TPD-NeRF: Temporally Progressive Reconstruction of Dynamic Neural Radiance Fields from Monocular Video (Supplementary Material)

Yu-Jie Yuan
Institute of Computing Technology, CAS
University of Chinese Academy of Sciences

Leif Kobbelt
RWTH Aachen University

Jie Yang
Institute of Computing Technology, CAS

Yu-Kun Lai
Cardiff University

Lin Gao
Institute of Computing Technology, CAS
University of Chinese Academy of Sciences
gaolin@ict.ac.cn

1. Overview

This supplementary document accompanies our main paper and provides the implementation details and more results. It contains five parts, including implementation details, more results, and some intermediate results.

- Section 2 provides the implementation details, including dataset details, network architecture and training details.
- Section 3 provides more results on dynamic novel view synthesis.
- Section 4 provides the visualization of the predicted 3D offsets on the image plane.
- Section 5 provides the visualization of the optical flow before and after the joint optimization.
- Section 6 provides the visualization of the predicted masks on both training views and test views.

2. Implementation Details

2.1. Datasets.

We use the monocular videos from the Nvidia Dynamic Scene Dataset [10] and Dycheck dataset [2]. We further generate some monocular videos from NeuralVideo [4] and ST-NeRF [11] datasets which are captured under a multi-camera system. We assign a camera trajectory for these cameras, which simulates the motion of a monocular camera and extract the images according to the time stamp to obtain the training set. Note that there is a 10 frame gap between the two cameras to avoid rapid camera movement and try to avoid the teleporting issue. We leave a camera alone to provide an evaluation view not seen during training. Other images are randomly selected as the test set. We use COLMAP [8, 7] to estimate the intrinsic and extrinsic parameters of the camera.

2.2. Network Architecture.

There are four networks in our method, including a local deformation network, a global deformation network, a rigidity network, and a space-time field network. All of them are multi-layer perceptron (MLP) networks. There are 4, 6, 5, and 9 fully connected layers in the local deformation network, global deformation network, rigidity network, and space-time field network, respectively.

2.3. Training details.

Our temporal progressive training consists of two deformation networks, which are trained in two stages. We first train the local deformation network for 300,000 iterations and then train the global deformation network with supervision from the fixed local network. During the training of the local network, we jointly optimize scene flow and optical flow, starting at 200,000 iterations. The rigidity network and space-time field network in the canonical space are shared in both stages and trained together with the corresponding deformation network. We use a warm-up strategy at the beginning of the training of the local deformation network. Specifically, we train from the first frame to the last frame in chronological order, with each frame trained for 500 iterations. After the warm-up training ends, we start to randomly sample frames for training. We randomly sample 1024 rays for training. Adam optimizer [3] with learning rate 0.0005 is used for the optimization.

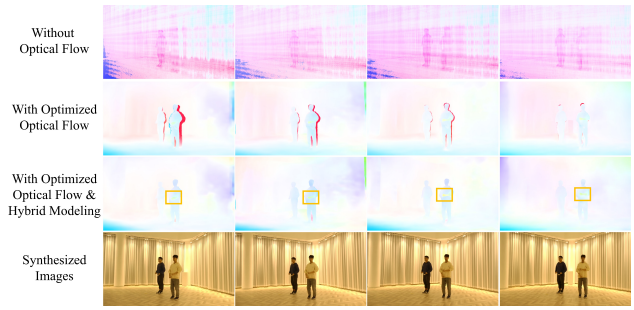


Figure 1. We visualize the aggregated 3D scene flow on the image plane. After adding the jointly optimized optical flow and the hybrid modeling of a space-time field network, the predicted scene flow conforms to the scene motions and is with fewer artifacts.

Figure 2. We visualize the optical flow before and after the joint optimization. The joint optimization can eliminate the inconsistent optical flow estimation in the static background and make the optical flow in the dynamic foreground sharp.

3. More Results

In this section, we show more comparison results on different monocular videos generated from [4] and [11] with existing methods (NSFF [5], D-NeRF [6], NR-NeRF [9], TiNeuVox [1]) in Figs. 3, 4, 5, 6 and 7. Among these figures, Fig. 3 shows the image synthesis results on the fully unseen views, and others show the results at different time stamps on the training views. These comparisons show that our method is superior to existing methods in maintaining dynamic details.

