Joint Motion Detection in Neural Videos Training

Niloufar Pourian Alexey Supikov
Intel Labs
Santa Clara, CA

Abstract

Neural radiance fields (NeRF) can produce photo realistic free-viewpoint images. Recently, incremental neural video training approaches took a step towards interactive streaming via a frame-by-frame approach naturally free of lag. Motion detection in neural videos via a frame-by-frame approach can provide valuable cues to enable temporally stable neural videos suitable for interactive streaming. In addition, motion cues can be used to guide the ray sampling phase to model dynamic regions more efficiently. Hence, motion detection can be a key component in telepresence/social networking and immersive cloud gaming applications. In this paper, we propose a novel approach that computes static/dynamic separation masks with high accuracy and spatial coherency across different views together with NeRF optimization process. This is enabled by using explicit deformation network instead of implicit motions/structure layers (novel network architecture) as well as novel specifically designed training schedule. To the best of our knowledge, this is the first work that enables motion estimation via a frame-by-frame approach in a neural video training. The proposed work is desirable as it does not require buffer chunks of frames available before processing and hence is suitable for interactive streaming scenarios. Experimental results shows the effectiveness of the proposed motion detection approach in neural videos.

1. Introduction

High quality view synthesis in long duration video streams can be a key component in telepresence, social networking, and immersive cloud gaming applications. The idea of using spatiotemporal NeRFs [4, 7, 8, 10] for synthesizing 3D videos has gained popularity due to their impressive photorealism. However, such techniques suffer from an inherent lag as they consume videos and thus have to wait for chunks of frames (often seconds) before processing, making them unsuitable for interactive streaming scenarios.

Recently, incremental neural video training [19] took a step towards interactive streaming via a frame-by-frame approach naturally free of lag by showing the possibility of training each frame in minutes while maintaining a streamable size. While incremental neural video training approaches present new possibilities in interactive 3D streaming, there is still a lot to be explored in this domain. In particular, incremental training approaches generally suffer from significant temporal artifacts in rendered 3D videos, like pixel flickering. In order to improve neural radiance field's temporal stability via an incremental training approach, researchers [19] rely on precomputed masks classifying static and dynamic pixels via a separate pipeline to split modeling the background and foreground content. Hence, having an effective approach to accurately define static/dynamic regions within the NeRF training pipeline in a frame-by-frame fashion is crucial and is the main focus of this paper.

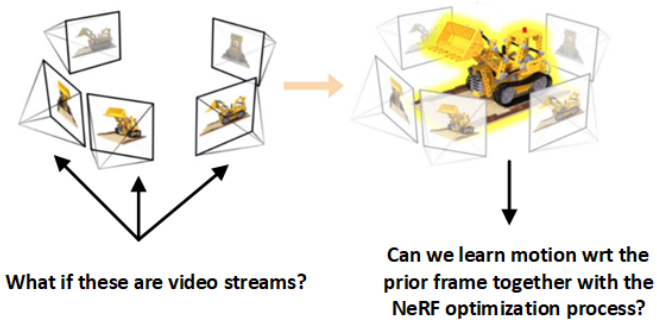

Figure 1. Overview of problem statement

In this paper, we propose a novel approach that computes static/dynamic separation masks with high accuracy and spatial coherency across different views together with NeRF optimization process. This is enabled by using explicit deformation network instead of implicit motions/structure layers (novel network architecture) as well as novel specifically designed training schedule. To the best of our knowledge, this is the first work that enables motion estimation via a frame-by-frame approach in a neural video training. The proposed work is desirable as it does not require buffer chunks of frames available before processing and hence is

