Fast novel view synthesis for multi-views video

Tuan Anh Tran
sid: 7015463
antr00001@stud.uni-saarland.de

Junaid
sid: 7023904
yyju00001@stud.uni-saarland.de

October 24, 2024

Abstract

We propose an approach for fast 3D video synthesis that takes a multiview video recordings of a dynamic scene as input and produces a video with novel view in each frame. Our method uses off-the-shelf fast neural radiance fields methods to reconstruct an expressive neural representation of each frame for a multiview video. We then use optimized model from each frame to produce a novel photorealistic view synthesis for that frame. We additionally use a continual training trick and a video denoise model to improve the consistency and smoothness of the final video. Our method can process a 10-second 30 FPS multiview video recording with 1K resolution in only under 1.5 hours to produce a novel-view video with an adequate level of photorealism. This is a huge boost from the prior state-of-the-art which requires 1300 GPUs hours for the same video. According to our knowledge, this is the first attempt that aims to solve the problem of fast 3D video synthesis from multiview video.

1 Introduction

Rendering novel photorealistic views of a scene from a sparse set of input images is important for many applications e.g. augmented reality, virtual reality, 3D content production, games, and the movie industry. Recent advances in the field of neural rendering, which optimize a neural scene representation that encodes both appearance and geometry, have shown impressive results [13, 21, 1, 2] that surpassed traditional methods like traditional Strutural-from-Motion [17]. One of the most prominent recent advances in neural rendering is Neural Radiance Fields (NeRF) [13] which learns an implicit volumetric representation of the scene given a number of 2D images of a static scene.

Extending to a more challenging direction, rendering photorealistic novel views of dynamic scenes is an interesting topic. Trivially including the temporal dimension into NeRF requires as large as 15000 GPU hours optimizing for 10 seconds, 30FPS multiview video [8]. Recently, DyNeRF [8] presents a solutions that produce a high level photorealism results which needs 1300 GPU hours for the same 10 seconds of video, which is a huge boost in terms of speed. Another line of research direction focuses on improving optimization speed for NeRF. For example, Instant-NGP [14] and Plenoxels [6] can speed up the optimization speed to the point that they allow real-time optimization and rendering.

In this work, we propose a solution for fast novel view video synthesis from a multiview camera that utilize fast NeRF models such as Instant-NGP and Plenoxels. Experiments show that our method can process a 300 frames video in only 1.5 hours (a boost of 850X compared to DyNeRF) and generate a novel view video with an adequate quality of 25.6 PSNR.

2 Related Works

Novel view synthesis for static scene: To reconstruct the static scene from a set of 2D images, mesh-based methods [3, 5], learning based methods [16],

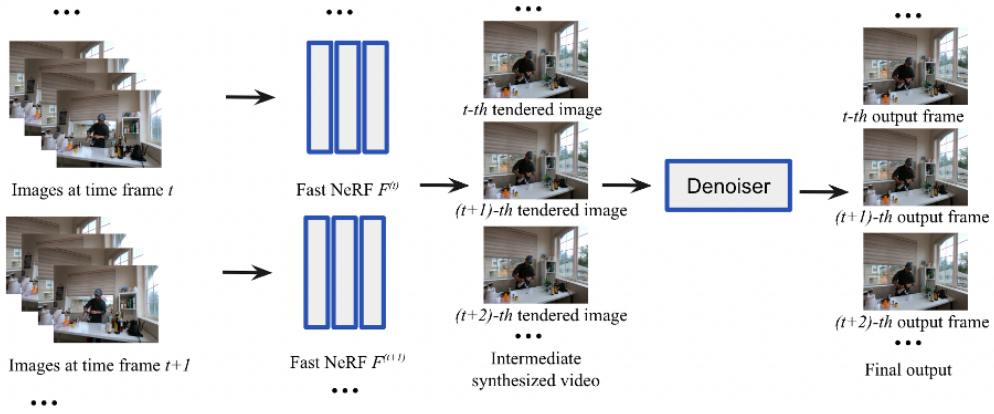

Figure 1: Proposed pipeline: we optimize each frame independently using a fast NeRF model. After finishing optimization, a novel view for each frame can be rendered to produce an intermediate novel view video. Finally, a denoiser is applied to produce a consistent and smooth output video.

Experiment	PSNR	JOD	Time (GPU-Hrs)
Setting 1	26.2	6.88	6.5
Setting 2	25.37	6.35	1.5
Denoised 1	26.34	7.01	6.51
Denoised 2	25.62	7.25	1.51
DyNeRF	29.59	8.07	1300

Table 1: Different Experiments Results

methods that use multi-planes images [18, 12] and voxels grids [9, 10] were introduced to accurately reconstruct the scene however they are either limited by in the small changes of viewpoints or facing memory consumption issue. These limitations are solved by NeRF which learns a volumetric implicit representation of the scene. However, while NeRF-based methods can reconstruct a scene with a high level of photorealism, often require a long processing time.