4. Visualizations of Offset

We project the aggregated 3D scene flow onto the image plane for visualization, and the results are shown in Fig. 1. Without the supervision of optical flow, the scene flow between two adjacent frames captured by the local deformation network is not meaningful. After adding the jointly optimized optical flow, the scene flow predicted by the network is more in line with the scene motion, but there are some artifacts at the edge of the human, which may be caused by the continuity of the network prediction. After adding the hybrid modeling strategy of a space-time field network, these artifacts disappear, and the scene flow still conforms to the human motions.

5. Visualizations of Optical Flow

Although the 2D-based optical flow estimation method has been extensively studied and is able to obtain good results, the lack of 3D perception will lead to artifacts in some details, which may affect the learning of scene flow. We propose a joint optimization strategy that allows 3D offsets and 2D optical flow to be jointly optimized. Although our goal is to achieve better image synthesis quality, we also bring in the multi-view constraint of NeRF to the optical flow. We visualize the optical flow before and after optimization in Fig. 2. As can be seen from the first row, we can eliminate the inconsistent optical flow estimation in the static background, while the result of the second row shows that we can make the optical flow in the dynamic foreground sharper.

6. Mask Visualizations

We use a rigidity network [9] to distinguish dynamic sampled points from static ones and decide whether to use the predicted offset to transform the sampled points. In order to verify that the rigidity network works well, we visualize the predicted mask on both training and novel views. Similar to the conversion from scene flow to optical flow, we first use volume rendering to aggregate the rigidity values at the sampled points of a ray and then project them onto the image plane. The visualization results are shown in Fig. 8. It can be seen that the predicted mask can better distinguish the dynamic part from the static part of the image.

References

[1] J. Fang, T. Yi, X. Wang, L. Xie, X. Zhang, W. Liu, M. Nießner, and Q. Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 2

[2] H. Gao, R. Li, S. Tulsiani, B. Russell, and A. Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022. 1

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

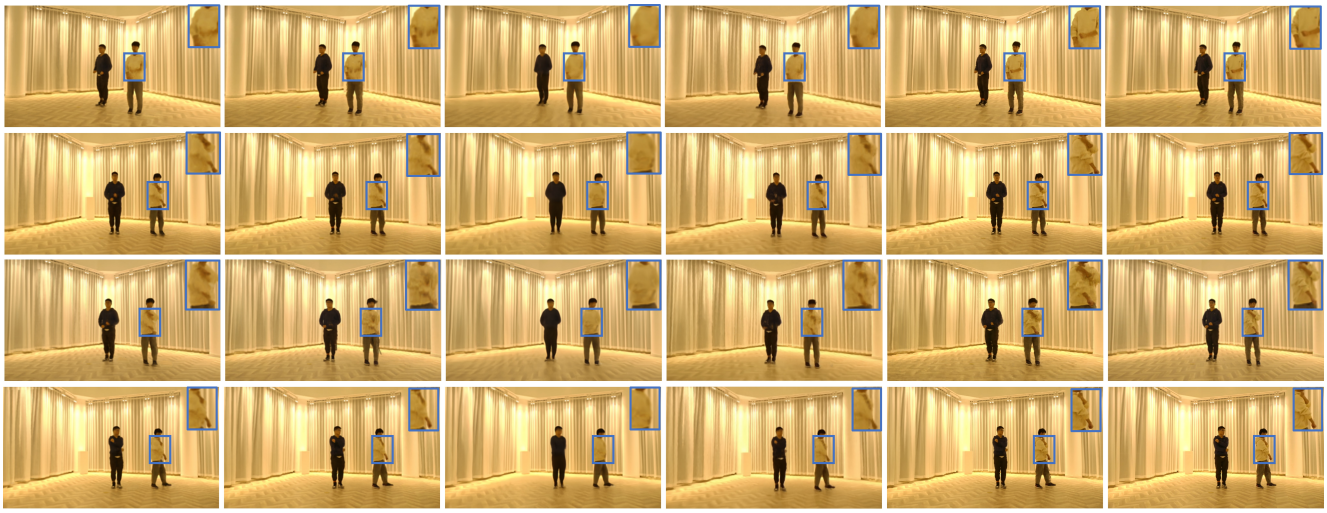
[4] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, and Z. Lv. Neural 3D video synthesis. 2021. 1, 2

[5] Z. Li, S. Niklaus, N. Snavely, and O. Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[6] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

(a) NSFF (b) D-NeRF (c) NR-NeRF (d) TiNeuVox (e) Ours (f) GT

Figure 3. More comparisons on the novel view synthesis with the left-alone camera.