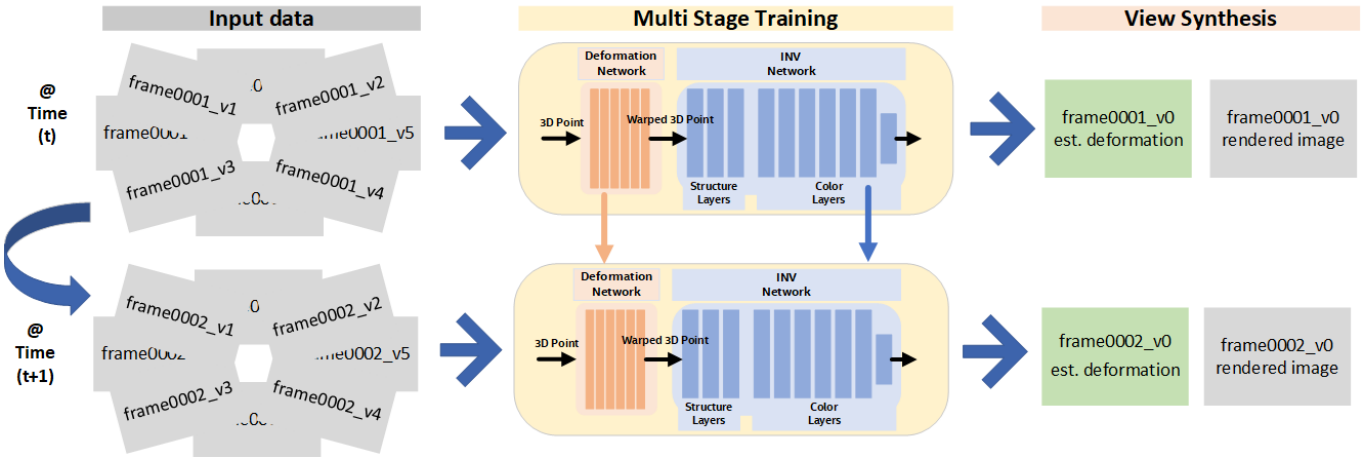

Figure 2. Overview of the proposed framework. As shown, our approach can estimate static/dynamic masks simultaneously within the NeRF training pipeline.

suitable for interactive streaming scenarios.

Experimental results show the effectiveness of the proposed motion detection approach in neural videos while reaching per-frame PSNR score similar to its baseline NeRF based approaches that require two NeRF networks, modelling static and dynamic content [19]. We show the applicability of the proposed work in enabling temporally stable neural videos suitable for interactive streaming. Further, our experiments indicate that such masks enable us to better model dynamic regions with the same number of iterations by guiding the ray sampling phase to slightly favor dynamic regions more.

2. Related Work

Traditional image based rendering techniques use pixel information from input images to synthesize views [1, 15, 16]. Recently, neural representations have shown high quality novel view synthesis [2, 10, 20, 24]. Specifically, Neural Radiance Fields (NeRF) [10] achieve an unprecedented level of fidelity by encoding continuous scene radiance fields within multi-layer perceptrons (MLPs).

One of the main challenges of applying static NeRF independently for long video streams is its slow training. To make NeRF applicable to dynamic scenes, researchers have explored various techniques [3, 4, 5, 8, 9, 12, 13, 14, 17, 18, 21]. In particular, [12, 13] use a deformation network for reconstructing non-rigid scenes via a learned deformation field mapping from coordinates in each input image into a canonical template coordinate space. [9] integrates classical image based rendering ideas into a volumetric rendering framework, rather than encoding 3D color and density directly in the weights of an MLP as in recent dynamic NeRF methods. Other methods represent scenes as time-varying NeRFs [4, 5, 8, 18, 21]. For instance, [8] uses neural scene flow fields that can capture fast and complex 3D scene motion for in-the-wild videos. However, all these approaches require buffer chunks of frames (often seconds) available before processing which are unsuitable for interactive streaming scenarios.

More recently, [19] introduces incremental neural video (INV) which allows interactive streaming via a frame-by-frame approach naturally free of lag. For each incoming new frame, INV improves on the knowledge from prior frames and thus reduces redundant learning. However, their approach suffers from flickering artifacts across frames. To help with temporal stabilization of the background, [19] relies on a set of static/dynamic separation masks precomputed using multi-frame optical flow map estimation techniques. To reach reasonable static/dynamic separation masks for each view, optical flow maps are computed between each frame and a set of prior frames independently. Such an approach results in inaccurate and inconsistent masks across views.

