Novel View Synthesis with Pixel-Space Diffusion Models

Noam Elata*
Technion
noamelata@campus.technion.ac.il

Bahjat Kawar
Apple
b_kawar@apple.com

Yaron Ostrovsky-Berman
Apple
yostrovskyberman@apple.com

Miriam Farber
Apple
m_farber@apple.com

Ron Sokolovsky
Apple
ronus@apple.com

Abstract

Synthesizing a novel view from a single input image is a challenging task. Traditionally, this task was approached by estimating scene depth, warping, and inpainting, with machine learning models enabling parts of the pipeline. More recently, generative models are being increasingly employed in novel view synthesis (NVS), often encompassing the entire end-to-end system. In this work, we adapt a modern diffusion model architecture for end-to-end NVS in the pixel space, substantially outperforming previous state-of-the-art (SOTA) techniques. We explore different ways to encode geometric information into the network. Our experiments show that while these methods may enhance performance, their impact is minor compared to utilizing improved generative models. Moreover, we introduce a novel NVS training scheme that utilizes single-view datasets, capitalizing on their relative abundance compared to their multi-view counterparts. This leads to improved generalization capabilities to scenes with out-of-domain content.[1]

1. Introduction

In novel view synthesis (NVS), we aim to recreate a snapshot of a given scene from an unseen perspective. A successful algorithm must consider both the 3D geometry of the given scene and the underlying distribution of real-world images. NVS algorithms have been researched for many years, covering several tasks which differ in the expected number of input and output views and the relative distance between inputs and outputs, among other things. In this work, we focus on the simplest form of NVS: single-image to single-image, hoping that it will serve as a foundation for more general NVS settings. In previous works, this task has been decomposed into a pipeline of several computer vision components, namely depth estimation, warping, and inpainting [3, 21, 38, 50]. However, modern advances in generative modeling offer a simpler and more robust end-to-end approach [29, 37, 44, 54].

Generative diffusion models [12, 39, 40] have emerged as a top class of image generators, excelling at many conditional generation tasks [18, 32], including NVS [37, 44, 54]. In diffusion modeling, we train a neural network for removing Gaussian noise, and iteratively apply it in several steps to generate an image. These models are relatively simple to train, owing their stability to a simple denoising regression loss. Modern diffusion architectures notably rely on transformer layers [46] with self- and cross- attention blocks. These attention mechanisms are highly effective for the conditional generation task. For instance, in the widely used text-to-image generators [30, 32], cross-attention is used to condition the denoising network's features on the textual tokens. In our work, we effectively harness cross-attention in diffusion models for the NVS task.

Some existing diffusion-based NVS works [7, 37] operate in the latent space of an auto-encoder, following [30]. This can result in an unnecessary loss of high-frequency details due to the auto-encoder's reconstruction error, leading to texture mismatches between source and target views. Therefore, we opt to apply our method directly in the pixel space using a cascaded diffusion model design [13], avoiding this issue altogether. We demonstrate the improved texture transfer capabilities of our model compared to latent space NVS models.

Furthermore, many recent generative model-based NVS works involve conditional diffusion training with some form of intricate 3D geometry encoding [37, 44, 54]. However, in many cases, the benefit of such complex encoding methods is unclear, with some researchers claiming they may not be needed at all [29]. Inspired by previous work,

*Work done as part of an internship at Apple.
[1]Code available at https://github.com/apple/ml-vivid.

Figure 1. Novel view synthesis results from our diffusion model. Source views are taken from RealEstate10K [59], and fed into our base and SR models to produce a 256 × 256-pixel prediction. Our end-to-end system implicitly learns to preserve the features in the source view, transform their position along with the camera movement, and generate realistic details in unseen areas.

we explore geometry encoding methods and perform an ablation study. In our experiments, we show that that the impact of these methods is minimal, especially when harnessing a powerful diffusion model architecture. Building on our conclusions, we train an NVS diffusion model, reaching state-of-the-art (SOTA) NVS capabilities for the commonly accepted RealEstate10K dataset [59]. Our model, which we call VIVID (**V**iew **I**nference **V**ia **I**mage **D**iffusion), excels in both image quality and fidelity to the ground-truth novel view, measured by FID [10] and PSNR, respectively.

Finally, we experiment with the generalization capabilities of our method, and attempt to overcome the limited availability of multi-view data. Using the insight that camera rotations can be accurately simulated with a simple 2D homography transform, we propose a novel data augmentation scheme that enables the use of single-view datasets in NVS training. Our proposed scheme unlocks the possibility of training NVS models on far richer image content distributions, without introducing data scale mismatches that commonly occur in multi-dataset NVS training [33]. We show the effectiveness of this scheme in generalizing to unseen scenes with out-of-distribution content relative to the multi-view training dataset.

