NViST: In the Wild New View Synthesis from a Single Image with Transformers

Wonbong Jang   Lourdes Agapito
Department of Computer Science
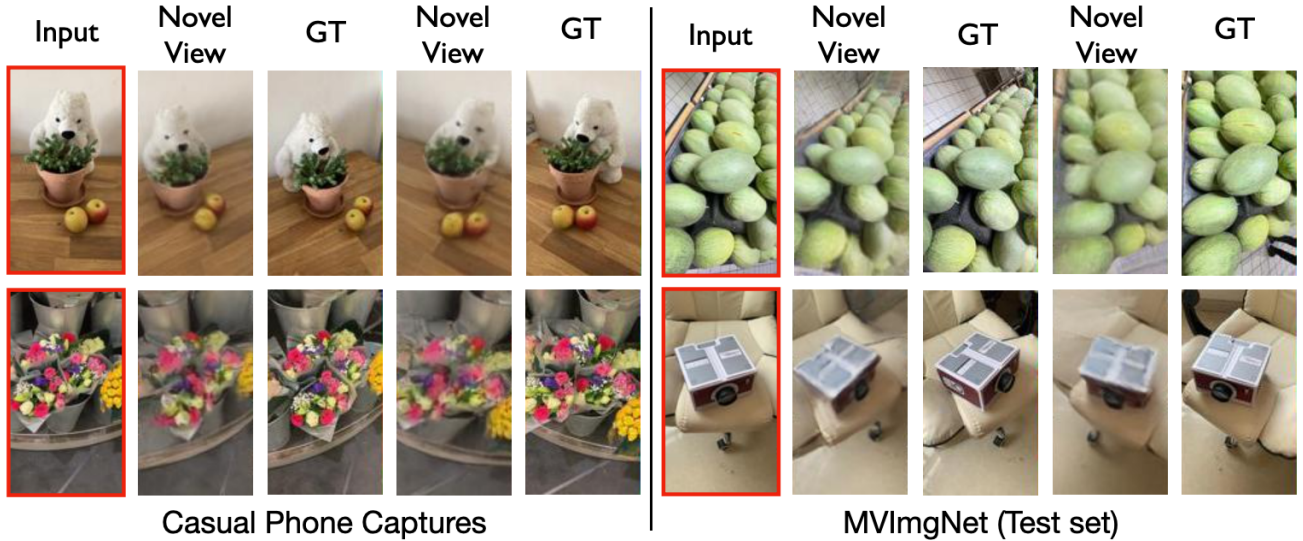University College London
{ucabwja,l.agapito}@ucl.ac.uk

Figure 1. We introduce NViST, a transformer-based architecture that enables synthesis from novel viewpoints given a single in the wild input image. We test our model not only on held-out scenes of MVImgNet, a large-scale dataset of casually captured videos of hundreds of object categories (Right) but also on out-of-distribution challenging phone-captured scenes (Left).

Abstract

We propose NViST, a transformer-based model for efficient and generalizable novel-view synthesis from a single image for real-world scenes. In contrast to many methods that are trained on synthetic data, object-centred scenarios, or in a category-specific manner, NViST is trained on MVImgNet, a large-scale dataset of casually-captured real-world videos of hundreds of object categories with diverse backgrounds. NViST transforms image inputs directly into a radiance field, conditioned on camera parameters via adaptive layer normalisation. In practice, NViST exploits fine-tuned masked autoencoder (MAE) features and translates them to 3D output tokens via cross-attention, while addressing occlusions with self-attention. To move away from object-centred datasets and enable full scene synthesis, NViST adopts a 6-DOF camera pose model and only requires relative pose, dropping the need for canonicalization of the training data, which removes a substantial barrier to it being used on casually captured datasets. We show results on unseen objects and categories from MVImgNet and even generalization to casual phone captures. We conduct qualitative and quantitative evaluations on MVImgNet and ShapeNet to show that our model represents a step forward towards enabling true in-the-wild generalizable novel-view synthesis from a single image. Project webpage: https://wbjang.github.io/nvist_webpage.

1. Introduction

Learning 3D scene representations from RGB images for novel view synthesis or 3D modelling remains a pivotal challenge for the computer vision and graphics communities. Traditional approaches, such as structure from motion (SfM) and multiview stereo pipelines, endeavor to op-

1

timize the 3D scene from RGB images directly by leveraging geometric and photometric consistency. The advent of Neural Radiance Fields (NeRF) [52] and its subsequent developments has marked a significant step forward by encoding the entire 3D scene within the weights of a neural network (or feature grid), from RGB images only. Although NeRF requires more than dozens of images for training, efforts have been undertaken to generalize NeRF across multiple scenes by taking a single image as an input [15, 31, 36, 46, 54, 63, 91, 99]. Nonetheless, generalizing NeRF-based models to multiple real-world scenes remains a challenge due to scale ambiguities, scene misalignments and diverse backgrounds. The huge success of 2D latent diffusion models [66] has sparked interest in 3D diffusion models. One trend is to make diffusion models 3D-aware by fine-tuning a pre-trained 2D diffusion model. However, so far these approaches are trained on centered objects, masked inputs, do not deal with backgrounds, and assume a simplified camera model (3-DOF) [48, 49, 61, 76]. Other approaches [13, 17, 28, 84] build the diffusion model on top of volume rendering in 3D space, but they are computationally expensive and slow to sample from.

Many recent breakthroughs in computer vision can be attributed to the emergence of very large datasets paired with the transformer architecture. For instance, MiDAS [62] exploits multiple diverse datasets, leading to robust performance in zero-shot depth estimation from a single RGB image while SAM [42] demonstrates that an extensive dataset coupled with a pretrained MAE and prompt engineering can significantly improve performance in segmentation. However, in 3D computer vision, scaling up real-world datasets has not been as straightforward. Synthetic datasets like ShapeNet [14] or Objaverse [21] have helped to promote progress, but there is a large domain gap.

The recent release of real-world large-scale multiview datasets such as Co3D [65] or MVImgNet [100], coupled with the availability of robust SfM tools such as COLMAP [71, 72] or ORB-SLAM [56] to enable camera pose estimation, has opened the door to large-scale training for new view synthesis. However, significant challenges remain to train a scene-level new view synthesis model on such real-world, large scale datasets due to the huge diversity of objects, categories, scene scales, backgrounds, and scene alignment issues. Motivated by this we propose NViST, a transformer-based architecture trained on a large-scale dataset to enable in-the-wild new view synthesis (NVS) from a single image. We exploit a subset of MVImgNet [100] which has one order of magnitude more categories and ×2 more objects than Co3D [65]. Our contributions can be summed up as follows:

- NViST can model general real-world scenes, including backgrounds, only requiring relative pose during training.
- Our novel decoder maps MAE features to 3D output tokens via cross-attention, conditions on camera parameters via adaptive layer normalisation and addresses occlusions with self-attention.
- Our qualitative and quantitative evaluations on MVImgNet test sequences show good performance on challenging real-world scenes.
- We demonstrate good generalization results on a zero-shot new-view synthesis task, on phone-captured scenes.