It is worth noting that while there is a wealth of published literature on optical flow estimation techniques [6, 22, 23], such approaches typically involve a separate pipeline and generally estimate a motion map between a single stereo pair via feature matching, hence they lack the ability to consider multiple views at once. In contrast, the proposed work introduces an approach to motion detection simultaneously within NeRF training pipeline (on a frame-by-frame basis) allowing for taking into account the multiview information.

3. Our Approach

Given a set of incoming video streams that are being captured from multiple viewpoints and are available only on a

Figure 3. Illustration of a rendered image and the estimated motion map using the proposed work

frame-by-frame basis, along with their known camera parameters, our goal is to learn motion with respect to the previous time stamp along with other scene attributes simultaneously with incremental NeRF training. The proposed approach requires only current timestamp frames to update radiance field and motion information as necessary information from previous frames is handled implicitly.

As illustrated in Figure 2, our framework enables motion detection within incremental neural video training and hence eliminates the need for a secondary pipeline to estimate motion. The details of the proposed work is as follows.

3.1. Network Architecture

As shown in Figure 2, our overall network structure consists of a deformation network and the INV Network as described in [19]. The deformation network is an MLP based deformation network that is added in the beginning of the INV network. For each 3D point (X, Y, Z) at its input, the deformation network estimates a rotation and translation with respect to the prior frame as $SE(3)$ field like in [13]. The estimated rotation and translation is used to define a transformation that is used in estimating a warped 3D point (X^w, Y^w, Z^w) at its output. The displacement/motion map is then defined as the normalized distance between the input 3D point and the estimated warped 3D point at the output of the deformation network. This displacement map provides the motion cues to define static/dynamic separation masks that can be used as a guide to improve flickering artifacts in the background and hence achieving temporally stable neural videos.

$$displacement = norm((X^w, Y^w, Z^w) - (X, Y, Z)) \quad (1)$$

The INV network used in our pipeline is similar to the network defined in [19] and has the task of estimating various scene attributes such as optical density and radiance.

3.2. Training Schedule

Our goal for adding the deformation network in the beginning of the INV pipeline is to separate the motion information from other attributes of the scene and to encode the motion information solely by the deformation network. This would allow INV to focus on learning other attributes of the scene.

We found that the learning the motion information between the current frame and the prior frame is not possible via an end to end training of the deformation and INV networks together. This is because concurrent training of the two networks would spill the motion information into other attributes and hence the parameters of the two models contribute to learning the motion mixed with other attributes of the scene.

To achieve our goal and prevent INV training from interfering with the deformation network trying to learn the motion, we follow a novel specifically designed training schedule. As illustrated in Figure 4, our multi-stage training procedure is as follows: At frame zero, we freeze the deformation network and train INV to learn the scene. Later for each incoming frame, the model has a good understanding of what the scene looks like, so we freeze the INV network parameters from prior frame and train the deformation network to learn the motion between the previous frame and the current frame. In the next stage, we freeze the deformation network, and train INV to account for the differences in the current frame that can not be modeled by applying the deformation map (learned by the deformation network) to the scene attributes of the prior frame (such as disappear-

ance of the flame from one frame to another). It is worth noting that the number of iterations to spend at each stage is found empirically.

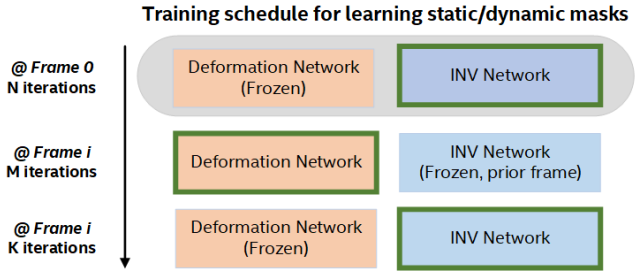

Figure 4. Illustration of the multi-stage training process for deformation+INV network