To summarize, we consider the task of novel view synthesis (NVS), characterize its different flavors, and focus on single-image to single-image NVS. We make use of a powerful diffusion model backbone [17], adapt it for NVS using the cross-attention layers, and apply it in the pixel space rather than a latent one. We ablate on different geometry encoding methods, and conclude that they offer minimal improvement over a simple scalar embedding of the camera poses. Our resulting model, called VIVID, achieves state-of-the-art NVS performance on the widely accepted RealEstate10K [59] benchmark. However, our method has some limitations that we discuss in section 5, which we hope to address in future work.

2. Related Work

While a wide variety of works are labeled "novel view synthesis", the exact tasks they are designed to solve can have subtle but important differences. We consider object-centric NVS techniques [2, 7, 8, 26, 53] to be a separate task, and instead focus on scene-level NVS, which poses different challenges such as more complex real-world scenes, multiple objects occlusions, and a high dynamic range of scene geometry. We categorize the different NVS tasks by the following 4 axes: (i) operating on static scenes [1, 29] vs. dynamic videos [6, 58]; (ii) having a single input view [9, 45] or multiple views [4, 51]; (iii) outputting a single novel view [37, 54] or multiple consistent views [24, 55]; (iv) the viewpoint difference between inputs and outputs can be short (most of the image content is shared) [1, 42, 50], medium (some of the image content is shared) [29, 37], or long (novel view contents are mostly unseen) [44, 54].

The distinction among these tasks is crucial for determining the choice of training data, scene representation, and model architecture. In this work, we focus on static scenes with a single input view and a single novel view in the mid- or long- range. Under this setting, NeRFs [23] and 3DGS [19] become less common as they struggle to extrapolate from a single viewpoint, even when combined with generative models [14, 15, 42, 43]. Instead, several works [3, 21, 38, 50] use monocular depth estimation (MDE), warp the pixels into the target view, and use an inpainting model to fill in missing details. This approach has several drawbacks [37], such as high sensitivity to depth estimation errors and loss of semantic image details. Gen-

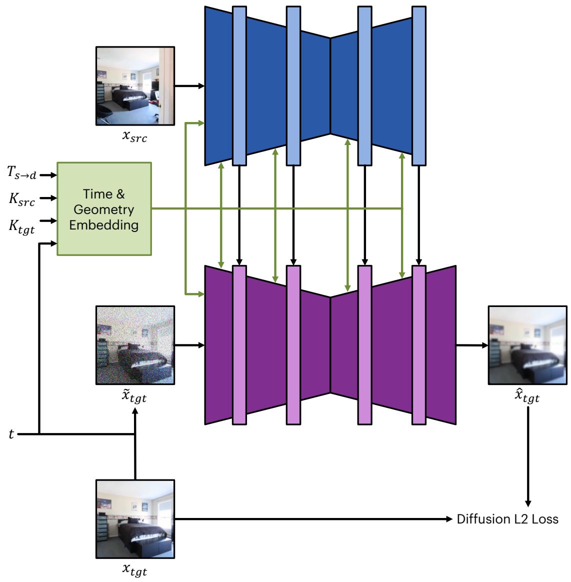

Figure 2. Overview of our model. The decoder (purple) learns to denoise the target view, using information from the source view provided by the encoder (blue) through cross-attention. Both models are aware of the diffusion timestep and scene geometry (green).

Figure 3. 256×256-pixel images from RealEstate10K [59], encoded and decoded using the autoencoder from Stable Diffusion v1.4 [30]. Some areas with severe loss of detail are highlighted.

erative models circumvent these issues by implicitly modeling the correlations among viewpoints. Thus, they are increasingly becoming a vital element of state-of-the-art approaches. GeoGPT [29] was among the first NVS methods to use generative modeling: they apply an autoregressive transformer to sample novel views conditioned on a single input view.

More recently, with the advent of powerful diffusion models [12, 17, 30], many techniques have incorporated them into NVS pipelines. In Pose-Guided Diffusion Models [44], the authors utilize a pre-trained MDE [28] to extract features from the input source view, which are then fed via cross-attention layers into a diffusion model that generates the target view. The cross-attention layers are constrained to pass information only along the epipolar lines, defined by the requested target viewpoint pose. Photometric-NVS [54] proposes a latent diffusion model with a two-stream architecture: two identical networks with shared weights process the source view and the noisy target views, exchanging information through cross-attention layers with pose information acting as query tokens. GenWarp [37] fine-tunes two copies of a pre-trained latent text-to-image diffusion model [30]: one for encoding the source view, and another for generating the target. They use a CLIP [27] image embedding of the source view instead of the text embedding, and further augment the network inputs with 2D coordinate embeddings, warped using MDE to match the source view coordinates.