NeRF follow up for dynamic scene: Several works have been introduced for dynamic scenes. Given a monocular video, D-NeRF [15] learns a scene flow field that wraps the current frames into the next frames, then uses this flow field as additional input for the NeRF model to produce a novel view for the desired time frame. DyNeRF [8], the current state of the art, learns a temporal neural representation for several key frames of the video and uses interpolation to get the encoding of intermediate frames. This representation is used as additional input for a NeRF architecture to optimize the whole video. However, like NeRF, DyNeRF requires a long time for optimization, eg. 1300 GPU hours for 10 seconds, and 30 FPS videos with 1K resolution.

Faster NeRF rendering: Neural Sparse Voxel Recently, Plenoxels [6] introduces learnable 3D spare grids that store the scene as spherical harmonics and achieved a two-order of magnitude speed-up compared to NeRF. Instant-NGP [14] proposed a multi-scale hashing representation for NeRF with low-level code optimization and get several order of magnitude speed-up compared to NeRF.

3 Method

Our proposed pipeline is shown in Fig. 1. Given a multiview video, at time frame t, we are giveen a set of images $I = \{I_0^{(t)}, I_1^{(t)}, ..., I_N^{(t)}\}$. We optimize a fast NeRF model $F^{(t)}$ that reconstruct the scene at time frame t. At this step, the scene each time frame is encoded in a volumetric neural representation. To synthesize a novel view video, each $F^{(t)}$ is

then used for rendering an image from a novel view at t. This collection outputs from independently optimized NeRF are noisy when put together as a video. Therefore, we apply a denoiser on top of this output to make it more consistent. To further increase the consistency, we further apply a continual training trick which reduce most of the flickering noises from the intermediate output video.

4 Experiments

4.1 Dataset and Experiment Setups

We use the Neural 3D Video Synthesis dataset from Li et al. [8]. The dataset consists of several multiview videos each consisting of 18 views. Like DyNeRF we used one center view for testing and the remaining view for training. To get the camera poses of a video, we apply SfM [17] on the first frame. All the videos are processed and output at 1K resolution. The video `coffee_martini` and `cut_roasted_beef` are used for the remaining experiments.

4.2 Fast NeRF methods

Plenoxels and Instant-NGP methods are considered to be the fastest novel view synthesis methods right now. We randomly take several time frames from the `flame_steak` and optimize Ploxenal and Instant-NGF on them. Average PSNR scores at different optimization times are reported in Fig. 3. The result shows that both Plenoxel and Instant-NGP converge very fast and Instant-NGP consistently outperforms Plenoxels. We, therefore, use Instant-NGP as our main scene reconstruction method for the remaining experiments.

We further trained Instant models with two different settings. The results from this experiment is shown in Fig. 4 and Tab. 1

Setting 1: Optimization frame-by-frame independently. In this setting, we optimize each frame independently for 5000 iterations. The 30 FPS 10 seconds video took around 6.5 hours to process.

Setting 2: Optimization using the pre-trained checkpoint. In this setting, we optimize NeRF model for 3000 iterations on the first frame. For the second frame onward, we use the pretrained checkpoint from the last frame to optimize the current around 1500 iterations. Only 1.5 hours are needed for a video. This setting is our continual training trick.

4.3 Post Processing: Denoiser

We applied different denoisers to reduce the noise in the reconstructed scenes. The denoisers included FastDVDNet [19], TecoGAN [4], Maxim [20], and RefVSR [7].

FastDVDNet: It is a fast video denoising method that enforces temporal coherence by using multiple frames to denoise a single frame.

TecoGAN: A conditional video generation framework. With the help of adversarial loss and Pingpong loss, it avoids the temporal accumulation of artifacts which benefits the recurrent architecture for video denoising.

Maxim: Maxim presented a multi-axis MLP-based architecture called MAXIM that allows for efficient spatial mixing of local and global features, and a cross-gating block, an alternative to cross-attention, which accounts for cross-feature conditioning.

RefVSR: RefVSR introduced a reference-based network that utilizes the reference image/videos to perform superresolution on input.

5 Results and Analysis

Evaluation Metrics: To assess the performance of reconstructed results we employ the following metrics: PSNR and JOD [11]. JOD is a video quality metric that is sensitive to temporal aspects such as flickering. Higher PSNR scores and JOD scores indicate better-reconstructed results to the reference video.

