3D Snapshot: Invertible Embedding of 3D Neural Representations in a Single Image *–Supplementary Materials–*

Yuqin Lu, Bailin Deng *Member, IEEE*, Zhixuan Zhong, Tianle Zhang, Yuhui Quan, Hongmin Cai, Shengfeng He, *Senior Member, IEEE*

Here, we first provide more implementation details of our framework in Sec. 1, including some specific network designs and hyperparameters. In Sec. 2, we demonstrate the capability of our model to synthesize arbitrary views by showing more views restored from the corresponding 3D snapshot, with comparisons to two video-based invertible methods, Video Snapshot [1] and IICNet [2], which can only restore a limited number of views due to their constrained concealing capabilities. Moreover, we show results on the Tanks and Temples dataset [3] to further validate our model's ability on embedding and faithfully restoring more complicated scenes. In Sec. 3, we further discuss a potential real-world application of our model, where we develop a mobile application demo to restore views from the printed 3D snapshot. We also provide a narrated overview of our method in video form.

1 Implementation Details

The invertible architecture of our proposed method consists of 16 invertible blocks. For each block, we leverage the coupling layer architecture used in [4]–[6] to build up the invertible mapping. The transformation functions $\phi(\cdot)$, $\eta(\cdot)$, and $\rho(\cdot)$ used in each invertible block can be arbitrary functions to add more non-linearity to the network. In our experiments, we deploy a 5-layer densely connected convolutional network to represent $\phi(\cdot)$, $\eta(\cdot)$, and $\rho(\cdot)$. Additionally, $\rho(\cdot)$ is followed by a centered sigmoid function and a scale term to prevent numerical explosion due to the $\exp(\cdot)$ computation.

The work is supported by the Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097) and the National Research Foundation Singapore under the AI Singapore Programme (No. AISG3-GV-2023-011). (Shengfeng He is the corresponding author.)

Yuqin Lu, Zhixuan Zhong, Tianle Zhang, Yuhui Quan, and Hongmin Cai are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China; Yuqin Lu, Zhixuan Zhong, and Tianle Zhang are also with the School of Computing and Information Systems, Singapore Management University, Singapore. E-mail: spaceluyq@gmail.com, zxzhong20@gmail.com, terryjoy0111@gmail.com, csyhquan@scut.edu.cn, hmcai@scut.edu.cn.

Bailin Deng is with the School of Computer Science and Informatics, Cardiff University, UK. E-mail: DengB3@cardiff.ac.uk.

Shengfeng He is with the School of Computing and Information Systems, Singapore Management University, Singapore. E-mail: shengfenghe@smu.edu.sg.

Moreover, we initialize the target constant noise tensor $\mathbf{P}$ to be all zeros and dynamically update it after 1000 iterations. This helps stabilize training at the beginning. We use the Adam optimizer [7] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to train our model. We set the initial learning rate to be 2×10^{-4} and halve the learning rate for every 5K iterations.

2 More Results

2.1 Arbitrary View Synthesis

In Fig. 1, we provide a more detailed view of the embedding images and showcase additional qualitative results from various viewpoints to demonstrate the ability of our method to synthesize views from arbitrary perspectives. The embedding images reveal that information is preserved as imperceptible patterns within the images, having little impact on the clarity of the hosts. As we are the first to achieve neural rendering from a single image, we compare our method to video-based invertible approaches in this paper. Video-based methods aim to embed a short video into a single image and restore it back. However, the restoration quality significantly decreases as the number of frames increases, limiting its capacity to recover arbitrary views. Fig. 2 presents a comparison of the embedding images and the limited restored views for these two video-based methods. In contrast, our model can generate high-quality views from arbitrary perspectives, and the qualitative results from Fig. 1 further demonstrate that the quality consistency is maintained across different views.

2.2 Comparisons with NeRF Models

In this section, we present additional comparisons with NeRF models on novel view synthesis, including NeRF [8], DVGO [9], TensoRF [10] and Plenoxels [11]. Fig. 3 shows the qualitative results of our method and other NeRF models for synthesizing novel views. We observe that our method can still maintain high-quality rendering compared to other NeRF models, even though our model size is significantly smaller.