(a) NSFF (b) D-NeRF (c) NR-NeRF (d) TiNeuVox (e) Ours (f) Ground Truth

Figure 4. More comparisons on the ST-NeRF dataset [11]. The synthesized images are obtained based on training cameras at specific times not included for training.

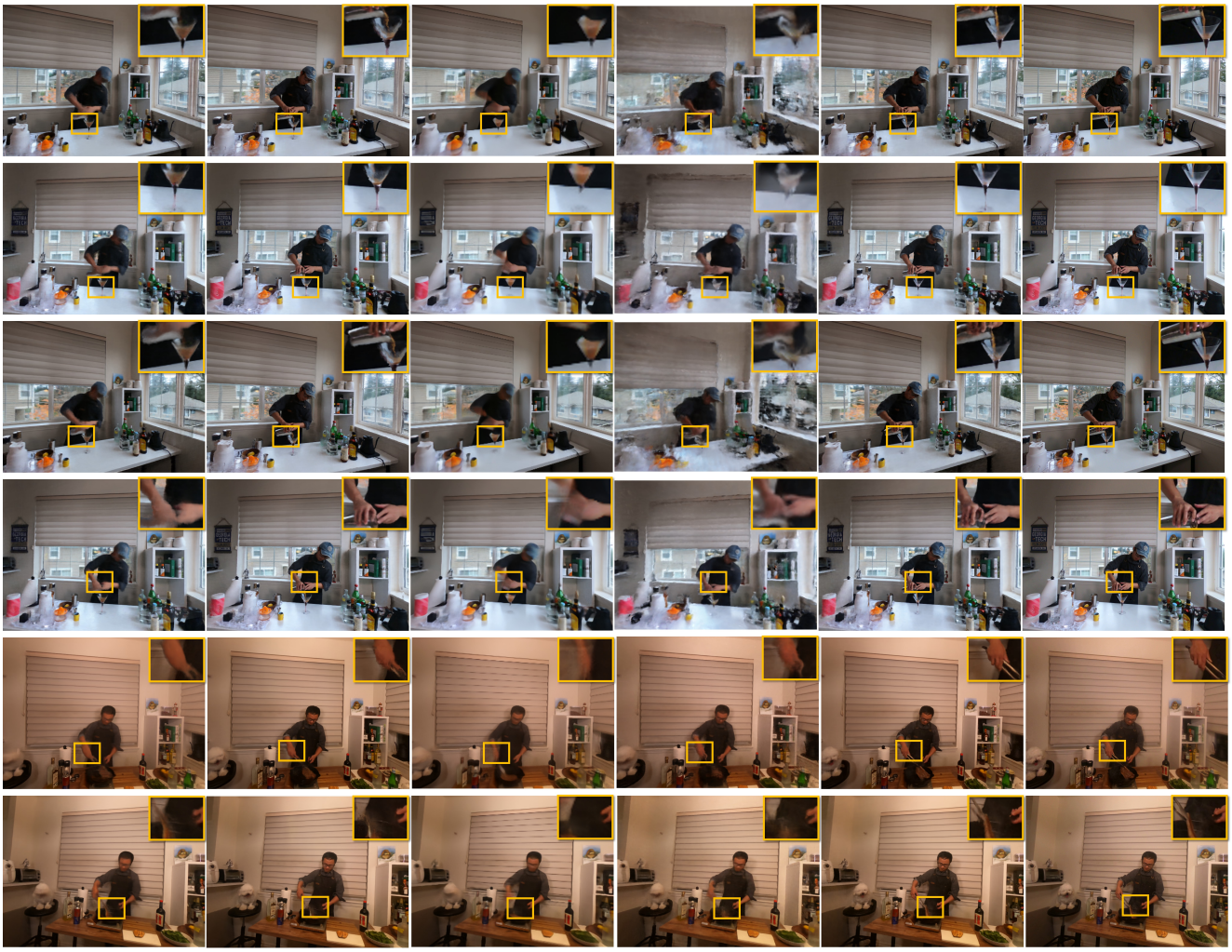
[7] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 1

[8] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. 1

[9] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhofer, C. Lassner, and C. Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2