3. NVS Diffusion Model

We formalize the NVS task as follows: for a given source image $\mathbf{x}_{src} \in \mathbb{R}^{C\times H\times W}$, a camera transformation (extrinsics matrix) $\boldsymbol{T}_{s\to t} \in \mathbb{R}^{3\times 4}$, and camera intrinsic matrices $\boldsymbol{K}_{src}, \boldsymbol{K}_{tgt} \in \mathbb{R}^{3\times 3}$, our algorithm should generate novel view samples $\mathbf{x}_{tgt} \sim p\left(\mathbf{x}_{tgt} \mid \mathbf{x}_{src}, \boldsymbol{T}_{s\to t}, \boldsymbol{K}_{src}, \boldsymbol{K}_{tgt}\right)$. Most multi-view datasets contain extrinsics matrices $\boldsymbol{T}_{src}, \boldsymbol{T}_{tgt}$ for each view. These matrices transform global coordinates to the view-specific camera coordinates. The transformation matrix between them can be obtained by $\boldsymbol{T}_{s\to t} = \boldsymbol{T}_{tgt}\boldsymbol{T}_{src}^{-1}$.[2]

3.1. Base Architecture

Diffusion models [12, 40] have emerged as a powerful tool to sample images from intricate conditional distributions such as text-to-image generation [30, 32] and image restoration [18, 60]. This makes them an ideal candidate for NVS, because they enable powerful sampling with a relatively simple training objective [12, 39, 48]. Diffusion model training requires a paired dataset of conditions $(\mathbf{x}_{src}, \boldsymbol{T}_{s\to t}, \boldsymbol{K}_{src}, \boldsymbol{K}_{tgt})$ and samples $(\mathbf{x}_{tgt})$, adding Gaussian noise to the samples, and training a network to remove the noise given the conditioning signals.

We propose using two parallel U-Net [31] architectures with attention layers [46]: an encoder to process $\mathbf{x}_{src}$, and a decoder to generate $\mathbf{x}_{tgt}$. The encoder acts as a feature extractor, while the decoder denoises the target image $\mathbf{x}_{tgt}$ as in common diffusion pipelines. To condition the denoising

[2]The inversion is done for a canonical 4×4 matrix representation.

Geometry Encoding	FID (↓)	PSNR (↑)
None	5.75 ± 0.04	13.39 ± 0.02
Epipolar	4.14 ± 0.06	17.43 ± 0.03
Pose	3.00 ± 0.02	21.11 ± 0.04
Pose + Epipolar	2.87 ± 0.04	21.15 ± 0.04
Pose + Depth	2.99 ± 0.01	21.29 ± 0.04
Pose + Coordinate	3.08 ± 0.04	21.06 ± 0.04

Table 1. Geometry encoding ablation. Metrics were computed 5 times for randomly sampled 10K source-target pairs.

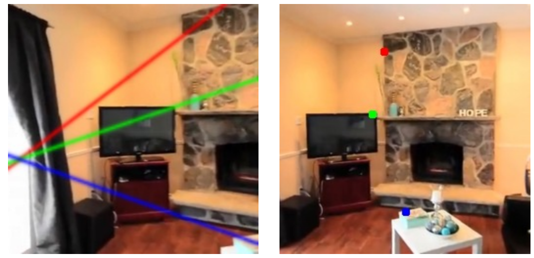

Figure 4. We select 3 random points in the target view, and show their epipolar lines in the source view in the corresponding color. In epipolar attention, we add a cross-attention bias relative to the proximity of the source view token to the target epipolar line.

process on the source image, we expand the self attention layers in the decoder to perform joint self and cross-view attention: query tokens from $\mathbf{x}_{tgt}$ attend to key and value tokens from both $\mathbf{x}_{tgt}$ itself and from $\mathbf{x}_{src}$ as extracted by the encoder. We train both U-Nets jointly, leading the encoder to best provide features for NVS in an end-to-end manner. Features are taken from the encoder at multiple layers and resolutions, enabling the transfer of global semantic information, as well as high resolution details. We choose not to share weights between the two U-Nets [54], nor build upon pre-trained models [37, 44] (*e.g.*, text-to-image models or depth estimators). This allows us to focus on end-to-end NVS without retaining artifacts from different objectives which may hinder NVS performance.