2. Related Work

Transformer, ViT and MAE: The transformer [86], a feed-forward, scalable, attention-based architecture, has brought a revolution to the field of natural language understanding. Inspired by it, the vision transformer (ViT) [24] uses image patches as tokens to achieve performance levels comparable to those of CNNs [29, 78] in many computer vision tasks. While the ViT is trained in a supervised way, masked autoencoders (MAE) [30] can be trained in a self-supervised way by randomly masking and in-painting patches, and be further fine-tuned on specific tasks.

Neural Implicit Representations: Neural implicit representations aim to learn a 3D representation without direct 3D supervision using neural networks. They have been employed for various tasks, including depth prediction [57] and scene representation through ray marching and coordinate-based MLPs [79]. *Mildenhall et al.* [52] proposed Neural Radiance Fields (NeRF), which integrates coordinate-based MLPs, positional encoding, and volume rendering to encode a scene in the weights of neural network. Upon optimisation, novel views can be rendered with impressively high quality. Beyond novel view synthesis, NeRF has found utility in a diverse range of computer vision tasks such as segmentation [10, 26, 87, 102], surface reconstruction [7, 53, 58, 89, 90], and camera registration [18, 20, 37, 45, 47, 50, 92, 96].

Grid-based representations: A limitation of the original NeRF method is its lengthy training time. As an alternative to coordinate-based MLPs, grid-based approaches [16, 25, 55, 81, 97, 98] have been proposed to expedite training. TensoRF [16] proposed the vector-matrix (VM) representation as an efficient and compact way to represent the 3D grid. In the context of 3D-aware Generative Adversarial Networks (GANs), EG3D [12] introduced triplanes by projecting 3D features into three different planes. Several approaches use the triplane representation to learn to generate 3d implicit representations from ImageNet [64, 70, 74, 80].

Learning multiple scenes using NeRF: Generalising NeRFs to multiple scenes remains a challenging problem. Several methods associate 2D features with the target views [15, 19, 31, 35, 65, 85, 91, 99], while others condition the network on latent vectors with a shared MLP across the dataset [3, 27, 32, 36, 54, 63]. Adding an adversarial loss to NeRF leads to 3D-aware GANs, which allow consis-
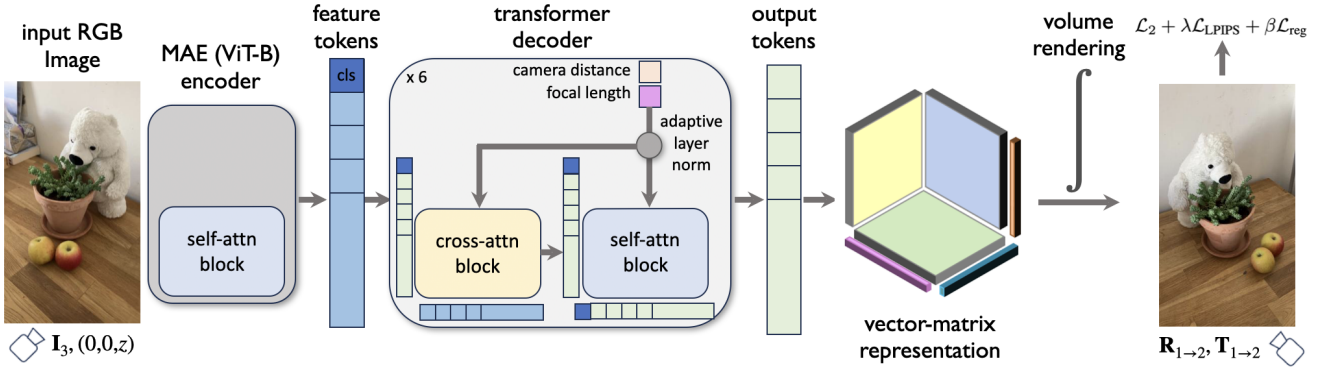
Figure 2. **Architecture.** NViST is a feed-forward transformer-based model that takes a single in-the-wild image as input, and renders a novel view. The encoder, a finetuned Masked Autoencoder (MAE), generates feature tokens, which are translated to output tokens via cross-attention by our novel decoder, conditioned on normalised focal length and camera distance via adaptive layer normalisation. Self-attention blocks allow reasoning about occlusions. Output tokens are reshaped into a vector-matrix representation that is used for volume rendering. NViST is trained end-to-end via a balance of losses: photometric L_2, perceptual L_{LPIPS}, and a distortion-based regulariser L_{reg}.

tent rendering from different viewpoints [8, 11, 12, 44, 73]. Most approaches that use a single input image require aligned datasets like ShapeNet [14, 79] or FFHQ [40], and fail to deal with scale ambiguities or diverse backgrounds.

Diffusion for NVS: Latent diffusion [66] and its open-source release Stable Diffusion have transformed the field of image generation. However, applying diffusion models to learn 3D implicit representations is not straightforward as there is no access to 3D ground-truth. 3DiM [94] proposes a pose-conditioned image-to-image approach. Dreamfusion [60] introduced score distillation sampling (SDS) to train NeRF with a 2D diffusion model. SDS has been used in many follow-up works [51, 77, 83, 88, 93]. Fine-tuning Stable Diffusion on a large-scale synthetic dataset [21] allows diffusion models to be 3D-aware such as Zero-1-to-3 and its follow-ups [48, 49, 61, 76]. Other approaches denoise directly in 3D and supervise the model in 2D space after rendering, [1, 2, 13, 28, 39, 82, 84], use the 2D diffusion model as a prior [22], or optimise NeRF jointly and regard it as the ground truth [17]. Similar to latent diffusion [5, 41, 74] adopt 2-stage training.

3D Representation with Transformers: Geometry-free methods employing transformer architectures have been explored as seen in [43, 67–69]. Others build NeRF representations using transformers [38, 46, 70, 75, 95]. The concurrent work LRM [33] extracts image features through DINO [9], and refines learnable positional embeddings via attention mechanisms. Unlike NViST, LRM focuses on object-centric scenes without background and employs triplanes instead of vector-matrix representation.

3. Methodology

NViST is a feed-forward conditional encoder-decoder model built upon transformers to predict a radiance field from a single image to enable synthesis from novel view-

points. As Figure 2 shows, our architecture is structured into three key components: two transformer-based modules (encoder and decoder) and a NeRF renderer, which is composed of a shallow multi-layer perceptron (MLP) and a differentiable volume rendering module.