3.3. Motion Prior and Cost Function

To ensure smooth motion across frames and to avoid jittering artifacts in the rendered frames, we train motion network incrementally using the estimated motion from prior frame as an initialization for the motion parameter of the current frame. Further, at each frame f, a deformation loss $Loss^{motion}(f)$ is added to the cost function $Loss(f)$ penalizing large deviations from estimated motion of the prior frame, in addition to $Loss^{rgb}(f)$ penalizing the differences in rgb color values:

$$Loss(f) = Loss^{rgb}(f) + \alpha Loss^{motion}(f) \quad (2)$$

$$Loss^{rgb}(f) = MSE(I^{est}(f), I^{GT}(f)) \quad (3)$$

$$Loss^{motion}(f) = MSE(M^{est}(f), M^{est}(f-1)) \quad (4)$$

with I representing the rgb image, M representing the motion/displacement map across frames, and f denoting the frame index. In our experiments, α is set to 10^{-5}.

3.4. Positional Encoding

Similar to the NeRF optimization pipeline, it is important to define a positional encoding at the beginning of the deformation network to lift the input data to higher dimensions and to make it easier for the network to learn things. This will allow the network to preserve details and sharp edges in the image.

3.5. Modeling dynamic regions

In the last stage of our training phase, after the deformations are learned, we train the INV network to better model scene attributes within dynamic regions. This allows our end to end pipeline to accommodate for differences across frames that can not simply be modeled by moving points with respect to the prior frame, in other words it allows us to handle newness within the same network. In this stage, we take advantage of the additional motion cues learned and use a higher sampling ratio for dynamic rays compared to the static rays.

3.6. Application to video stabilization

Our proposed work provides the additional cues to effortlessly reach video consistency in static regions without introducing motion blur artifacts. In particular, one can use the average of the current frame and the prior frame within the static regions to significantly reduce the typically seen fluctuations in static regions and to reach improved temporally stable rendered videos.

4. Experimental Results

4.1. Dataset

In our experiments, we use a publicly available multi-camera Plenoptic Video dataset [7] and show experimental results on *flame salmon* video consisting of 19 different viewpoints along with their corresponding calibrations, and showing a mix of static and dynamic, opaque and volumetric content. For fair comparison, we adhere to the same training and evaluation pipelines, designating 18 views for training and one for evaluation.

4.2. Motion detection + View synthesis

Figure 3 illustrates the rendered image and the estimated motion map using the the proposed approach. As shown, our method achieves high quality view synthesis results while providing highly accurate motion mask.

Table 1 shows the comparison between PSNR scores using different NeRF methods for Plenoptic Video dataset. As shown, the proposed work maintains PSNR comparable to other state of the art techniques that are trained per-frame on videos while providing motion detection masks simultaneously. While training time associated with the proposed work is slightly higher than its baseline INV, we believe that deformation + INV network optimization will allow for further reduction in training time and this is one of the directions that we are currently exploring. In addition, it is good to note that the baseline INV network used in the proposed work can be replaced with other NeRF related techniques that allow for improved NeRF model training and inference such as [11].

Figure 6 shows some qualitative comparison between the estimated static/dynamic separation masks for different views via the optical flow pipeline of [23] and [22] versus the proposed work. As shown the proposed work can achieve more accurate and spatially consistent static/dynamic separation masks across different views.

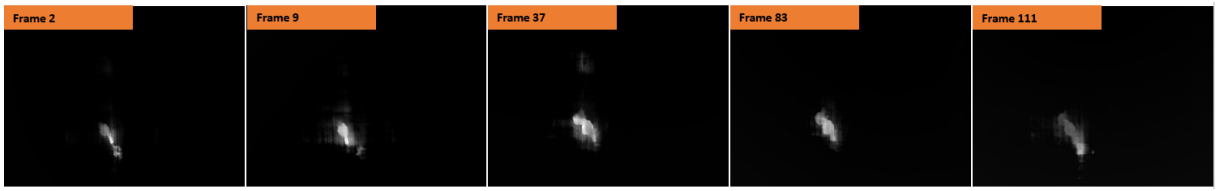

Figure 5. Illustration of stable static/dynamic mask estimation over time via the proposed work.