Additionally, we opt to perform the NVS diffusion process in the pixel space, and not in a latent space [37] of a pre-trained autoencoder model [30]. We believe this to be an essential prerequisite for accurate correspondence to source image details, since simply encoding an image and then decoding it leads to a significant loss of detail. We show an example of this in Figure 3. Similar observations about autoencoder latent spaces were also made in different contexts [49]. As an alternative to the speedup that latent space compression provides, we use a cascaded diffusion design [13]: we train a base model that receives $\mathbf{x}^{LR}_{src}$ at a low resolution and generates a novel view $\mathbf{x}^{LR}_{tgt}$, and a super-resolution (SR) model that receives $\mathbf{x}^{LR}_{tgt}$ (with noise conditioning augmentation [13]), and generates a higher-resolution version of it $\mathbf{x}^{HR}_{tgt}$. The SR model has the same architecture as the base model, differing only by having a smaller number of channels, and receiving the higher-resolution source $\mathbf{x}^{HR}_{src}$ as an additional input, allowing it to retain high-frequency details from the source view. We provide a depiction of our base model architecture in Figure 2, and of our SR model and end-to-end system in Appendix A. While cascaded diffusion can theoretically operate with a latent diffusion base model, we only use pixel-space diffusion and defer more complex options to future work.

We use the UNet architecture proposed in EDM2 [17], due to its impressive performance and training stability. In addition to the attention expansion and some hyperparameter choices (detailed in Appendix A), we perform two additional changes to the architecture: (i) we add a single attention layer at the second-highest resolution to enhance fine detail fidelity; and (ii) we encode the geometric conditioning information ($\boldsymbol{T}_{s\to t}, \boldsymbol{K}_{src}, \boldsymbol{K}_{tgt}$) into the network in several ways, as described in the next section.

3.2. Geometry Encoding Ablation

GeoGPT [29] demonstrated a reasonably good end-to-end generative NVS autoregressive model, using a simple encoding of the camera pose as model input. The authors make the claim that there is no need for further geometry-specific modules such as depth estimation or geometry-aware feature-matching layers. Despite these findings, many recent diffusion-based NVS methods [37, 44, 54] reach superior results by incorporating advanced forms of geometric bias, questioning whether this claim still holds. Here, we conduct thorough ablations on four types of 3D geometry encoding methods, and assess their impact. We train all model variants on RealEstate10K [59], and provide more details in Appendix B.

Pose Embedding. This is the simplest form of geometry encoding, often used in conjunction with other methods [29, 54]. We simply take the scalars of the camera pose matrices $\boldsymbol{T}_{s\to t}, \boldsymbol{K}_{src}, \boldsymbol{K}_{tgt}$, and encode them into a shallow perceptron that shifts the model's activations, similar to diffusion timestep embedding [12, 16, 17]. In our work, we also normalize these scalars by their dataset-wide mean and standard deviation before embedding them.

Epipolar Attention. In the cross-attention layers, target view tokens attend to their source view counterparts based on their visual features. With epipolar attention, we attempt to enrich the source-target correlation with camera pose information by modifying the cross-attention maps. For each position in the target view, we can use the camera pose information to find the relevant source view positions along its

Figure 5. Different NVS results for the same input sampled from our model. Details from the source view are kept in all samples, and diverse realistic options are generated in newly visible areas.

epipolar line, guiding the model towards features that constitute geometrically correct information. However, these correlations may amplify irrelevant information, as only a handful of points on the epipolar line will correspond to the query's target view position. Thus, we prefer the epiopolar correlations to act as a learnable bias for the attention matrix, enabling the model to select where the epipolar information is useful during training. Specifically, we use an implementation similar to [44], with two small but critical modifications: First, our epipolar correlations act as an additive bias to the attention matrix, instead of a multiplicative transformation. This does not zero out correlations outside the epipolar line, which may have critical semantic significance. Second, each attention head has its own independently learned mixing parameters. This enables the model to produce both geometrically dependent features, and semantically significant features, based on the parameters of each attention head. We illustrate this method in Figure 4.

Monocular Depth Estimation. MDE is used in many key NVS works [29, 37], either as an additional input, or for warping intermediate predictions and network features. However, we believe the information contained in a monocular depth prediction can be learned internally within the NVS model, given enough model capacity, training data, and iterations. We conduct experiments using a DepthAnythingV2 [52], a recent state-of-the-art MDE model, attaching its prediction on the source view as an additional input to our encoder.

Method	Mid-range FID ↓	Mid-range PSNR ↑	Long-range FID ↓	Long-range PSNR ↑
GeoGPT [29]	6.43	14.06	7.22	13.13
PhotoNVS [54]	7.12	13.32	9.22	12.05
GenWarp [37]	5.91	13.43	7.38	12.10
VIVID (Ours)	**2.89**	**17.36**	**3.89**	**15.21**
Source View	2.58	13.12	3.00	11.91

Table 2. Comparison to previous methods. Evaluation is done on 10K source-target pairs from RealEstate10K [59].

Coordinate Warping. MDE can also be used to warp coordinate embeddings, creating additional inputs for both the encoder and decoder, which can assist in feature-matching. This technique was proposed in [37], and we test it using the DepthAnythingV2 depth prediction [52].

