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 [4] with ViT-B [3] from the original MAE implementation. Those weights are trained for ImageNet [2] 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 [7] 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 [4]. 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 [7], 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 [1] 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 [5] on ShapeNet-SRN [6] dataset as depicted in Figure 3. VisionNeRF, recognised as one of top-performing models on ShapeNet-SRN, employs ViT [3] 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 decoder 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 1. **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.

References

[1] 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. 1

[2] 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. 1

[3] 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. 1

[4] 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. 1

[5] 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. 1

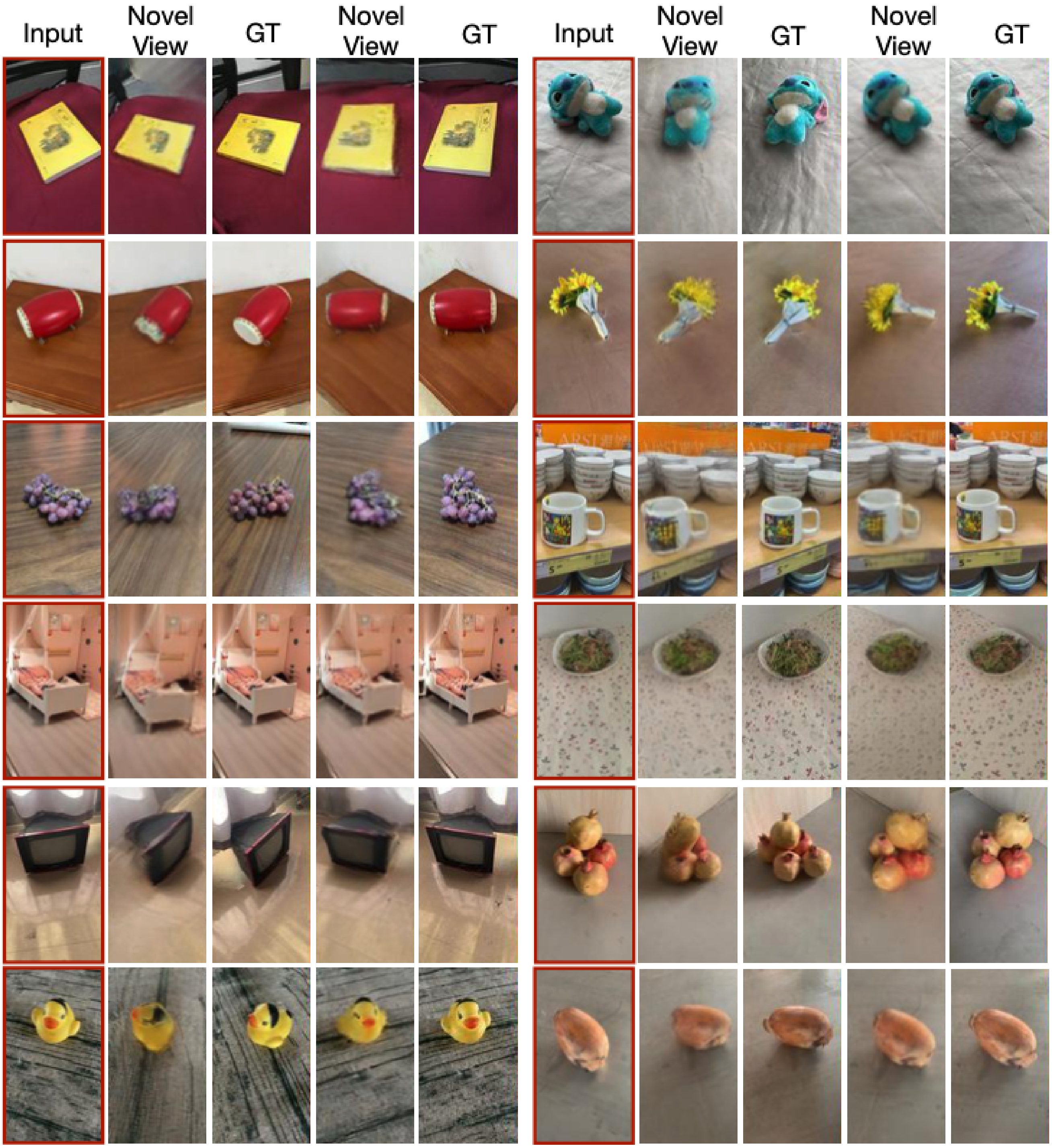

Figure 2. **Qualitative Results on Test (Unseen) Scenes of MVImgNet [7]:** NViST can synthesize high-quality novel view on challenging scenes from single in-the-wild input images.

[6] 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. 1, 3

[7] 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. 1, 2

Figure 3. **Qualitative Comparison on ShapeNet-SRN [6]:** 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.