3.1. Encoder

Input images are first split into a set of fixed size non-overlapping patches before being fed to the encoder, which predicts feature tokens using a ViT-B transformer architecture. We formulate the encoder $\mathcal{E}$ as $F, C = \mathcal{E}(I)$ where I are input images and F and C are the feature and class tokens respectively. In practice, we use the encoder of a pre-trained MAE [30], a self-supervised vision learner trained to predict masked image patches, which we further finetune on the training dataset to adapt to the different image resolution from ImageNet, on which MAE [30] is trained. As illustrated in Figure 3, we find that the self-supervised features learnt by the MAE [30] encapsulate a powerful intermediate representation of the scene's geometric and appearance properties that can be leveraged by the decoder to predict the radiance field. Note that the weights of the encoder are continuously updated during end-to-end training, since we found that this enhances the encoder's ability to generate smoother and more segment-focused features, as shown in Figure 3, and better performance as shown in Table 2.

3.2. Decoder

The goal of the decoder $\mathcal{D}$ is to take the class and feature tokens, C and F, predicted by the encoder along with the camera parameters used as conditioning (normalized focal length f and camera distance z) and predict the radiance field, which we encode using a vector-matrix (VM) representation, parameterized with three matrices M and three vectors V. A key feature of our decoder is the use of cross-
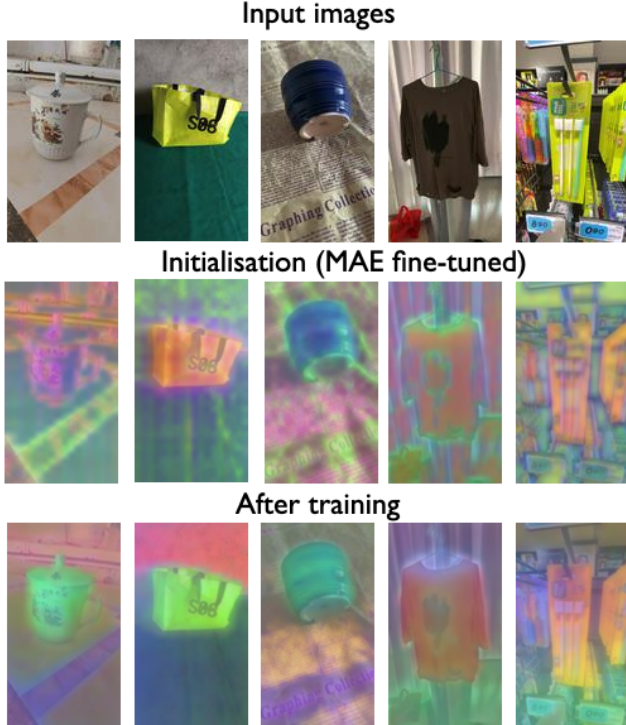
Figure 3. **Encoder Output Visualisation:** (Top) Input images. (Middle) Features from a fine-tuned MAE, which serve as initialisation to our encoder. (Bottom) Features after end-to-end training. Features shown after reducing to 3 dimensions with PCA. Optimised features appear smoother and more segment-focused, supporting the fact that updating encoder weights significantly improves the performance (see also ablation in Table 2).

attention and self-attention mechanisms. Unlike previous models [99] which project spatially aligned ResNet [29] features onto target rays, our decoder learns this 2D to 3D correspondence by learning to associate feature tokens with the relevant output tokens through cross-attention.

The decoder $\mathcal{D}$ has output tokens O, which are learnable parameters, initialised following a random normal distribution. For simplicity, from here on we will use the term output tokens O to refer to the concatenation of output tokens and class token C, to which we further add positional embeddings [86].

$$M, V = \mathcal{D}(F, C, O, f, z). \tag{1}$$

The decoder applies cross-attention between feature tokens F and output tokens O, and self-attention amongst output tokens O. The attention mechanism allows the network to reason about occlusions, where the class token C acts as a global latent vector. For each attention block, we apply adaptive layer normalisation [34, 59], instead of standard layer normalisation, to condition on camera parameters.

3.2.1 Relative Camera Pose

A key strength of our approach is that it does not require all objects/scenes in the dataset to be brought into alignment by expressing their pose with respect to a canonical reference frame. Instead, we assume the rotation of the camera associated with the input image I_i to be the identity and we express the rotation of any other image I_j as a relative rotation $R_{i\rightarrow j}$. We assume that the camera location of the input image is $T_{i\rightarrow i} = (0, 0, z)$, where z is the normalized distance of the input camera from the origin of the world coordinate frame, located at the centroid of the 3D bounding box containing the point cloud, and $T_{i\rightarrow j} = R_i^T(T_j - T_i)$. Zero-1-to-3 [49] also adopts a relative-pose based approach, however it assumes objects are located in the centre and uses a 3-DoF camera pose model (radius, elevation and azimuth), while NViST uses a full 6-DoF model for the camera pose.

3.2.2 Conditioning on Camera Parameters

There are several ways to apply conditioning such as concatenating camera parameters to output tokens O or feature tokens F. However, in NViST we apply adaptive layer normalization, as the camera parameters influence the overall scale of the scene. Conditioning camera parameters improves the model performance as seen in Table 2.

Positional encoding: We apply the positional encoding from NeRF [52] for f and z, concatenating up to 4-th sine and cosine embeddings with the original values z and f as $M = \oplus_{k=1}^{4}(\sin 2^k(f), \cos 2^k(f), \sin 2^k(z), \cos 2^k(z))$.

Adaptive Layer Normalisation: We employ an additional MLP $\mathcal{A}$ to regress the shift δ, scale α and gate scale γ from the conditioning inputs (z, f, M) for each attention block.

$$\alpha, \delta, \gamma = \mathcal{A}(z, f, M) \tag{2}$$

An alternative to adaptive layer normalisation would be to concatenate M to output tokens O, but as seen in our ablation (Table 2) this strategy does not lead to better results.