Results. In Table 1, we compare the various geometry encoding methods on our base diffusion model, handling 64 × 64-pixel images, uniformly sampled from the RealEstate10K [59] test set. We start with a geometrically-uninformed baseline (no use of $\boldsymbol{T}_{s\to t}, \boldsymbol{K}_{src}, \boldsymbol{K}_{tgt}$), and then add pose embedding and epipolar attention to it. While both methods help, pose embedding produces far better results. Thus, we utilize pose embedding in conjunction with each of the other methods. While they mostly improve FID [10] and PSNR, the improvement is often small and not noticeable in qualitative evaluations. As a result, we prefer the pose embedding technique for its simplicity and effectiveness, and use this variant in the following sections.

3.3. Comparison to Previous Methods

In our ablations, we train the EDM2 [17] architecture with a limited batch size and number of iterations, and without using exponential moving average (EMA). In this section, we scale up our model (with the pose embedding technique): we double the batch size and quadruple the number of training iterations, use EDM2's Power function EMA [17], and apply classifier-free guidance (CFG) [11] for the base model using a separate unconditional denoiser. We provide more comprehensive details in Appendix A.

We compare our results in terms of image quality using FID [10] and distortion from the ground truth using PSNR. For a thorough evaluation, we generate 10K novel views from each tested method based on randomly sam-

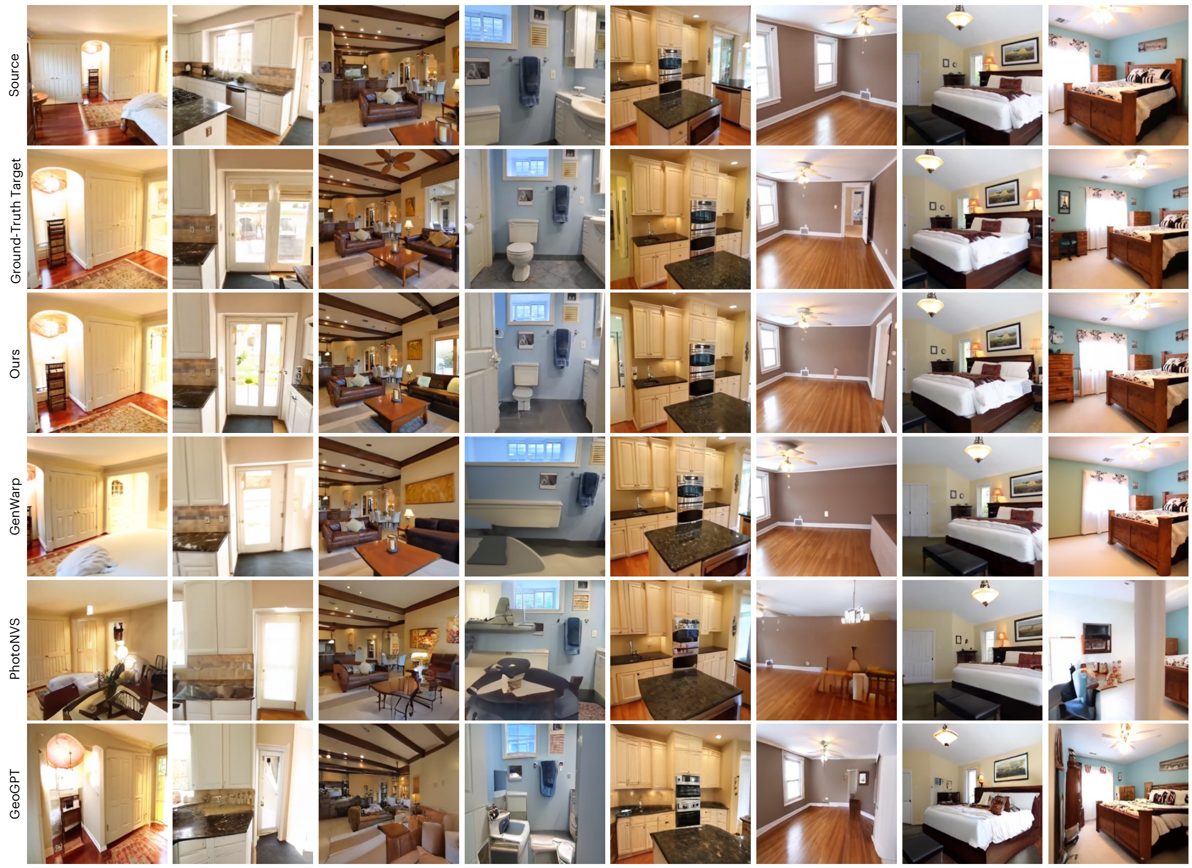

Figure 6. Comparison between our method and previous state-of-the-art approaches in NVS on RealEstate10K [59].