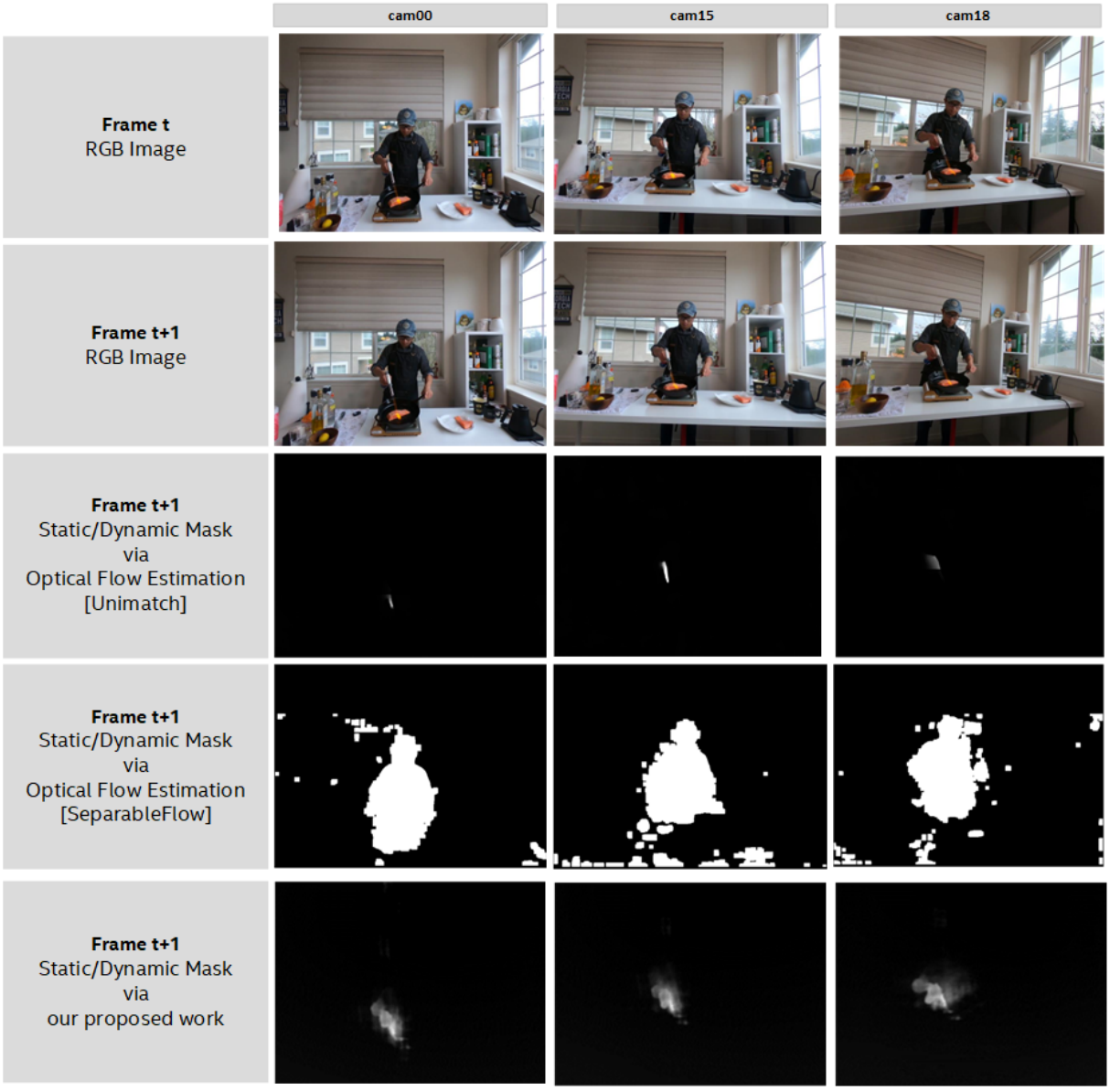

Figure 6. Estimated static/dynamic separation masks using the proposed approach vs. the optical flow-based approaches of Unimatch [22] and SeparableFlow [23]. Here, for [23] we show the accumulated motion over a set of nearby past frames to ensure recovering all dynamic regions. As shown, the proposed work achieves static/dynamic separation masks that are spatially consistent across views and are more accurately reflecting the motion of the hand/tool across frames.