3.2.3 Attention Blocks

NViST uses self-attention blocks between output tokens O and cross-attention blocks between output tokens O and feature tokens F. The embedding dimension for both O and F is e. While standard layer normalization is applied to feature tokens F, we apply adaptive layer normalization to output tokens O using the shift δ and scale α values regressed by $\mathcal{A}$, the MLP described in Equation 2.

$$\begin{aligned} O_n &= \delta + \alpha \times \text{Norm}(O) \\ F_n &= \text{Layer Norm}(F). \end{aligned} \tag{3}$$
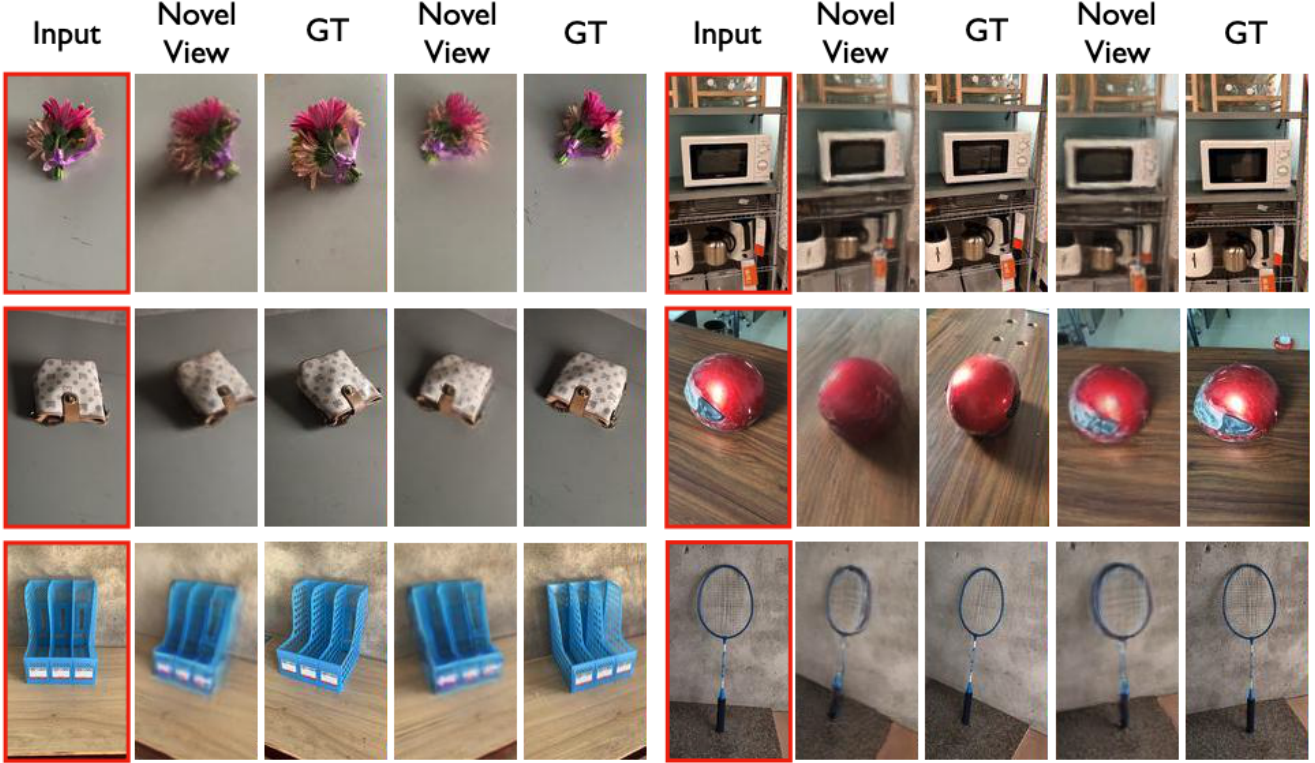
4

Figure 4. **Qualitative Results on Test (Unseen) Scenes:** We show the capabilities of NViST to synthesize novel views of unknown scenes. The model correctly synthesizes images from different viewpoints of various categories with diverse backgrounds and scales.

Cross-attention can then be expressed as

$$\text{Attn} = \text{Softmax}(\frac{O_n F_n^T}{\sqrt{e}}). \tag{4}$$

Finally, output tokens are updated via residual connection $O \leftarrow O + \gamma \times \text{Attn} \cdot F_n$, where γ is the gate scale also regressed by $\mathcal{A}$. Note that self-attention is obtained in an equivalent way, just between output tokens.

Reshaping: We use MLPs to reshape the output tokens into the vector-matrix representation that encodes the radiance field, adapting their respective dimensionalities. This is followed by unpatchifying into the 3 matrices and 3 vectors that form the VM representation. Please refer to the supplementary material for more details.

3.3. Rendering

The VM representation of the radiance field predicted by the decoder is used to query 3D point features which are then decoded by a multi-layer perception into color **c** and density σ and finally rendered via volumetric rendering.

Vector-Matrix Representation: For a compact yet expressive representation of the radiance field, we adopt the vector-matrix decomposition proposed by TensoRF [16] which expresses each voxel as the sum of three vectors and matrices, one pair per axis. Specifically, a 3D feature grid

($\mathcal{T}$), is decomposed into three vectors $(V_{r_1}^X, V_{r_2}^Y, V_{r_3}^Z)$ and three matrices $(M_{r_1}^{Y,Z}, M_{r_2}^{Z,X}, M_{r_3}^{X,Y})$, each pair sharing the same channel dimensions (k) such that

$$\mathcal{T} = \sum_{r_1=1}^{k} V_{r_1}^X \circ M_{r_1}^{Y,Z} + \sum_{r_2=1}^{k} V_{r_2}^Y \circ M_{r_2}^{Z,X} + \sum_{r_3=1}^{k} V_{r_3}^Z \circ M_{r_3}^{X,Y}. \tag{5}$$

While the value of density σ is obtained by applying ReLU activation directly to the feature value $\mathcal{T}_\mathbf{x}$ at point **x**, the colour **c** is predicted with a shallow MLP, conditioned on the viewing direction. The VM representation outperforms using a triplane as shown in our ablation study (Table 2).

Volume Rendering: For each sampled ray r, we obtain its final color $\hat{C}(r) \in \mathbb{R}^3$ using volumetric rendering, following the methodology of NeRF [52]. The transmittance T_i is first computed at each point **x** along the ray as $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$, where δ_i is the distance between adjacent points $\delta_i = t_{i+1} - t_i$. The pixel color is calculated by integrating the predicted color at each i-th point $\mathbf{c}_i$, weighted by light absorbance T_i - T_{i+1}.

$$\hat{C}(r) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i \tag{6}$$

5

3.4. Training Losses

We employ a combination of losses to train our architecture in an end-to-end manner including the L2 photometric rendering loss, a Learned Perceptual Image Patch Similarity (LPIPS) loss [101], and the distortion-based regulariser proposed by [4]. Given ground-truth pixel colors $\mathbf{v}$, estimates $\hat{\mathbf{v}}$ and accumulated transmittance values $\mathbf{w}$ along the points $\mathbf{x}$ on the ray, our overall loss function is defined as

$$\mathcal{L} = \mathcal{L}_2(\hat{\mathbf{v}}, \mathbf{v}) + \lambda \mathcal{L}_{\text{LPIPS}}(\hat{\mathbf{v}}, \mathbf{v}) + \beta \mathcal{L}_{\text{dist}}(\mathbf{w}, \mathbf{x}). \quad (7)$$

For MVImgNet $\lambda = 0.1$ and $\beta = 0.01$. We found LPIPS to be extremely effective on real world datasets (see Table 2).