pled source images from the RealEstate10K [59] test set. To sample a ground-truth novel view and its corresponding camera transformation, two distinct ranges are identified, following GenWarp [37]: mid-range and long-range, corresponding to target views that are 30-60 and 60-120 frames away from the source, respectively. We use our base and SR models to generate 256×256 images. We compare FID and PSNR of images generated from our model with the results of state-of-the-art generative NVS methods: GenWarp [37], Photometric-NVS [54], and GeoGPT [29]. We additionally report the FID and PSNR of simply outputting the source view for any requested transformation as a naïve frame of reference. We expect the source view to achieve high perceptual quality, as the source and target images are sampled from the same distribution, yet have low PSNR. We present the results in Table 2, showing a substantial improvement of our method (more than 24%) over the previous state-of-the-art across all ranges and metrics. We show qualitative results in Figure 1 and Figure 6, and provide additional metrics (*e.g.*, joint FID [5]) and discussion in Appendix D.

Furthermore, since our approach uses a probabilistic diffusion model, it can sample from the distribution of novel views given a source view and a camera transformation. This distribution can have significant variance in its results, especially for large camera transformations. In Figure 5, we demonstrate our model's ability to estimate this distribution by drawing multiple samples for the same input. The wide variety of plausible completions may occasionally generate unrealistic features, as exhibited in general diffusion.

Our findings demonstrate that applying the diffusion process in the pixel space (rather than a latent space) is crucial for NVS. Using latent diffusion causes a loss of fine details in the source image that is not recoverable, as demonstrated in Figure 3, and validated in previous methods' results in Appendix D (as they use latent diffusion). Additionally, we hypothesize that the use of a dedicated encoder, implicitly trained to encode relevant NVS information, is a major advantage of our method over alternatives such as Photometric-NVS [54].

4. NVS Training on Single-View Data

We have shown impressive NVS results for a model trained on the RealEstate10K [59] training set, and tested on its corresponding test set, following the standards set in previous literature. However, towards the goal of achieving a

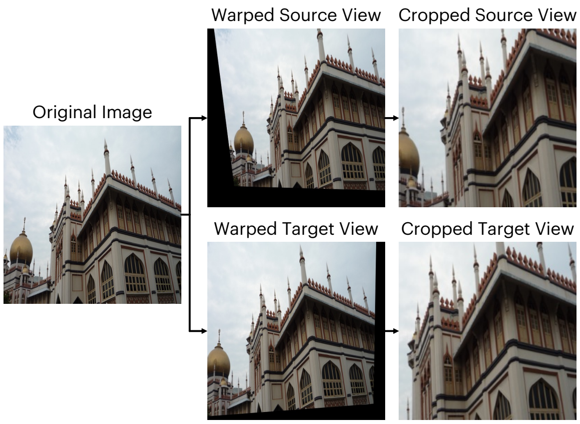

Figure 7. Example of our proposed homography augmentation. We randomly sample two camera rotation matrices, apply the transforms, and then perform the same crop to both views. In this example, the camera rotates to the right and slightly up from the source to the target view.

general-purpose NVS model, these training and evaluation strategies suffer from a few key shortcomings.

First, both training and evaluation data consist almost entirely of scenes inside houses. This is a biased and limited class of scenes, and does not cover the general real-world image distribution. Other large-scale multi-view datasets [21, 56] similarly suffer from limited diversity of both semantic contents and camera trajectories. This is understandable, since capturing multiple views of static scenes at a large scale is challenging, especially compared to the quantities achieved in single-view image datasets [20, 36].

Second, the multiple views in a given scene originate from frames inside a 640×360-pixel video touring a house. The tour videos predominantly capture empty houses with no movement, which is a good feature as different frames in the video can simulate multiple cameras capturing the same static scene. However, the data is low in resolution, and it also undergoes significant video compression that further degrades the quality of the model input and supervision.

Third, the camera positions in the different frames are not measured during video capture. Existing multi-view datasets offer camera extrinsics matrices which are synthetically generated through Structure-from-Motion (SfM) algorithms. While SfM algorithms often produce impressive results, they have their own margin of error and failure cases.