Method	PSNR	Trained Per-Frame	Motion Detection	Training Time
NeRF [10]	24.62	✓	×	8min
DyNeRF [7]	29.58	×	×	260min
INV [19]	29.62	✓	×	8min
Ours	29.83	✓	✓	10min

Table 1. Accuracy of NeRF methods on Plenoptic Video dataset. Please note that, for fair comparison, our method should be compared with approaches that are trained per frame. As shown, the proposed work offers PSNR comparable to other state of the art techniques that are trained per-frame, while providing motion detection mask simultaneously.

Figure 5 shows that our motion detection approach can robustly estimate motion masks across frames.

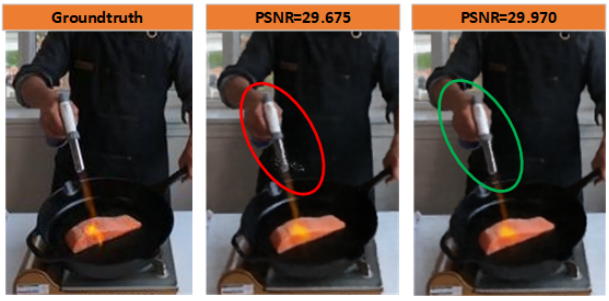

Figure 7. Groundtruth (left) and synthesized views using the proposed approach via (middle) random sampling of static and dynamic points versus (right) sampling points in dynamic regions at a higher rate in the last stage of INV network training. As shown, by using a higher rate of sampling for points in dynamic regions, we can improve the quality of rendered views.

Figure 10 shows the importance of updating the INV network parameters after learning the deformation across frames to account for the differences that can not simply be modeled by applying the motion field to the scene attributes of the prior frame. As shown, our approach can recover the disappearance of the flame or cases where the deformation field can not fully model the motion.

4.3. Improved Modeling of Dynamic Regions

In the last stage of the training, we take advantage of the additional motion cues learned and use a higher sampling ratio for dynamic rays compared to the static rays. This allows for a more accurate synthesized views within dynamic regions as shown in Figure 7.

4.4. Applications to Video Stabilization

As we describe in section 3, our end to end framework provides the additional motion cues necessary to stabilize static regions in the rendered video frames on the fly. For

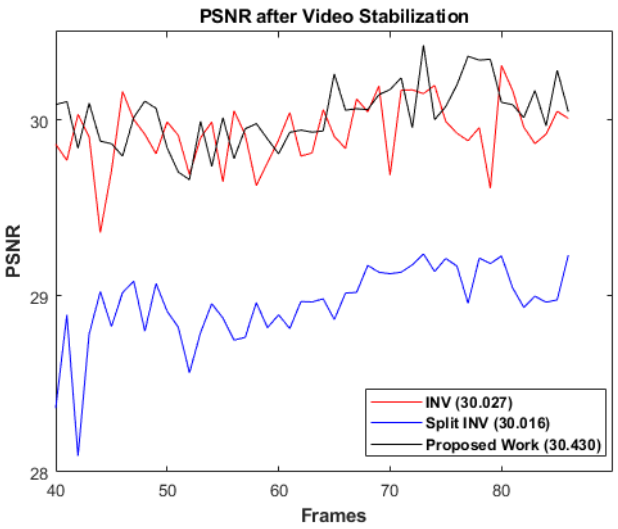

Figure 8. Comparison of the PSNR across frames for rendered videos via the proposed work after video stabilization and baseline methods of INV (without including the temporal stabilization) and Split INV (including the temporal stabilization). As shown the proposed work can provide video stabilization of the static regions without degrading the PSNR score.