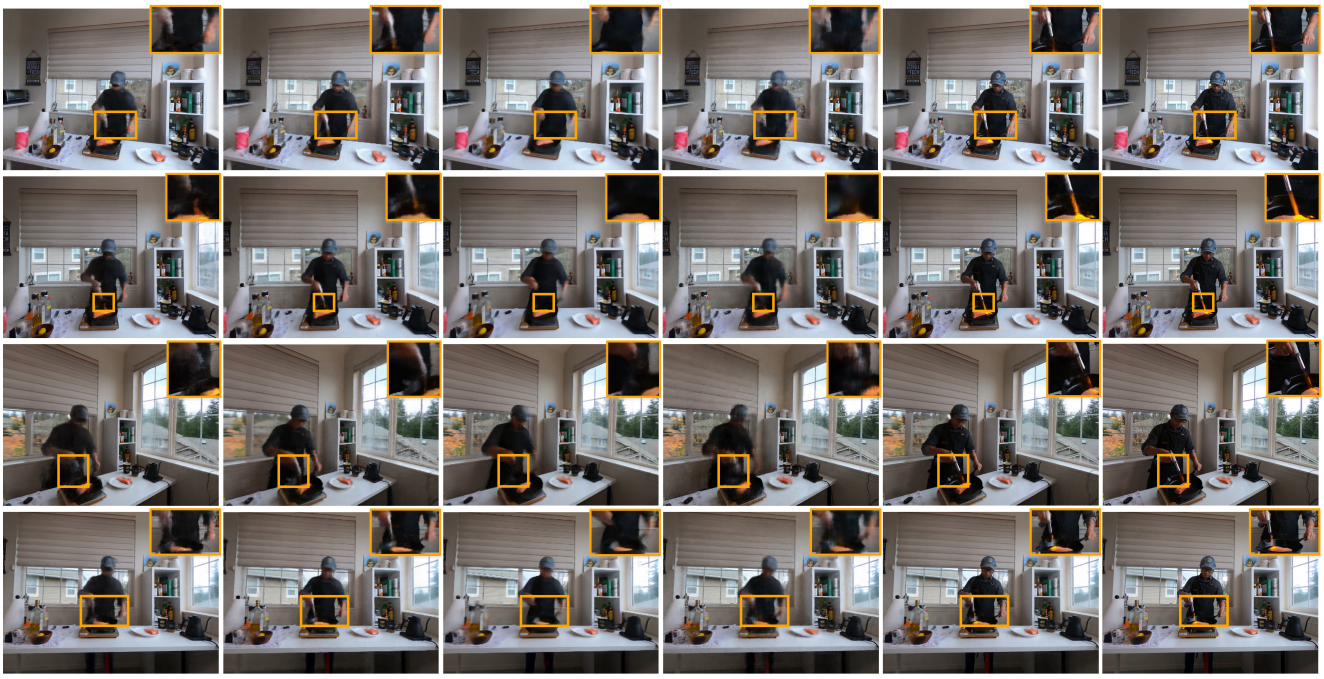
[10] J. S. Yoon, K. Kim, O. Gallo, H. S. Park, and J. Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5335–5344, 2020. 1

[11] J. Zhang, X. Liu, X. Ye, F. Zhao, Y. Zhang, M. Wu, Y. Zhang, L. Xu, and J. Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021. 1, 2, 3