4. Experimental Evaluation

4.1. MVImgNet

Train/Test Split: MVImgNet contains videos of over 6.5 million real-world scenes across 238 categories. Our training set, contains a subset of 1.14M frames across 38K scenes of 177 categories. For the test set, we hold out every 100th scene from each category to a total of $13,228$ frames, from 447 scenes and 177 categories.

Pre-processing: MVImgNet uses COLMAP to estimate camera matrices and generate 3D point clouds. Since each scene has its own scale from SfM, we rescale point clouds to a unit cube, such that the maximum distance along one axis equals 1. Then, we center the point clouds in the world coordinate system. We downsample the original images by $\times 12$ while preserving their aspect ratio. Camera intrinsics are recalibrated accordingly.

Implementation Details: NViST has approximately 216M parameters: 85M for the encoder, 131M for the decoder, and 7K for the renderer. The encoder takes images of size 160×90, using a patch size of 5, so the total number of feature tokens is 576. The resolution of the VM representation is 48. The patch size for the decoder is 3, the total number of output tokens 816, and 16 heads. The embedding dimension is 768 for both encoder and decoder, and we sample 48 points along the ray. We train NViST with $2\times$ A100-40GB GPUs for approximately one week, using a batch size of 22 images and rendering 330K pixels, up to one million iterations. The initial learning rates are 6e-5 for encoder, and 4e-4 for decoder and renderer and we decay them following the half-cycle cosine schedule.

4.1.1 Qualitative Results

Results on Test (Unseen) Scenes: As depicted in Figure 4, our model demonstrates its capability to synthesise new views of unknown scenes. Figure 4 highlights the model's ability to handle objects from diverse categories, such as flowers, bags, helmets and others. We show that NViST can
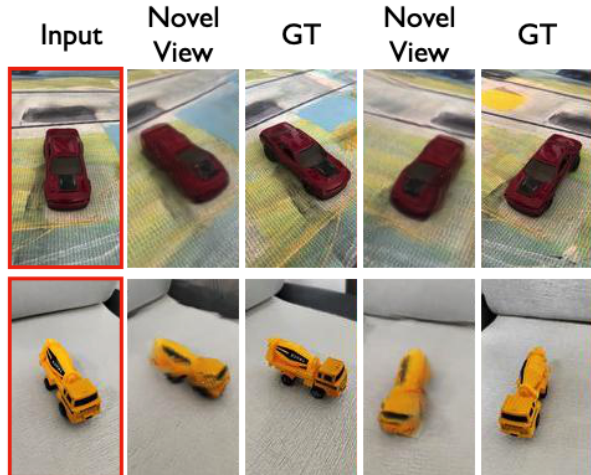

Figure 5. **Results on Unseen Category:** This figure shows how the model generalises to a novel category unseen at training. We validate our model with a held-out category (toy-cars).

Figure 6. **Casual Phone Captures:** We demonstrate NViST in OOD scenarios. First row: the model can capture different categories. Second row: outdoor setting. Third row: many objects.

deal with a variety of backgrounds, such as tables, wooden floors, textureless walls or more cluttered environments.

Results on Unseen Category: We take a held-out category (toy-cars), and test the ability of our model to generalize to categories unseen at training. Figure 5 shows that NViST can synthesize new views of objects from a category not seen at training time given a single input image.

6

	PSNR↑	SSIM↑	LPIPS↓
PixelNeRF [99]	17.02	0.41	0.54
VisionNeRF [46]	19.82	0.51	0.47
Ours	**20.83**	**0.57**	**0.29**

Table 1. **Quantitative Tests on MVImgNet:** We compare NViST with PixelNeRF and VisionNeRF on new view synthesis from single in-the-wild images. NViST outperforms both on all metrics.

Casual Phone Captures: Our motivation for this paper was to train a feed-forward model that we could easily use on casually captured images, for instance acquired with our own mobile devices. We test the ability of NViST to deal with out-of-distribution scenes/images by performing zero-shot new-view synthesis on scenes captured by us on a mobile-phone device. We pre-process the images in the same manner as MVImgNet, using COLMAP to estimate the focal length and camera distance parameters to be used as conditioning. Figure 6 shows results on out-of-distribution scenes. The top row highlights the model's ability to process a scene with multiple objects from diverse categories. The second row reveals its competence on outdoor scenes, despite their limited presence in the training set. The third row illustrates NViST's ability to learn scenes with a large number of objects. Figure 1 shows two further examples of results on phone captures.

Depth Estimation: Figure 7 shows qualitative results of the depth predicted by NViST. Since there is no ground-truth depth for these images, we qualitatively compare NViST with the recent MiDASv3.1 with Swin2-L384 [6]. As MiDAS predicts depth from RGB images, we provide the GT novel view image as input. This shows that depth estimation with NViST performs well even though it is not trained with depth supervision, while MiDAS uses direct depth supervision from multiple datasets.

4.1.2 Comparisons with baseline models

We compare NViST with PixelNeRF [99] and Vision-NeRF [46], all trained on MVImgNet using their official code releases and applying the same pre-processing as NViST. Figure 8 and Table 1 show that NViST outperforms both models by a large margin. Both models rely on aligned features, but have limitations dealing with occlusion. We qualitatively compare with the pre-trained Zero-1-to-3 [49] on a phone capture in Figure 9, using the approximate camera pose as it assumes a centered object. Since Zero-1-to-3 and its follow-up works assume a 3DoF camera and do not model the background, we could not conduct quantitative comparisons.

We could not conduct quantitative comparisons with generative models such as GenVS [13] or with the large-scale model LRM [33] as their models are not publicly
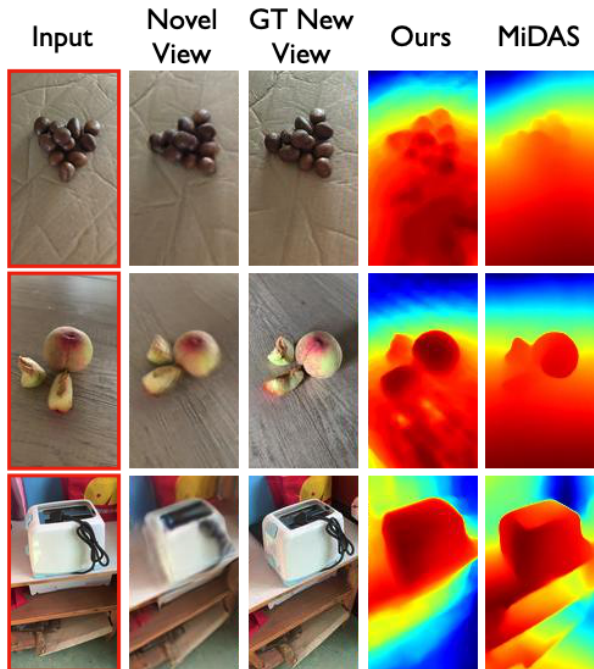