(a) Mic

(b) Hotdog

(c) Lego

Fig. 1: Qualitative results of arbitrary view synthesis on NeRF-Synthetic. The leftmost column displays the 3D snapshot, while the right side shows the synthesized views from arbitrary perspectives.

(a) Chair

(b) Drums

(c) Ship

Fig. 2: Qualitative comparisons on the embedding images and restored views. From top to bottom are results from Video Snapshot [1], IICNet [2], and our method, respectively.

2.3 Choice of Host Image

Given a trained NeRF model and a host image (usually an image that depicts the 3D scene from a specific viewpoint), our method aims to embed the NeRF model into the host image to create an embedding image called 3D snapshot. 3D snapshot can serve as a perceivable carrier for 3D scene and can be stored and transmitted in the form of image, which later allows inverting back to the 3D model for novel view synthesis. In this section, we investigate the impact of the host image on the reconstruction quality. We conduct experiments to validate our method's robustness to different choice of host image. Specifically, for the same scene that we use for embedding, we choose three images from different viewpoint of the scene to serve as the host image and embed the scene into each of them using the same network. The underlying scene models are then respectively inverted back to synthesize novel views. The synthesis results and the embedding image are presented in Fig. 4. We observe that the reconstruction quality is not significantly affected by the variance of the host image, and the PSNR for each variant does not deviate too much, indicating that our method is robust to the choice of host image.

2.4 Results on Tanks and Temples

In this section, we present additional results on the Tanks and Temples dataset [3], which comprises larger scenes with more intricate geometry and lighting effects. Qualitative results are depicted in Fig. 5, including comparisons to the original model, as well as two video-based invertible methods, Video Snapshot [1] and IICNet [2]. Detailed illustrations of the embedding images and additional reconstructed views from the restored neural planes are shown in Fig. 6, which further demonstrate our method's capability to handle more complex scenes.

3 Application and Video Demonstration

The idea of representing a NeRF model in a 3D snapshot brings about a potential application in the metaverse, where the 3D snapshot can be printed and users can have access to the underlying 3D model by capturing the printed 3D snapshot with a mobile device. To enable this capability on a mobile device, we reduce the resolution to adapt to the limited computing resources for responsiveness and fluency. Fig. 7 shows a demo of our mobile application. Here the user scans an image on the wall using a mobile app, and the 3D model embedded in the image is decoded by the app for interactive viewing from different angles. The full demo with a screen recording is shown in our supplementary video to demonstrate this potential. Note that, for efficiency, the image size in our demonstration is reduced to 380×380, and only 20 images are rendered (versus 200 in normal cases).

We have also provided a narrated overview of our method in video form, where our method is illustrated in an animated way for better understanding, and the reconstructed results are shown in the form of a video. For more details, we refer the reader to the attached video, where the narrated overview is followed by a screen recording of our mobile application demo.

References

[1] Q. Zhu, C. Han, G. Han, T.-T. Wong, and S. He, "Video snapshot: Single image motion expansion via invertible motion embedding," *IEEE TPAMI*, vol. 43, no. 12, pp. 4491–4504, 2020.

[2] K. L. Cheng, Y. Xie, and Q. Chen, "Iicnet: A generic framework for reversible image conversion," in *ICCV*, 2021, pp. 1991–2000.

[3] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[4] L. Ardizzone, J. Kruse, C. Rother, and U. Köthe, "Analyzing inverse problems with invertible neural networks," in *ICLR*, 2018.

[5] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[6] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