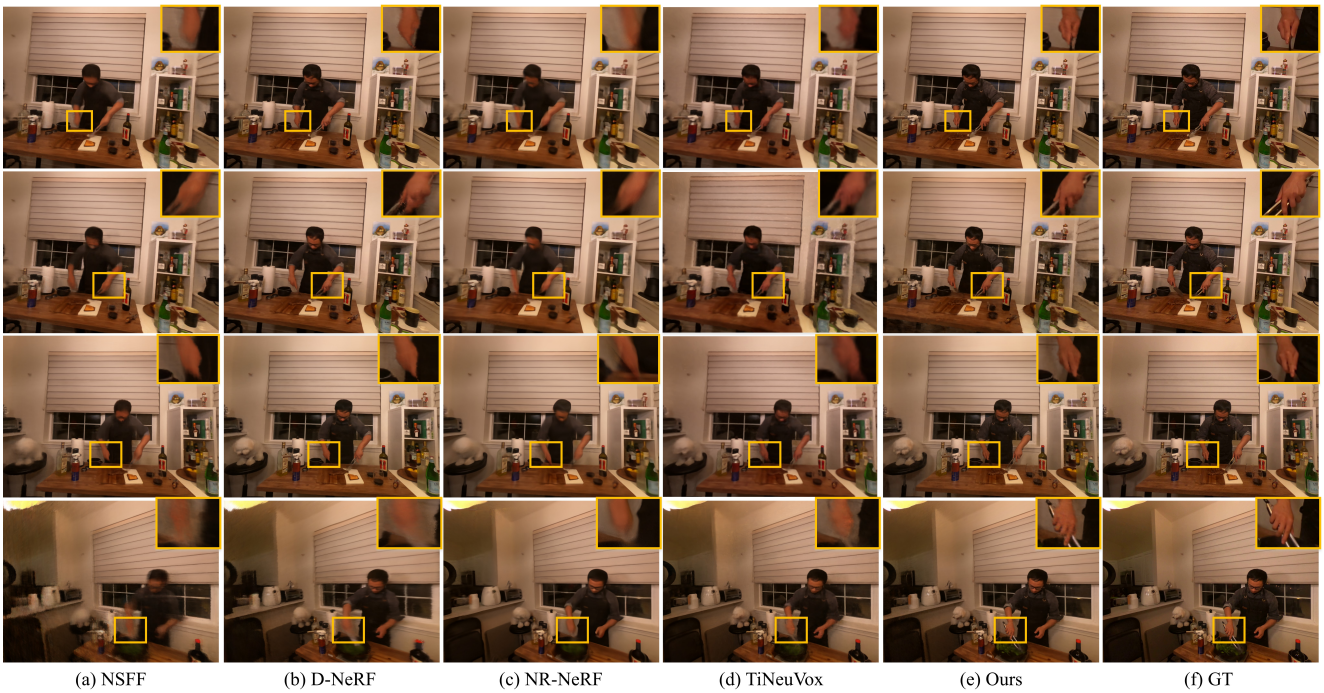
(a) NSFF (b) D-NeRF (c) NR-NeRF (d) TiNeuVox (e) Ours (f) GT

Figure 5. More comparisons on the novel view synthesis with training cameras at specific times not included for training.



(a) NSFF (b) D-NeRF (c) NR-NeRF (d) TiNeuVox (e) Ours (f) GT

Figure 6. More comparisons on the novel view synthesis with training cameras at specific times not included for training.



(a) NSFF (b) D-NeRF (c) NR-NeRF (d) TiNeuVox (e) Ours (f) GT

Figure 7. More comparisons on the novel view synthesis with training cameras at specific times not included for training.

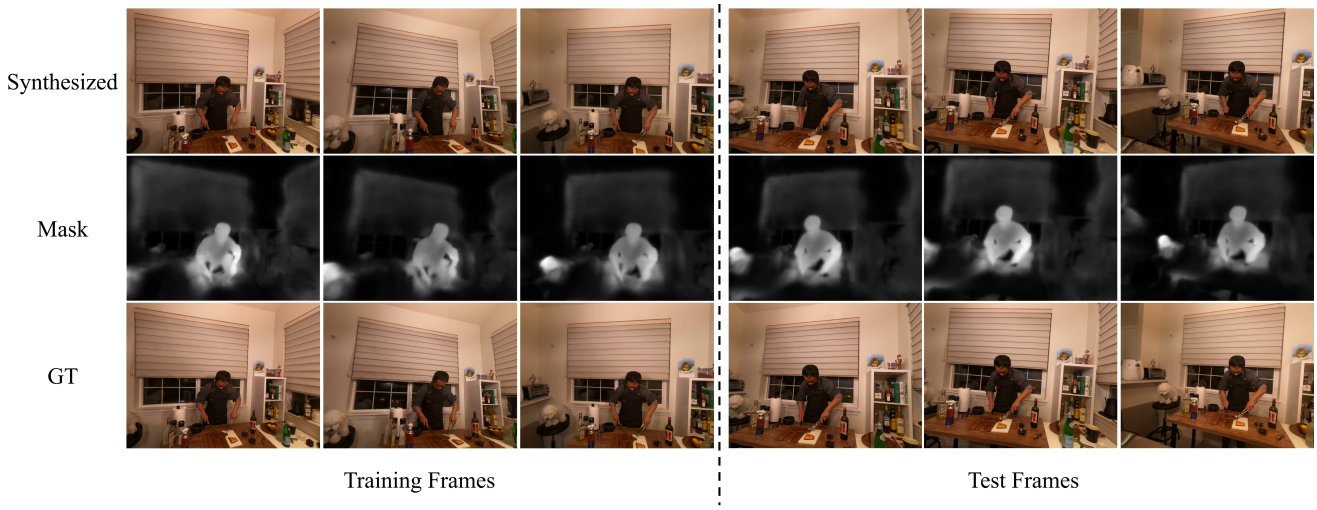


Figure 8. We visualize the predicted mask on both training views and novel views.