Performance: We have trained our models in two different settings to observe the performance of our model. The first setting produces strong flickering noise with high inconsistency. Our continual training tricks make use of the reconstructed scene from the

Figure 2: Result of Different Denoisers

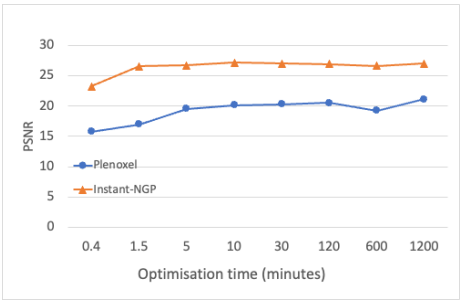

Figure 3: Comparison of novel view synthesis methods. **X-axis**: The training time of each model. **Y-axis**: PSNR achieved during different training point

previous frame which requires much less processing time Tab. 1. Fig. 4 also show that setting 2 results in smoother and more consistent video (green area) which lead to better reconstruction (red area) compared to inconsistency from setting 1 (red box). We tried different denoiser (Tab. 2 and Fig.2) in our post-processing setting to reduce the visible noise. FastDVDNet showed the best improvement in video quality and strongly reduce inconsistencies with a 15% improvement in JOD score. Furthermore, the FastDVDNet benefit most from our setting 2 (Denoised 2) which is compared to setting 1 (Denoised2) 1. This suggest that our use of continual trick works not only allow significantly lower time processing but also allow better final output.

Compared to DyNeRF which is the SOTA, we achieved 850x improvement in speed while still allow reasonable reconstruction.

Denoiser	PSNR	JOD	Time (Seconds)
FastDVDnet	**25.6**	**7.25**(+14.2%)	**40**
TecoGAN	25.3	6.25(-1.5%)	120
Maxim	25.2	6.48(+2.1%)	1500
RefVSR	25.5	6.5(+2.4%)	2000

Table 2: Different Denoiser Results

6 Conclusion

We introduced a methods that use Instant-NGP for synthesizing novel view video synthesis of dynamic scene. It can produce result with an adequate quality while require a small processing time (850x faster than DyNeRF). The results suggest this is a promissing research direction. Further direction for improvement can be finding a way to encode time dimension into the model which will allow a single model for the whole video and save a lot of space. To reduce flickering noises, dynamic and static areas can be treated differently like in DyNeRF. Also, further geometry bias (eg. local smoothness in depth map) can be introduced in the loss function to reduce floating artifacts. For more animated results, see our presentation slide here: slide

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Confer-*

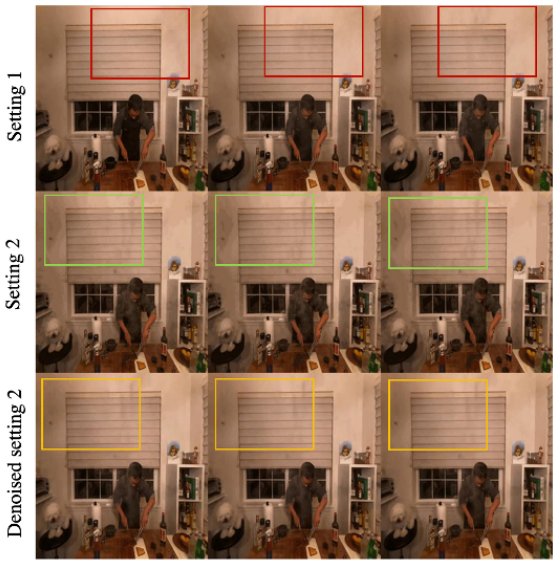

Figure 4: Results on 3 consecutive frames: **First Row (Setting 1)**: Result of frame-by-frame optimization, **Second Row(Setting 2)**: Result of using pretrained checkpoint for training other frames, **Third Row(Denoised)**: Denoised Results after applying FastDVDnet

ence on Computer Vision, pages 5855–5864, 2021. 1

[2] 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. 1

[3] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 425–432, 2001. 1

[4] Mengyu Chu, You Xie, Laura Leal-Taixé, and Nils Thuerey. Temporally coherent gans for video super-resolution (tecogan). *arXiv preprint arXiv:1811.09393*, 1(2):3, 2018. 3

[5] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20, 1996. 1

[6] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 1, 2

[7] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17824–17833, 2022. 3

[8] 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, 2, 3

[9] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2

[10] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 2

[11] Rafał K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021. 3

[12] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[13] 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. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1

[14] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives

with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 1, 2

[15] 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

[16] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, pages 623–640. Springer, 2020. 1

[17] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 3

[18] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019. 2

[19] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1354–1363, 2020. 3

[20] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022. 3

[21] 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. 1