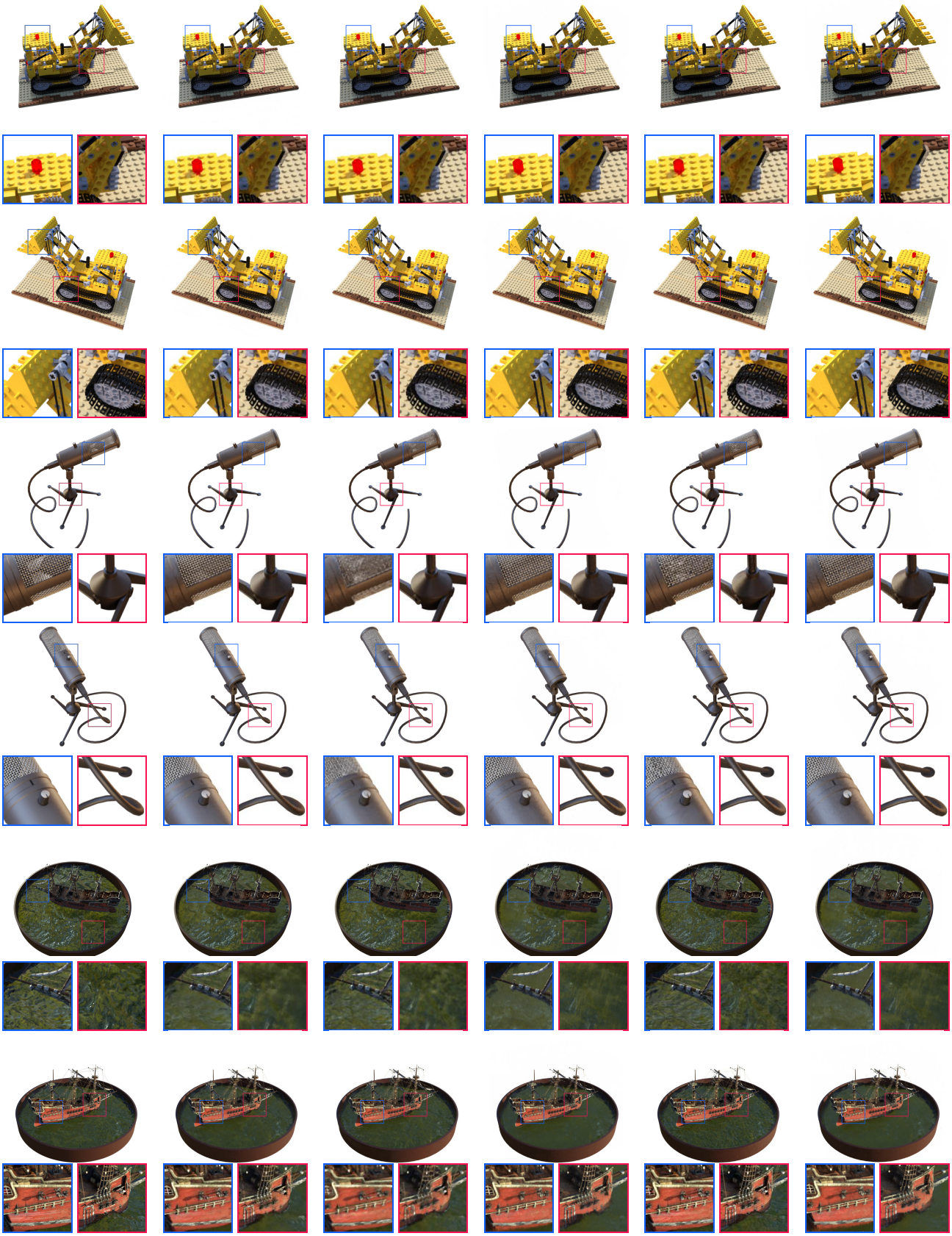
[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[9] C. Sun, M. Sun, and H. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *CVPR*, 2022.

[10] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *ECCV*, 2022.

[11] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *CVPR*, 2022, pp. 5501–5510.



(a) GT (b) NeRF (c) DVGO (d) TensoRF (e) Plenoxels (f) Ours

Fig. 3: Comparisons with other NeRF models on novel view synthesis. Our method can achieve high-quality rendering while maintaining a significantly smaller model size.

