Please fill in the following information after you read the paper.
Name: Darshit Miteshkumar Desai, Dir ID: darshit, UID: 118551722

[Paper tile] DynIBaR: Neural Dynamic Image-Based Rendering

[Summary] Describe the key ideas, experiments, and their significance.

The paper addresses the problem of synthesizing novel views from a video clip which contains a host of dynamic scenes. Recent methods have shown advances in synthesis of novel view synthesis using Neural Radiance Fields techniques. Techniques such as HyperNERF use a hyperspace to ensure that it records changes in topology, a prior work by the author's of this paper called NSFF also uses neighboring views to perform spatial and temporal stabilization of the video feed. But all of these methods suffers to render high-quality views along with long video lengths. They also fail to account for complex camera and scene motion. These methods are also limited to its application only to structured videos and cannot be applied on casual and in the wild videos. The author's previous work on NSFF also struggles from long duration videos and is only effective for 1 second length of videos which are supposed to be only forward facing.

In this paper, the author's plan to alleviate all the above disadvantages by proposing an approach that is scalable to dynamic videos which has a long duration, have uncontrolled camera trajectories and can be applied to unbounded scenes and for fast and complex object motion. In the related works the authors talk about various works in the field of novel view synthesis, one such method is image-based rendering using classical techniques. Other notable works include IBRNet which combines the classical technique with the latest Nerf based volume rendering techniques. The authors conclude that the focus of this paper would be on synthesizing high quality novel views for long videos which the above methods fail to do.

The methodology is described in a three-pronged approach. The authors consider a set of image frames with know parameters of the cameras. The model is trained per video which is then later used to render new views. The first section which is termed as motion adjusted feature aggregation describes the author's method of extracting colour and density from a set of temporally nearby source views. They consider a set of views in a predefined view radius and extract a 2D feature map from a shared convolutional encoder network. The author's then use this to form an input tuple of Image, Camera Parameters and Feature vector which is then aggregated using motion trajectory fields. The author's come up with a motion estimation scheme where each point in the target ray is represented by a basis vector using a MLP. This in turn helps in predicting the motion trajectories of the 3d point which can be used to find the color and the feature vector from the warped points. This is then passed through a ray transformer network along with a time embedding which gives colors and densities per sample. This is then rendered using volume rendering which is used in NeRF to get the final pixel colour.

The authors also enforce temporal coherence to account for overfitting of the novel views by using cross time rendering for temporal consistency, Apart from this the model is further bifurcated to process static and dynamic objects in the form of segmentation masks. The training and evaluations are done one Nvidia Dynamic scene dataset and

UCSD dynamic scene dataset. The camera poses are estimated using colmap. The authors also come up with a complicated view selection technique for in-the-wild videos for evaluation.

In conclusion a novel view synthesis approach using monocular video is proposed in this paper which can be applied on standard and in the wild dynamic scene videos and has achieved significant improvements over prior state of the art methods and dynamic scene benchmarks.

[Strengths] Consider the aspects of key ideas, experimental or theoretical validation.

The strengths are listed as below:

1) A key insight from this method is that a complete compression of the scene contents of a video is not required into a giant MLP. As from IBRNet the pixel data from neighboring pixels are directly used to render new views.
2) On quantitative metrics like PSNR signal to noise ratio, LPIPS for perceptual similarity and SSIM for structural similarity this method outperforms each of the other methods like HyperNeRF, NSFF and Dynamic view synthesis.
3) Prior dynamic NeRF methods fail to reconstruct photo realistic scale reconstructions when compared qualitatively for synthesis of novel views.
4) In this method the authors have improved the ability of view synthesis to be done on long duration, dynamic scene monocular videos and achieved state of the art performance.

[Weaknesses] Consider the aspects of key ideas, experimental or theoretical validation, writing quality, and data contribution (if relevant). Explain clearly why these are weak aspects of the paper

The weaknesses are mentioned below:

1) The method is limited relatively small viewpoint changes compared to other methods designed for static and quasi-static scene.
2) The synthesized views are not multi-view consistent and the rendering quality is highly dependent on the selection of source views.
3) The method also fails when rendering dynamic contents which are only visible in distant frames
4) The method also fails to estimate fast moving small objects as it initializes the the initial depth and optical flow which might contain noise or are incorrect.

[Reflection] Share your thoughts about the paper. What did you learn? How can you further improve the work?

In this paper I learnt about the state-of-the-art monocular video-based view synthesis algorithms like HyperNeRF, NSFF. I also learnt about two additional dynamic scene datasets categorized by the authors as NVIDIA and UCSD dynamic scene datasets. I also learnt other applications of NeRF like flow predicition and a new evaluation metrics like LPIPS and SSIM which are used frequently in the domain of view synthesis and Volume rendering. Additionally I learn about various applications of this algorithms like Video

Stabilization and Video Bokeh which were new to me. A possible extension of this work in the future could be to use this in different robotics applications like SLAM and scene reconstruction.