Still To Reel: Visualising the Frame

Chintan Shah

***Abstract*—We introduce a novel technique for 3D cinemagraph generation that synthesizes realistic animations by seamlessly combining visual content animation with camera motion. Starting from a single still image, our method addresses key challenges such as depth inaccuracies, motion field errors, and artifacts that arise when naively merging existing 2D animation and 3D photography techniques. Leveraging advanced depth estimation and inpainting, our approach converts input images into layered depth representations and adapts them into robust 3D motion fields using a hybrid Runge-Kutta Eulerian flow. To address issues like depth discontinuities and the emergence of visual artifacts, we implement customized smoothing, depth clipping, and bilateral filtering to ensure consistent and natural results. By bidirectionally displacing features in the 3D scene and synthesizing novel views through projection and blending, our method achieves visually coherent and artifact-free animations. Quantitative evaluations and user studies demonstrate the effectiveness and robustness of our approach, offering a significant step forward in 3D cinemagraphy.**

Keywords— 3D Motion, Novel Views, Euler Motion, Runge Kutta, Gaussian Blur

I. Introduction

Images have been a captivating medium for storytelling and expression for centuries. Their ability to capture a moment in time has made them a cornerstone of art and communication. But what if we could breathe life into still images and make them move? This intriguing idea has gained traction in recent years with the advent of cinemagraphs. Cinemagraphs combine the timeless allure of photography with the dynamism of motion, creating mesmerizing short videos where only a part of the image moves while the rest remains static. Despite their aesthetic appeal and growing popularity, creating cinemagraphs is not without challenges. Issues such as parallax distortions, inaccurate motion field estimation, erroneous depth predictions, camera motion inconsistencies, and visual artifacts often hinder the seamless transformation of a still image into a compelling cinemagraph. Recognizing these challenges, we sought to build upon this fascinating concept to create more robust and visually striking cinemagraphs. Our objective was to develop a robust 3D cinemagraph generation method capable of addressing the limitations of existing approaches. Specifically, our goals included:

1. Producing accurate motion fields to ensure smooth transitions between moving and static regions.

2. Mitigating issues arising from erroneous depth predictions, particularly in areas involving thin structures and intricate geometries.

3. Enhancing the quality of RGB and depth inpainting for a more seamless and visually cohesive output.

To achieve this, we propose a framework inspired by Xingyi Li et al.'s innovative approach, which utilizes a joint task of image animation and novel view synthesis. Our enhancements include replacing the original Eulerian flow method with a hybrid Runge-Kutta-Eulerian flow approach, which offers greater precision in motion field generation. Additionally, we refined inpainting techniques to improve the quality of reconstructed regions, further elevating the realism and visual appeal of the generated cinemagraphs. By addressing these challenges and building on an established foundation, our project aims to find innovative ways for 3D cinemagraph generation, turning a single still image into a captivating moving visual experience.

II. Related Works

In this section, we discuss foundational works related to single-image animation, novel view synthesis, and space-time view synthesis. Several methods have been proposed for animating still images, including approaches centered on physical object simulation. However, these approaches often exhibit limitations when applied to real-world photographs and complex scenes.

Motion transfer [1,2] has emerged as a significant development in this domain, utilizing reference videos to animate static objects. Despite their utility, such methods are highly dependent on reference footage or prior object knowledge, which limits their applicability to this research. Generative adversarial networks (GANs) have also contributed notably to image and video synthesis. Although GANs [3,4] can generate realistic transformations, they often lack the ability to decouple motion from appearance. Recently, diffusion models [5] have shown potential to produce high-quality videos from images or text, albeit at the expense of considerable computational overhead. Efficient techniques leveraging learned motion [6] priors have demonstrated promise in animating single images into video textures.

In the realm of novel view synthesis[7], earlier methods required dense multi-view[8] datasets to capture diverse camera angles. More recent techniques have shifted towards learning 3D scene representations

from single images, which align more closely with this research's objectives. However, these methods typically assume static scenes, limiting their realism in dynamic environments. Space-time view synthesis, on the other hand, explores the generation of novel