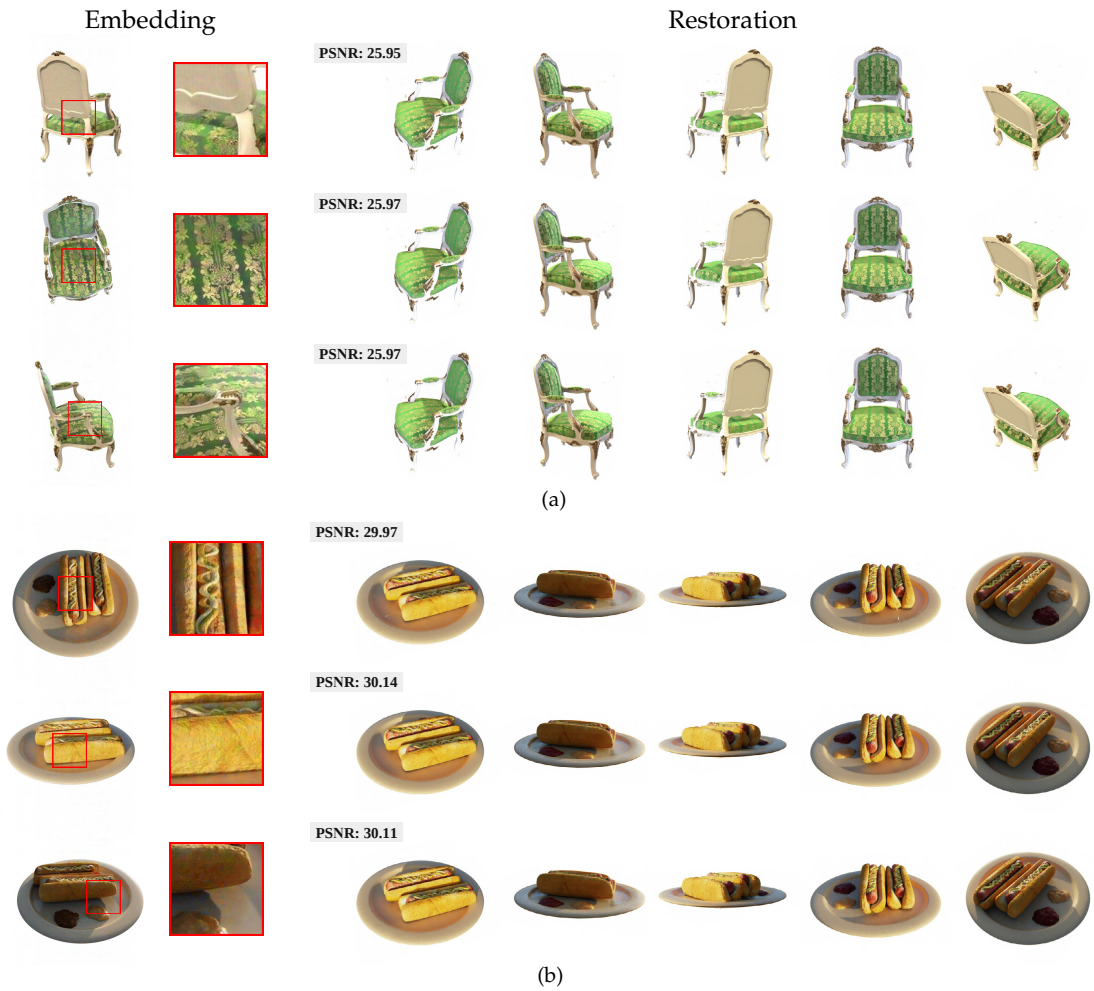