Lastly, even if we ignore their inaccuracies, SfM algorithms have another fundamental limitation: scale ambiguity – they cannot accurately recover the metric scale of the entire scene. Furthermore, because we focus on NVS with a single image input, the NVS model will also be limited, as it cannot accurately infer the scale of its input view in relation to the scale of the camera transformation matrix. RealEstate10K [59] and others [33, 37] have addressed this

Percentage	FID (↓)	PSNR (↑)
0%	36.14	14.38
10%	**31.98**	**14.52**
25%	35.27	14.45

Table 3. NVS performance on 20 out-of-domain scenes, with varying percentages of training data stemming from single-view images with our proposed augmentation.

problem by normalizing the scales of the translation vectors inside the camera matrices, either at the scene level or at the single image level. These methods stabilize the NVS model behavior and measurably improve its results, but they are still based on heuristic techniques and are prone to error, resulting in mismatches between the camera transformation $\boldsymbol{T}_{s\to t}$ and the ground-truth target view we compare against. These errors and mismatches have an increasing effect when considering data from multiple sources, cameras, dynamic ranges, lighting conditions, and SfM algorithms.

To mitigate some of these issues, we propose a novel NVS training scheme that makes use of both multi-view and single-view data. We apply it to our model, and demonstrate its benefits on out-of-domain test data.

4.1. Single-View Augmentation

Compared to multi-view data, high-resolution single-view images are abundantly available. Given a single image, we would like to simulate multiple views of a scene, beyond the trivial identity transform. To do so, one might consider taking smaller crops of high-resolution images. However, note that different overlapping crops do not represent any geometrically grounded camera transform. Camera translations can result in dis-occluded areas depending on the depth difference of different objects in the scene, with contents that are unseen in the original image. Therefore, we focus on simulating camera rotations.

To that end, we apply a homography transform on a high-resolution image, which warps the image's contents according to a randomly sampled camera rotation matrix. We apply the transform twice, for two randomly chosen rotation matrices, generating a source view and a target view. This produces warped images with undefined areas near the edges, as they should contain contents that were not visible in the original view. However, for a crop far enough from image boundaries, and small enough camera rotations, the resulting view is contained entirely in the original image. We then compare this to the corresponding crop in the target image, successfully simulating a camera rotation. We provide an illustrative example of this homography transform in Figure 7.

With this data augmentation strategy and a single-view image dataset, we produce a large amount of source-target

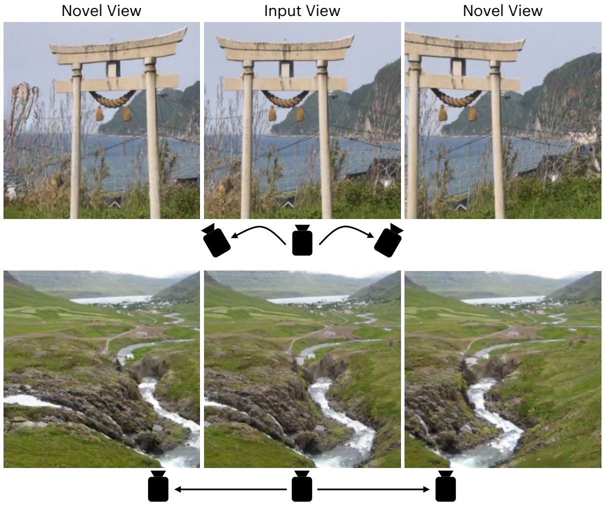

Figure 8. Our NVS model performs camera rotation (top) and translation (bottom) on OpenImages, a single-view dataset.

view pairs with diverse and accurately-labeled camera rotations. We then train our NVS model on a mixture of multi-view and single-view datasets, attempting to capture the benefits of both: real-world camera movement with ground-truth views from multi-view datasets, and rich diverse image contents with non-trivial camera transforms from single-view datasets. While the fact that homography transforms can simulate camera rotations has long been known, to the best of our knowledge, this is the first work to utilize it as part of a NVS model training scheme.

4.2. Experimental Results

We enrich our original model training scheme presented in section 3 with the use of single-view data with our proposed augmentation. We choose the OpenImages v5 dataset [20] as our source of single-view images, owing to its diversity and large size (more than 9 million images), and continue with RealEstate10K [59] as the source of multi-view images. To avoid our model collapsing to only learning rotations, we need to strike a balance between the two datasets. To that end, we retrain our base model with a percentage of the training data coming from single-view OpenImages scenes: 0% (original, RealEstate10K only), 10%, and 25%. We provide more training details in Appendix A.

Then, to evaluate the generalization capabilities of our NVS model, we test it with input images and camera transformations that are out-of-domain with respect to RealEstate10K. We perform qualitative evaluation with input source views from the OpenImages test set and camera transformations that include both rotation and translation, and show results in Figure 8 (using the 10% model).

Moreover, to obtain quantitative NVS performance metrics on out-of-domain data, we draw source-target view pairs from 20 scenes coming from 3 multi-view datasets (LLFF [22], MipNeRF-360 [1], and Ref-NeRF [47]), and evaluate FID and PSNR for each of our models. As evident Table 3, the 10% model exhibits a significant improvement over the 0% one in both FID and PSNR. This improvement shows that the model manages to benefit from the diverse single-view training data, while remaining true to the NVS framework due to our proposed augmentation. However, increasing the percentage to 25% leads to degraded performance, possibly due to an over-representation of rotation transforms in the data.