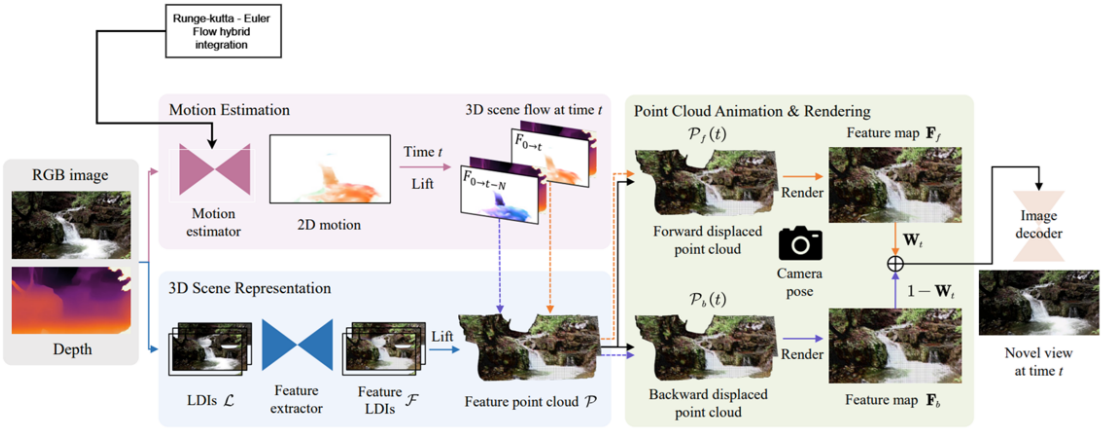


Fig. 1: The pipeline generates novel views of a scene by combining RGB images and depth maps. It estimates 3D scene flow using motion estimation and represents the scene as a feature point cloud. Forward and backward point clouds are rendered, blended, and decoded to produce a realistic image from a new camera perspective at a specific time.

viewpoints for dynamic scenes using neural rendering techniques [9,10]. Despite advancements, the use of real-world data remains a challenge for many of these approaches. The work presented in [11] introduces the novel task of creating 3D cinemagraphs from single images by connecting image animation with novel view synthesis. This innovative method enables plausible scene animations while allowing camera movements, demonstrating flexibility and customizability for user control. By integrating masks and flow hints into the motion estimation process, users can dictate how the scene is animated. Furthermore, the approach generalizes well to diverse input types, including in-the-wild photos, paintings, and synthetic images generated by diffusion models. Building on these foundational works, our research addresses critical limitations in existing 3D cinemagraph generation methods. Specifically, we focus on producing accurate motion fields to ensure smooth transitions between moving and static regions, mitigating challenges arising from erroneous depth predictions, particularly in thin structures and intricate geometries, and enhancing the quality of RGB and depth inpainting to achieve seamless and visually cohesive outputs.

III. Methodology

A. Overview

Our aim is to generate synthetic novel views from a single frame. We base our method on the work of Xingyi Li. Motion estimation is crucial for generating smooth cinemagraphs. Following [11], we utilize motion estimation and depth maps. While adhering to their framework, we have focused on improving key areas, including accurate motion fields and mitigating erroneous depth fields. Additionally, we improved the quality of RGB and depth inpainting to ensure seamless results. These improvements were applied to various components of the existing framework.

B. Motion Estimation

Motion estimation is critical to video generation and rendering. In [12], the authors propose a framework that uses Eulerian motion fields to generate velocities in scenes. Eulerian representations, as used in [13], observe fluid motion by tracking changes at fixed spatial points. However, challenges such as numerical diffusion may arise, causing inaccuracies in capturing steep gradients and sharper interfaces [14].

Accurate motion flows are essential for cinemagraphs to enhance visual quality. To achieve this, methods that mimic dynamic natural motion flows are necessary. While Euler's methods provide good approximations, they struggle with rapid and minute changes in motion. To address this, we implemented a higher-order Runge-Kutta method, which is better suited to capture fine motion changes [15].

According to the current framework, the motion field is defined as:

$$F_{t\to t+1}(.) = M(.),$$

where $F_{t+1}(.)$ represents the optical flow map from time t to $t+1$. The next frame is computed as:

$$X_{t+1} = X_t + M(X_t),$$

where X_t represents the coordinates of a pixel at time t.

We implemented a 4th-order Runge-Kutta method to solve the ordinary differential equation (ODE) of the form:

$$\frac{d\text{flow}}{dt} = f(\text{flow}, t),$$

where flow represents the motion field, and $f(\text{flow}, t)$ is the velocity of the flow field over time.

The 4th-order Runge-Kutta method is expressed as:

$$\text{flow}(t+\Delta t) \approx \text{flow}(t) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

where:

$$\begin{aligned} k_1 &= \Delta t \cdot f(\text{flow}, t), \\ k_2 &= \Delta t \cdot f(\text{flow} + \frac{k_1}{2}, t + \frac{\Delta t}{2}), \\ k_3 &= \Delta t \cdot f(\text{flow} + \frac{k_2}{2}, t + \frac{\Delta t}{2}), \\ k_4 &= \Delta t \cdot f(\text{flow} + k_3, t + \Delta t). \end{aligned}$$

This method integrates the flow field more accurately over each time step, providing smoother motion translation. Comparatively, Runge-Kutta exhibits reduced truncation error:

- Euler Method: Local $O(h^2)$, Global $O(h)$,
- Runge-Kutta Method: Local $O(h^5)$, Global $O(h^4)$.