Figure 7. **Depth Estimation:** Examples of the depth estimates on test images. Although NViST focuses on novel view synthesis and is trained with RGB losses only, depth estimation is consistent and finds good object boundaries. We show a comparison with the state-of-the-art disparity estimator MiDAS v3.1 with Swin2-L384 [6]. We provide MiDAS the GT new view images as input, as it cannot do novel view synthesis. NViST performs well even though it is not trained with depth supervision, unlike MiDAS.
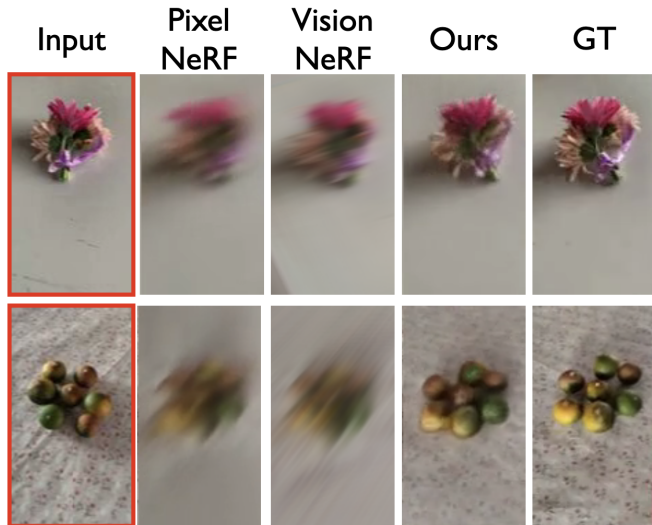

Figure 8. **Qualitative Comparison with PixelNeRF and Vision-NeRF:** NViST displays better performance than PixelNeRF [99] and VisionNeRF [46], especially when the target view is far away from the input view.

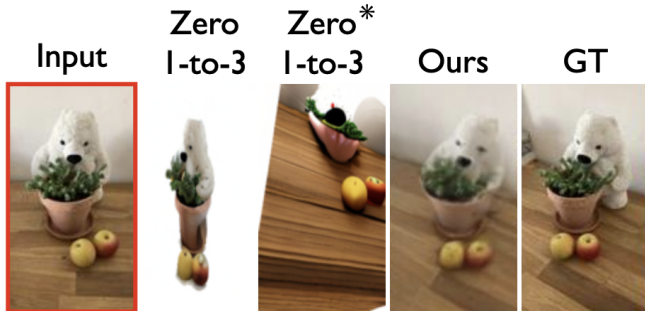available. Moreover, GenVS is category-specific and LRM

Figure 9. **Qualitative comparison with Zero-1-to-3 [49] on a phone captured input image:** Zero-1-to-3 shows result w/masked input, Zero1to3* shows result w/full image input. None result in good novel-view synthesis, except NViST.

	PSNR↑	SSIM↑	LPIPS↓
Ours	**20.83**	**0.57**	**0.29**
w/o LPIPS	20.72	0.54	0.46
w/o Camera Conditioning	20.24	0.49	0.36
Concat Camera Parameters[(1)]	20.81	**0.57**	0.30
w/ Self-Attention Decoder[(2)]	20.74	**0.57**	0.31
w/o VM Representation[(3)]	19.60	0.49	0.44
w/o Updating Encoder	18.54	0.47	0.49

Table 2. **Ablation:** For (1) we concatenate high dimensional camera feature tokens to output tokens instead of adaptive layer normalisation. For (2) we update all tokens via self-attn (no cross-attn). For (3), we use triplane instead of VM as the representation.

does not model backgrounds.

4.1.3   Ablation Study

We conducted an ablation study to analyse the effect of design choices on the performance of NViST as summarised in Table 2. While the perceptual LPIPS loss [101] is not commonly employed in NeRF-based methods, it appears to play a crucial role in improving quality. Conditioning on normalized focal length and camera distance helps the model deal with scale ambiguities, and adaptive layer normalisation performs better than concatenating camera parameters to output tokens. Instead of employing cross-attention in Figure 2, we can concatenate feature tokens and output tokens and update both of them using self-attention. However, it increases memory consumption and does not lead to a better result in Table 2. While projecting features onto triplanes has been extensively used before [12, 33, 44, 70], our experiments show that the use of a vector-matrix (VM) representation [16] improves performance. Note that for the triplane representation, we use a 1-layer MLP to regress occupancy, while occupancy is directly calculated using Equation 5 in our VM representa-

	Cars		Chairs	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
PixelNeRF [99]	23.17	0.146	23.72	0.128
VisionNeRF [46]	22.88	0.084	24.48	0.077
VD [82]	23.29	0.094	-	-
SSDNeRF [17]	23.52	**0.078**	24.35	**0.067**
Ours	**23.91**	0.122	**24.50**	0.090

Table 3. **Quantitative Evaluation on ShapeNet-SRN:** NViST performs similarly to other baselines on ShapeNet Cars/Chairs, although our method only requires relative pose. For qualitative comparisons, see supplementary materials.

tion. Updating encoder weights also improves the performance as the encoder output is used for cross-attention.

4.2. ShapeNet-SRN

ShapeNet-SRN [79] has two categories (cars and chairs) and is a widely used benchmark to compare models that perform novel view synthesis from a single input image. Since all objects are aligned and there is no scale ambiguity, pre-processing is not needed and we do not use the LPIPS loss as it is a synthetic dataset. Table 3 shows that NViST performs similarly to baseline models on ShapeNet-SRN dataset. Although other models apply absolute pose, we only employ relative pose, which means we do not fully exploit the alignment of objects in ShapeNet-SRN.

5. Conclusion

We have introduced NViST, a transformer-based scalable model for new view synthesis from a single in-the-wild image. Our evaluations demonstrate robust performance on the MVImgNet test set, novel category synthesis and phone captures of out-of-distribution scenes. Our design choices were validated via ablations and a quantitative comparison was conducted on MVImgNet and ShapeNet-SRN. Interesting future directions include extending NViST to adopt a probabilistic approach and to multiview inputs.

Limitations: Some loss of sharpness could be due to our computational constraints, which led us to downsample images by ×12 and train on a fraction of the original dataset. We pushed for a transformer-based architecture, without GAN losses or SDS [60], which eased and sped up training, but may have also contributed to some loss of detail.