5. Conclusion

In this work, we explored and analyzed the many options proposed in the literature for designing an end-to-end novel view synthesis approach based on a generative model. We utilized a leading diffusion model architecture [17] as our backbone network, adapted it for the NVS task, and ablated on options for encoding the geometric information input. Our experiments show that while some sophisticated options can slightly enhance performance, the simple scalar embedding option works very well and the differences are mostly negligible. In fact, we show that our generative modeling choices such as utilizing a powerful architecture, passing information across views through attention, and operating in the pixel space, substantially upgrade the quality of NVS results compared to previous state-of-the-art methods.

Our approach achieves SOTA results in the widely accepted RealEstate10K [59] benchmark. Nevertheless, we highlight the shortcomings and potential pitfalls of this benchmark towards a comprehensive general-content NVS approach. We address some of these shortcomings with a novel training scheme enabling NVS model training on the abundantly available single-view image datasets, benefiting from their content diversity while maintaining the geometric capabilities learned from the multi-view datasets. We demonstrate that our proposed scheme can significantly enhance NVS performance on out-of-domain data.

In future work, we anticipate the continued improvement of NVS capabilities through the utilization of more modern and advanced generative models. We also hope that our method's limitations, which are also common in NVS literature, can be addressed. These limitations include generalization to camera trajectories that are uncommon or nonexistent in the training set, as our proposed augmentation does not introduce any camera translation. In addition, our method is inherently dependent on the scene scales provided in its training set, which could be problematic if they are obtained using Structure-from-Motion (SFM) techniques without some form of normalization [33, 59]. Finally, we hope that the foundations laid in this work will be extended to consistent multi-view generation, for example in the form of video NVS.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2, 8, 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. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 2

[3] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 1, 2

[4] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[5] Terrance DeVries, Adriana Romero, Luis Pineda, Graham W Taylor, and Michal Drozdzal. On the evaluation of conditional gans. *arXiv preprint arXiv:1907.08175*, 2019. 6, 3

[6] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2

[7] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 1, 2

[8] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 2

[9] Jiatao Gu, Qingzhe Gao, Shuangfei Zhai, Baoquan Chen, Lingjie Liu, and Josh Susskind. Control3diff: Learning controllable 3d diffusion models from single-view images. In *2024 International Conference on 3D Vision (3DV)*, pages 685–696. IEEE, 2024. 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2, 5

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5, 1

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 4

[13] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 1, 4, 2

[14] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[15] Wonbong Jang and Lourdes Agapito. Nvist: In the wild new view synthesis from a single image with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10181–10193, 2024. 2

[16] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 4

[17] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024. 2, 3, 4, 5, 8, 1

[18] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 1, 3

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 7, 8, 3

[21] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 1, 2, 7

[22] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 8, 3

[23] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[24] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10258–10268, 2024. 2

[25] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 3

[26] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 3500–3509, 2017. 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[28] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3

[29] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021. 1, 2, 3, 4, 5, 6

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 4

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3, 1

[32] 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. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 3

[33] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9420–9429, 2024. 2, 7, 8

[34] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[35] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 7

[37] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *arXiv preprint arXiv:2405.17251*, 2024. 1, 2, 3, 4, 5, 6, 7

[38] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 1, 2

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1, 3

[40] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1, 3

[41] George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[42] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024. 2

[43] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *Advances in Neural Information Processing Systems*, 36: 12349–12362, 2023. 2

[44] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 1, 2, 3, 4, 5

[45] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 1, 3

[47] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 8, 3

[48] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 3

[49] Yikai Wang, Chenjie Cao, and Yanwei Fu. Towards stable and faithful inpainting. *arXiv preprint arXiv:2312.04831*, 2023. 4

[50] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7467–7477, 2020. 1, 2

[51] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 2

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 5, 2

[53] 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. In *2024 International Conference on 3D Vision (3DV)*, pages 664–674. IEEE, 2024. 2

[54] Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7094–7104, 2023. 1, 2, 3, 4, 5, 6

[55] Jason J Yu, Tristan Aumentado-Armstrong, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Polyoculus: Simultaneous multi-view image-based novel view synthesis. *arXiv preprint arXiv:2402.17986*, 2024. 2

[56] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 7

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[58] Xiaoming Zhao, Alex Colburn, Fangchang Ma, Miguel Ángel Bautista, Joshua M. Susskind, and Alexander G. Schwing. Pseudo-Generalized Dynamic View Synthesis from a Video. In *ICLR*, 2024. 2

[59] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2, 3, 4, 5, 6, 7, 8

[60] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023. 3