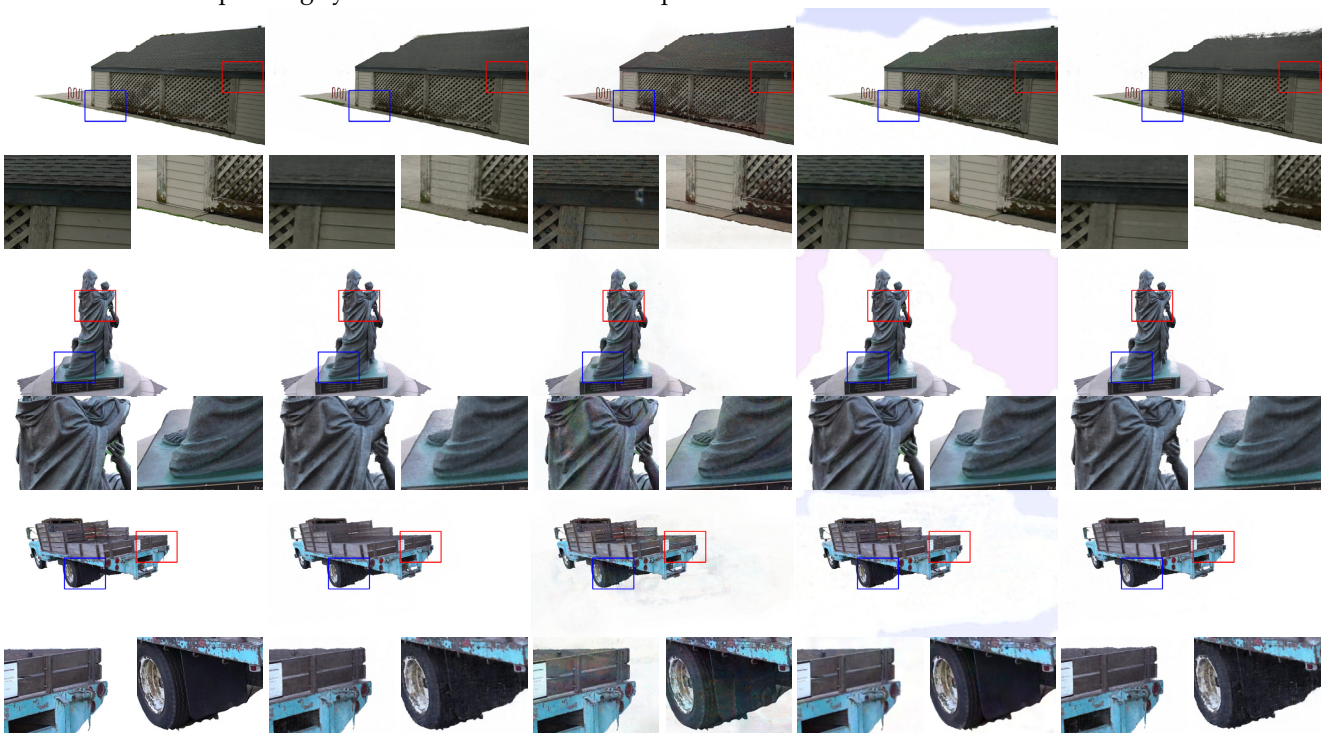