comparison, Figure 8 illustrates the PSNR across a subset of the frames for the *flame salmon* dataset [7] via the proposed work, the INV network without any video stabilization [19], and split INV which is a proposed solution by [19] to reach improved temporal stability. It is good to note that split INV introduces two separate MLP pipelines for foreground and background content guided by the precomputed static/dynamic masks via a separate pipeline based on optical flow [23]. As shown, our proposed work achieves rendered video with PSNR scores close to the baseline work of [19] even after stabilization. Figure 9 shows the absolute difference between two consecutive rendered frames for the proposed work and the methods mentioned above. We can see that a simple video stabilization technique guided by our estimated static/dynamic separation masks is sufficient to remove many of the background flickering artifacts. Together, Figure 8 and Figure 9 prove the efficacy of our proposed work in achieving temporally stable rendered videos.

5. Conclusion & Future Directions

In summary, this paper presents an approach to estimating spatially coherent static/dynamic separation masks within the incremental neural video pipeline by proposing a multistage training process. In essence, we extract the displacement map between frames by adding a deformation network in the beginning of the INV pipeline following a multistage training that avoids INV from interfering with the deforma-

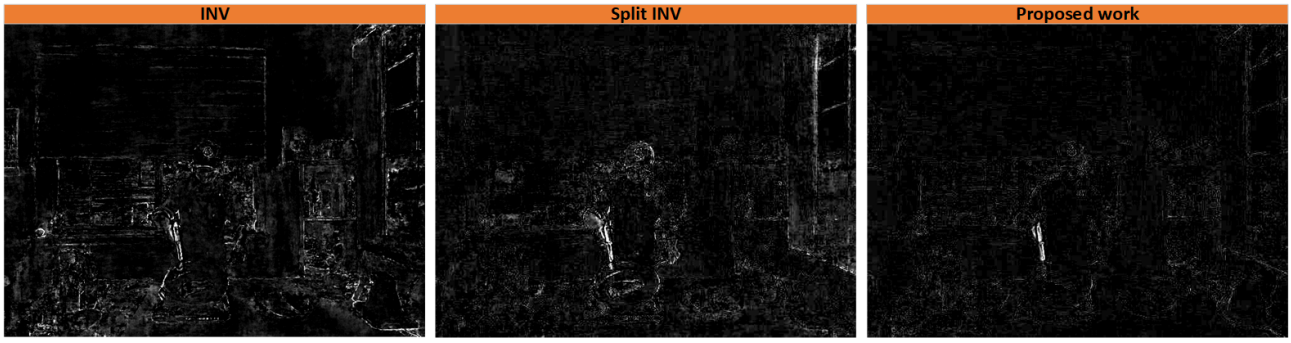

Figure 9. Comparison of the difference between two consecutive rendered frames via the proposed work and baseline methods. As shown, the proposed work can provide improved temporal stability for static regions.