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, in-painting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12608–12618, 2023. 3

[2] Titas Anciukevičius, Fabian Manhardt, Federico Tombari, and Paul Henderson. Denoising diffusion via image-based rendering. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[3] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[4] 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. 6

[5] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Josh Susskind. Gaudi: A neural architect for immersive 3d scene generation. *arXiv*, 2022. 3

[6] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 7

[7] Hongrui Cai, Wanquan Feng, Xuetao Feng, Yan Wang, and Juyong Zhang. Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[8] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3981–3990, 2022. 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[10] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Chen Yang, Wei Shen, Lingxi Xie, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. In *NeurIPS*, 2023. 2

[11] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020. 3

[12] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2, 3, 8

[13] 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. *arXiv preprint arXiv:2304.02602*, 2023. 2, 3, 7

[14] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. In *arXiv preprint arXiv:1512.03012*, 2015. 2, 3

[15] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2

[16] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022. 2, 5, 8, 14

[17] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. *arXiv preprint arXiv:2304.06714*, 2023. 2, 3, 8

[18] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. *arXiv preprint arXiv:2211.11505*, 2022. 2

[19] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. *arXiv preprint arXiv:2304.12294*, 2023. 2

[20] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 264–280. Springer, 2022. 2

[21] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3

[22] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023. 3

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 14

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 14

[25] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[26] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224*, 2022. 2

[27] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 2

[28] 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, 3

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3, 14

[31] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4709, 2021. 2

[32] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2

[33] 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. 3, 7, 8

[34] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[35] Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Vitor Guizilini, Thomas Kollar, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. 2023. 2

[36] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 2

[37] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021. 2

[38] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3

[39] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18423–18433, 2023. 3

[40] Tero Karras, S. Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 3

[41] Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[42] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[43] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 198–216. Springer, 2022. 3

[44] Eric-Tuan Le, Edward Bartrum, and Iasonas Kokkinos. Stylemorph: Disentangled 3d-aware image synthesis with a 3d morphable styleGAN. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 8

[45] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 2

[46] Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 2, 3, 7, 8, 14

[47] Yunzhi Lin, Thomas Müller, Jonathan Tremblay, Bowen Wen, Stephen Tyree, Alex Evans, Patricio A Vela, and Stan Birchfield. Parallel inversion of neural radiance fields for robust pose estimation. *arXiv preprint arXiv:2210.10108*, 2022. 2

[48] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 2, 3

[49] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tok-makov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2, 3, 4, 7, 8

[50] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Jo-hannes Kopf, and Jia-Bin Huang. Robust dynamic radi-ance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[51] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 3

[52] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-thesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceed-ings, Part I 16*, pages 405–421. Springer, 2020. 2, 4, 5

[53] Mirgahney Mohamed and Lourdes Agapito. Dynamicsurf: Dynamic neural rgb-d surface reconstruction with an opti-mizable feature grid, 2023. 2

[54] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2022. 2

[55] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[56] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[57] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervi-sion. In *Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[58] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2

[59] William Peebles and Saining Xie. Scalable diffusion mod-els with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 4

[60] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Milden-hall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3, 8

[61] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Sko-rokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123:

One image to high-quality 3d object generation us-ing both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2, 3

[62] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocu-lar depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 2

[63] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition*, pages 1558–1567, 2022. 2

[64] Pradyumna Reddy, Ismail Elezi, and Jiankang Deng. G3dr: Generative 3d reconstruction in imagenet, 2024. 2

[65] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Com-mon objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2

[66] 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. 2, 3

[67] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view syn-thesis through set-latent scene representations. In *Proceed-ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022. 3

[68] Mehdi S. M. Sajjadi, Daniel Duckworth, Aravindh Mahen-dran, Sjoerd van Steenkiste, Filip Pavetić, Mario Lučić, Leonidas J. Guibas, Klaus Greff, and Thomas Kipf. Ob-ject Scene Representation Transformer. *NeurIPS*, 2022.

[69] Mehdi S. M. Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, and Klaus Greff. RUST: Latent Neural Scene Representations from Unposed Imagery. *CVPR*, 2023. 3

[70] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and De-qing Sun. Vq3d: Learning a 3d-aware generative model on imagenet. *arXiv preprint arXiv:2302.06833*, 2023. 2, 3, 8

[71] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Com-puter Vision and Pattern Recognition (CVPR)*, 2016. 2

[72] Johannes Lutz Schönberger, Enliang Zheng, Marc Polle-feys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[73] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware im-age synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3

[74] Katja Schwarz, Seung Wook Kim, Jun Gao, Sanja Fidler, Andreas Geiger, and Karsten Kreis. Wildfusion: Learning

3d-aware latent diffusion models in view space. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3

[75] Bokui Shen, Xinchen Yan, Charles R Qi, Mahyar Najibi, Boyang Deng, Leonidas Guibas, Yin Zhou, and Dragomir Anguelov. Gina-3d: Learning to generate implicit neural assets in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4913–4926, 2023. 3

[76] 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, 2023. 2, 3

[77] 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. 3

[78] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[79] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 8, 14, 16

[80] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *International Conference on Learning Representations*, 2023. 2

[81] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 2

[82] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3D generative models from 2D data. In *ICCV*, 2023. 3, 8

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

[84] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Frédo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *arXiv*, 2023. 2, 3

[85] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2

[86] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[87] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021. 2

[88] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Morpheus: Neural dynamic 360 {\deg} surface reconstruction from monocular rgb-d video. *arXiv preprint arXiv:2312.00778*, 2023. 3

[89] Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In *2022 International Conference on 3D Vision (3DV)*, pages 433–442. IEEE, 2022. 2

[90] 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. *arXiv preprint arXiv:2106.10689*, 2021. 2

[91] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2

[92] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[93] 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. 3

[94] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[95] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3D reconstruction. *arXiv:2301.08247*, 2023. 3

[96] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 2

[97] Brent Yi, Weijia Zeng, Sam Buchanan, and Yi Ma. Canonical factors for hybrid neural fields. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[98] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021. 2

[99] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2, 4, 7, 8

[100] 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. 2, 14, 15

[101] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 8

[102] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 2

NViST: In the Wild New View Synthesis from a Single Image with Transformers

Supplementary Material

A. Implementation Details

Finetuning MAE Encoder: We use the pre-trained MAE [30] with ViT-B [24] from the original MAE implementation. Those weights are trained for ImageNet [23] which has a resolution of 224×224 pixels with a patch size 16. This means that the model divides the image into 196 feature tokens. Our image resolution for MVImgNet [100] is 160×90, and we use an encoder patch size of 5, resulting in 576 patches in the encoder. During fine-tuning, we initialise the weights of attention blocks with the pre-trained MAE, as the Transformer architecture allows for arbitrary attention matrix shapes as long as the embedding dimension remains the same. We fine-tune by randomly masking out and inpainting patches with L2 reconstruction loss, similar to the approach used in MAE [30]. The process converges within a single epoch.