Fig. 4: Synthesis results for different choice of host image. The embedding images are shown in the left side and the right side shows the corresponding synthesized views. PSNR is reported for each variant in each row.

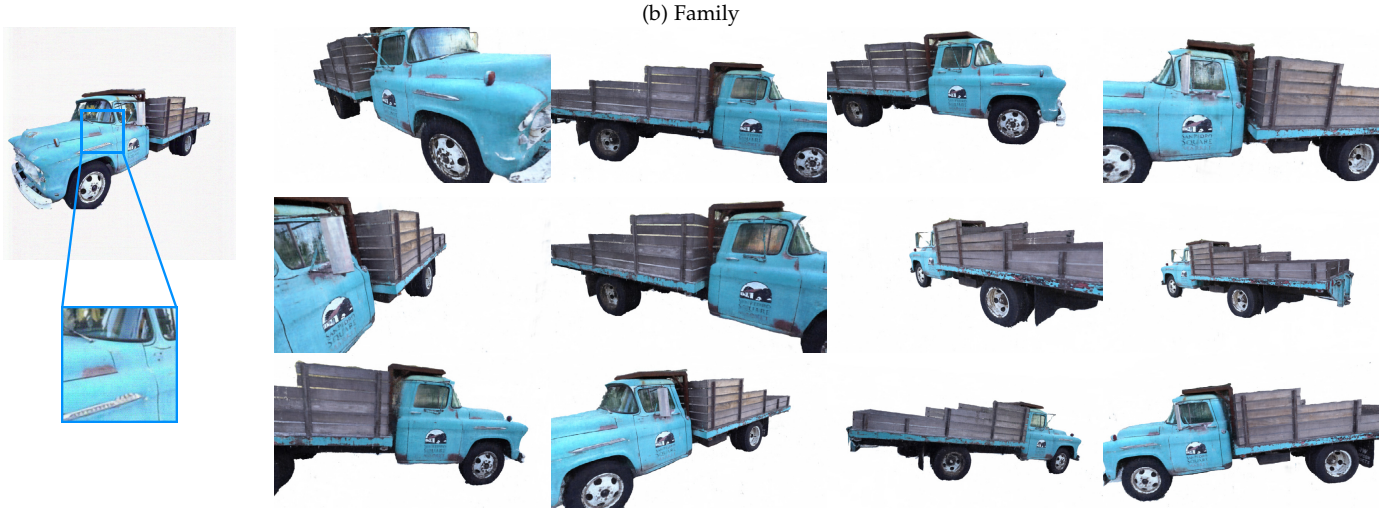


(a) GT (b) Original (c) Video Snapshot (d) IICNet (e) Ours

Fig. 5: Reconstruction qualitative comparison on Tanks and Temples [3] with the original NeRF model (b), and other invertible video methods (c & d).

(a) Caterpillar

(b) Family



(c) Truck

Fig. 6: Qualitative results of arbitrary view synthesis on Tanks and Temples [3].The leftmost column displays the 3D snapshot, while the right side shows the synthesized views from arbitrary perspectives.

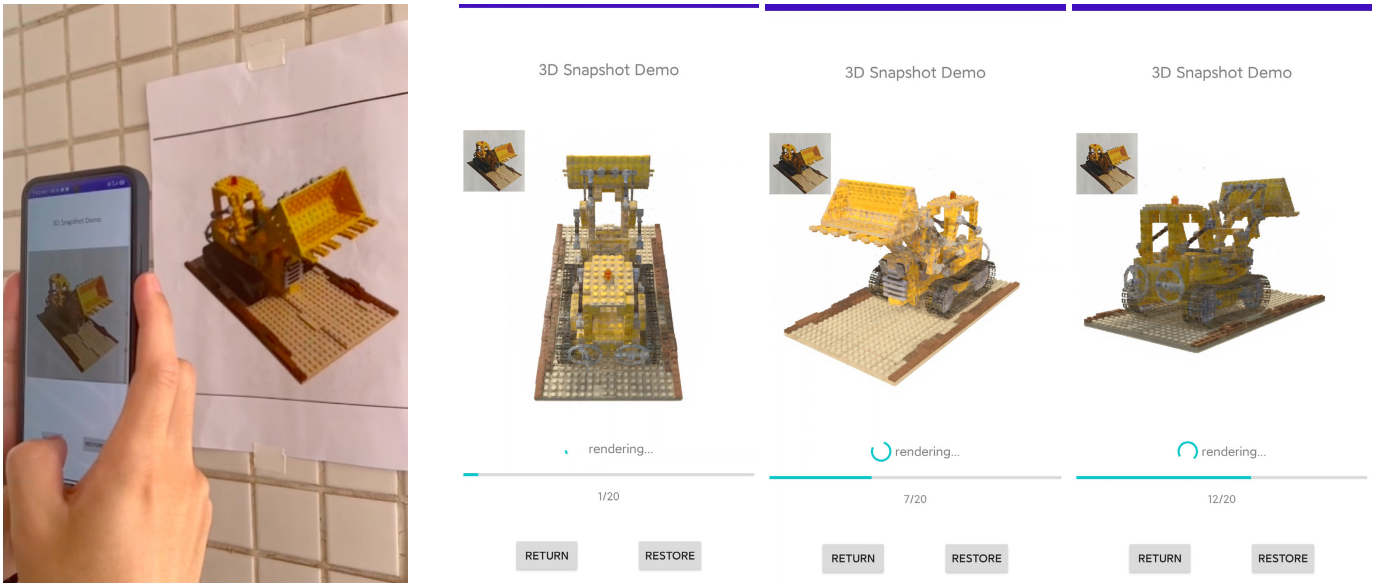


Fig. 7: Demo of a potential application of our technique on a printed paper. Left: a user using our mobile app. Right: screenshots of the mobile app, where a 3D model can be decoded in real-time.