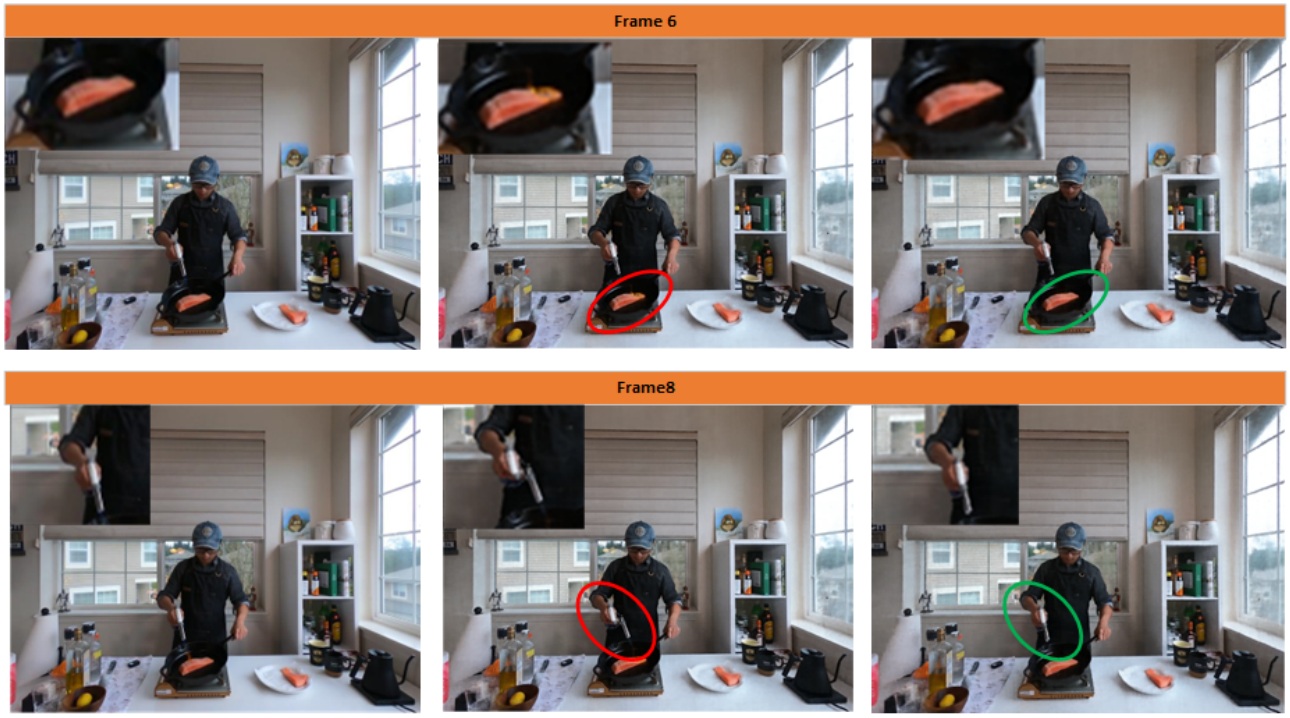

Figure 10. Illustration of the Groundtruth images (top) along with the synthesized views using the proposed work without (middle) and with (bottom) the INV network update stage after learning the deformation across frames. To better visualize the differences a cropped zoomed in version of the images are added as well.

tion network aiming to learn the motion information across frames. The proposed work is favorable as it allows estimating the deformation between frames without requiring the entire or large chuck of video to be available before processing. This is particularly important for enabling interactive streaming via a frame-by-frame approach naturally free of lag. Accurate static/dynamic separation masks allow temporal stabilization of the static background for rendered images. Further, one can improve the accuracy of synthesized views in dynamic regions by allocating a higher sampling rate for rays within dynamic regions. In future, we plan to explore INV network optimization and/or bandwidth reduction by possibly updating the INV network parameters at a slower rate (not per frame).

References

[1] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 307–318, 2000. 2

[2] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7781–7790, 2019. 2

[3] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314. IEEE Computer Society, 2021. 2

[4] 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. 1, 2

[5] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022. 2

[6] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 2

[7] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 1, 4, 6

[8] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver 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. 1, 2

[9] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023. 2

[10] 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. 1, 2, 6

[11] 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. 4

[12] Sinha U. Barron J.T. Bouaziz S. Goldman D.B. Seitz S.M. Park, K. and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[13] Sinha U. Hedman P. Barron J.T. Bouaziz S. Goldman D.B. Martin-Brualla R. Park, K. and S.M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. In *arXiv preprint arXiv:2106.13228*, 2021. 2, 3

[14] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc 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

[15] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing 2000*, pages 2–13. SPIE, 2000. 2

[16] Richard Szeliski, Steven Gortler, Radek Grzeszczuk, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on computer graphics and interactive techniques (SIGGRAPH 1996)*, pages 43–54, 1996. 2

[17] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian 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

[18] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 2

[19] Supikov A. Ratcliff J. Fuchs H. Wang, S. and R. Azuma. Inv: Towards streaming incremental neural videos. In *arXiv preprint arXiv:2302.01532*, 2023. 1, 2, 3, 6

[20] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8543, 2021. 2

[21] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 2

[22] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 4, 5

[23] Woodford O.J. Prisacariu V.A. Zhang, F. and P.H. Torr. Separable flow: Learning motion cost volumes for optical flow estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10807–10817, 2021. 2, 4, 5, 6

[24] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2