Initialisation of Decoder: We initialise the decoder of NViST with the fine-tuned MAE weights. With the exception of the learnable parameters of positional embedding of output tokens and the last MLP layers, we initialise the weights of attention blocks with the fine-tuned MAE weights.

Number of output tokens: For MVImgNet [100], the resolution of vector-matrix(VM) representation is 48, and the channel dimension of each matrix and vector is 32. The patch size of the decoder is 3. Each 48×48 matrix M consists of non-overlapping 16×16 patches, and the 48 dimensional vector V is divided into 16 patches. Therefore, the total number of output tokens for VM representation is 818.

Decoder MLPs and Reshaping: The embedding dimension of the decoder is 768. We have 818 output tokens, and the channel dimension of VM representation [16] is 32, with a patch size of 3 for the decoder. For the output tokens corresponding to the matrices M in the VM representation, we deploy MLP to reduce the embedding dimension to 288. For those corresponding to vectors V, we reduce it to 96. Subsequently, we reshape them into VM representation.

B. Qualitative Results on ShapeNet-SRN

We perform a qualitative comparison with VisionNeRF [46] on ShapeNet-SRN [79] dataset as depicted in Figure 12. VisionNeRF, recognised as one of top-performing models on ShapeNet-SRN, employs ViT [24] as its encoder. Notably, VisionNeRF does not utilise any generative approaches, and was trained using 8 A100 GPUs. Similarly for MVImgNet, we fine-tune a MAE for the ShapeNet-SRN dataset and initialise the parameters of both encoder and de-

Figure 10. **Failure Cases** This figure illustrates when the model fails to do new view synthesis properly. The toilet scene shows that the model learns geometry in a distorted way. In the motorcycle scene, the model fails to estimate the occluded area and the proper scale.

coder of NViST with this fine-tuned MAE for ShapeNet-SRN. The ShapeNet-SRN images are of resolution 128×128, and we use an encoding patch size of 8, resulting in 256 feature tokens. The resolution of VM representation is 64, and the decoder patch size is 4, so we use 818 output tokens, each with an embedding dimension of the Transformer as 768. We still maintain the relative pose but do not condition on camera parameters as the dataset is aligned and does not have scale ambiguities. We train the model with a single 3090 GPU with $500,000$ and $700,000$ iterations, respectively for car and chair.
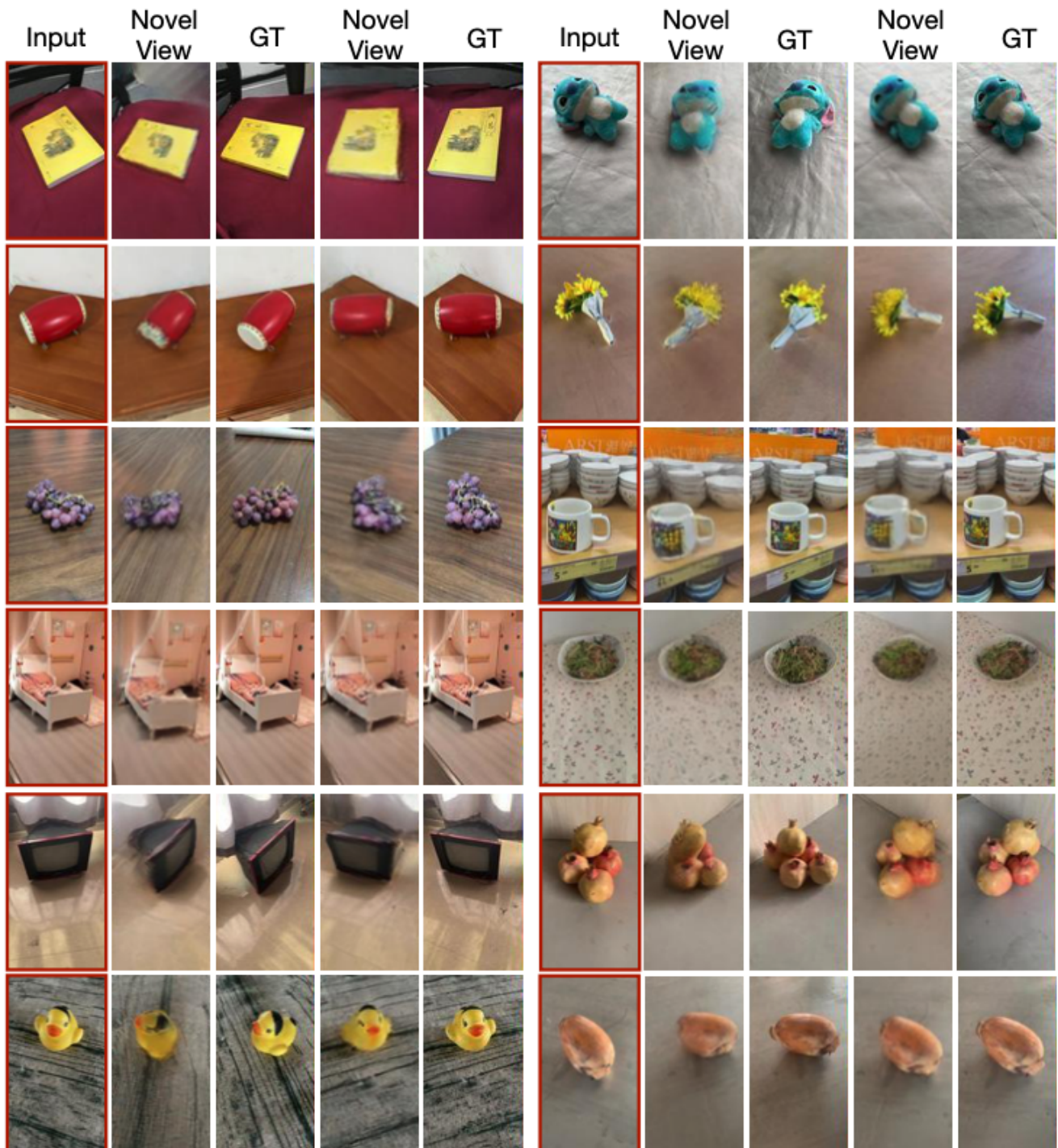
Figure 11. **Qualitative Results on Test (Unseen) Scenes of MVImgNet [100]:** NViST can synthesize high-quality novel view on challenging scenes from single in-the-wild input images.

Figure 12. **Qualitative Comparison on ShapeNet-SRN [79]:** NViST performs similar to VisionNeRF which is one of the top-performing models on ShapeNet-SRN dataset. Note that we do not employ LPIPS and do not condition on camera parameters for ShapeNet-SRN as it is a synthetic dataset, but we still use the relative pose even though objects are aligned in 3D.