different settings besides improving the consistency. Moreover, our method shows better image quality compared with SyncDreamer even in the 16-view setting with fixed camera pose.

**Input Consistency.** Input consistency terms whether the results align with the input image. Fig. 5 illustrates that both our method and SyncDreamer enhance multi-view consistency. However, the color and texture details of SyncDreamer's results diverge from the input image and appear

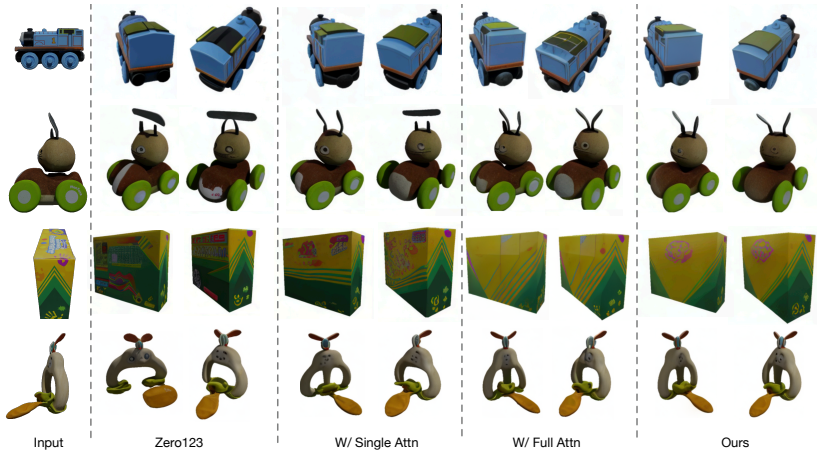


Figure 6. Qualitative Comparison for different design choices. Our method, employing multi-view epipolar attention, demonstrates the best consistency.

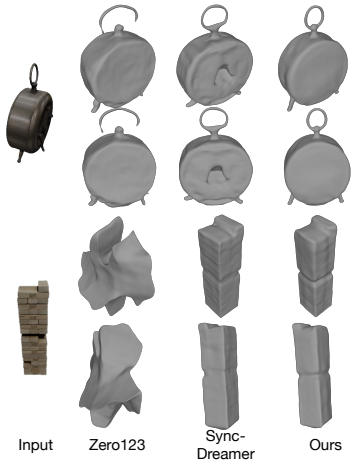


Figure 7. Our method shows better direct 3D reconstruction [34].

visually unnatural. This discrepancy is evident in the input consistency score presented in Tab. 3, indicating lower similarity with the condition image in the SyncDreamer results.

### 5.3. Ablation Study

The overall quantitative results are shown in Tab. 5, and the qualitative comparisons are shown in Fig. 6.

**Full Attention *vs*. Epipolar Attention.** The results presented in Tab.5 and Fig.6 demonstrate that our epipolar attention mechanism can synthesize more consistent multi-view images compared with full attention. Furthermore, our epipolar attention achieves a greater performance improvement compared to full attention when using multiple reference images. This could be attributed to the fact that our epipolar attention more effectively localizes target information, as depicted in Fig. 4, thereby reducing noise from the reference images. In the multi-view setting, where multiple reference images are utilized, this noise reduction becomes particularly crucial. Moreover, it is noteworthy that the epipolar attention mechanism consumes less GPU memory compared to our baseline, as discussed in Sec. 4.4.

**Attending Single-View *vs*. Multi-View.** Applying the epipolar attention significantly improves the consistency between the input and target views. However, the consistency between different views in the unobserved regions of the input view is not well preserved. After implementing our epipolar attention in the multi-view setting, the consistency across the generated multi-view images is further improved. The last row in Tab. 5 shows that after applying our multi-view epipolar attention, the consistency score is further improved compared with the single-view setting. Besides, the qualitative result in Fig. 6 also shows better consistency among different target views.

### 5.4. Downstream Application

To demonstrate the effectiveness of our method, we also applied it to the downstream 3D reconstruction task. Specifically, we trained the NeuS model [34] directly using images synthesized by our method, Zero123, and SyncDreamer, respectively. The quantitative results in Tab. 6 show that the consistent multi-view images synthesized by our method can significantly improve the 3D reconstruction quality. Additionally, our method exhibits similar performance to SyncDreamer which requires time-consuming re-training. The qualitative results in Fig. 7 show that it is challenging to train the NeuS model directly due to the lack of consistency in the images generated by Zero123. In contrast, our method generates more consistent multi-view images and, therefore, better reconstructs the geometry and texture details. We show improvements on other downstream applications such as image-to-3D in the Supplementary Material.

Table 6. Comparison of 3D reconstruction results. Our method significantly improves the reconstruction quality.

| | Chamfer Dist.↓ | Volume IoU↑ |
|---|---|---|
| Zero123 | 0.017 | 0.819 |
| SyncDreamer | **0.013** | **0.847** |
| Ours | 0.014 | 0.842 |

## 6. Conclusion

In this paper, we propose a method to improve the consistency of multi-view images synthesized by a pose-guided diffusion model without any training or fine-tuning. Specifically, for each pixel in the target view, we use epipolar attention to locate and retrieve features at corresponding locations in the input view and insert them into the generation process of the target view to enhance consistency. We also extend epipolar attention to the multi-view setting by synthesizing multiple views and retrieving information from the input and other target views. Experimental results show that our method can improve the consistency of the generated multi-view images and further benefit downstream applications such as 3D reconstruction.

**Acknowledgement.** Botao Ye is partially supported by the ETH AI Center.

## References


[1] 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 *ICCV*, pages 22560–22570, 2023. 3

[2] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. *arXiv*, 2023. 3

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3D content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2

[4] 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 *CVPR*, pages 13142–13153, 2023. 3, 7

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 3

[6] 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 *ICRA*, pages 2553–2560, 2022. 2, 6

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[8] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *CVPR*, pages 7779–7788, 2020. 3

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 1, 3

[11] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 3

[12] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3D content creation. In *CVPR*, pages 300–309, 2023. 2

[13] 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*, pages 9298–9309, 2023. 1, 2, 3, 6, 7

[14] 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. 2, 3, 6, 7

[15] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, pages 210–227, 2022. 2

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 2

[17] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 3

[18] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):1–15, 2022. 6

[19] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1, 2

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 1, 2

[24] 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. 2, 3, 6

[25] 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. 2, 3, 6

[26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015. 3

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1, 3

[28] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *ECCV*, pages 156–174, 2022. 3

[29] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023. 4, 5

[30] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, pages 16773–16783, 2023. 2, 3, 6, 7

[31] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3, 5

[33] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3D generation. In *CVPR*, pages 12619–12629, 2023. 2

[34] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, pages 27171–27183, 2021. 1, 2, 6, 8

[35] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. In *ECCV*, pages 573–591, 2022. 3

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[37] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[38] Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 1, 2, 6

[39] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 3

[40] Matthias Wödlinger, Jan Kotera, Manuel Keglevic, Jan Xu, and Robert Sablatnig. Ecsic: Epipolar cross attention for stereo image compression. *arXiv preprint arXiv:2307.10284*, 2023. 3

[41] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3D-aware image generation using 2D diffusion models. *arXiv preprint arXiv:2303.17905*, 2023. 1

[42] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3D view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023. 2, 3

[43] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023. 4, 5

[44] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3D scene generation with neural radiance fields. *arXiv preprint arXiv:2305.11588*, 2023. 1

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[46] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3D reconstruction. In *CVPR*, pages 12588–12597, 2023. 2, 3, 6