The higher order of Runge Kutta Truncation Errors leads to faster reduction in error rates, making it suitable for finer motions.

C. Data Augmentation

1) Depth Discontinuities: Previous approach used uses median filtering for noise reduction. Median filters, while robust against outliers, may cause block artifacts. We implemented a Gaussian Blur model for smoother transitions. Gaussian Blur, utilizing a weighted average of nearby pixels, smoothens uneven pixel values. However, Gaussian Blur may not preserve edges [16]. To address this, we employed a hybrid model combining Median Filtering for edge preservation and Gaussian Blur for noise reduction and smoothness.

2) RGB Inpainting: RGB inpainting is crucial for multi-frame generation. Current framework lacks smoothing and refinement steps, leading to visual artifacts. We employed a Bilateral Filter to address noise, edge sharpness, and blending issues. Additionally, we applied Gaussian Blur and residual sharpening to enhance RGB inpainting quality.

3) Depth Inpainting: Depth inpainting is essential for 3D video generation as depth impacts perceptual quality. Common methods, such as depth clipping, may lead to artifacts. We implemented consistent depth range clipping to ensure globally consistent values. Further, we refined Gaussian Blur parameters to balance noise reduction and edge preservation effectively.

IV. Results

A. Dataset

For this study, we utilized a dataset curated by Holynski et al., specifically designed for high-quality visual analysis. The dataset consists of videos, each with a duration of 60 frames, a consistent width of 1280 pixels, and variable heights to accommodate different aspect ratios. To preserve visual fidelity, the videos are encoded in the H.264 MP4 format with exceptionally high bitrates, effectively minimizing compression artifacts.

B. Implementation details

For our experiment, we used a system featuring a 13th Gen Intel(R) Core(TM) i7-13650HX 2.60 GHz processor, 16.0 GB of RAM, and a single RTX 4050 GPU operating on Ubuntu. The project required us to choose 15 videos from the initial dataset. We extracted only the first frame from each video to form the basis of our analysis. After implementing the process, we created two videos—one using the code provided by the original authors and the other from our implementation. These videos were then compared against the original ground truth, and the metrics obtained were analyzed to assess performance.

C. Quantitative Analysis

The results are shown in the table below:

TABLE I: Quantitative Metrics

Metric	Ours	Previous
PSNR	14.450	14.238
SSIM	0.205	0.186
LPIPS	0.409	0.457

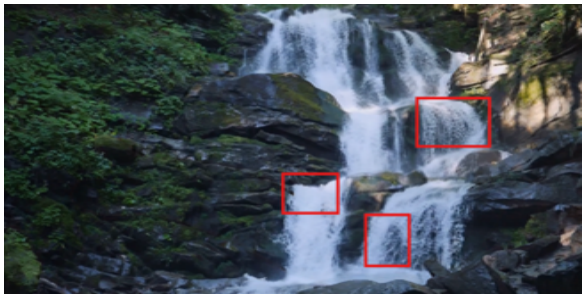
In this research, we assess our proposed methodology against the previous method developed by Xingyi Li et al. using quantitative metrics, as outlined in Table 1. The findings indicate that our approach exceeds the earlier method in fidelity, structural similarity, and perceptual quality. Enhancements in PSNR and SSIM demonstrate increased accuracy and structural integrity, while a reduced LPIPS score points to improved perceptual quality and visual realism. These results, backed by both quantitative and qualitative evaluations, affirm the effectiveness and robustness of our method in overcoming the shortcomings of existing techniques.

D. Qualitative Analysis

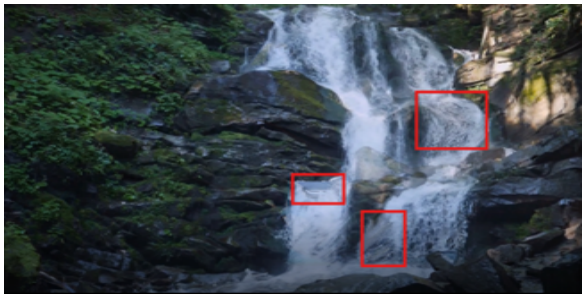
Figures 1, 2, and 3 present a comparison of the outputs generated by the earlier method from Xingyi Li et al. (shown on the left) versus our proposed approach (shown on the right).

In Figure 1, the image on the left exhibits various visual artifacts within the three outlined boxes. Notably, in the water regions, two sections of the video do not convincingly illustrate water flow, and the upper box indicates an unnaturally wavy flow. In contrast, the corresponding sections in our output on the right show precise rendering, offering a more authentic and natural depiction of water flow.

Likewise, in Figure 2, the left image uncovers additional limitations. Within the larger highlighted

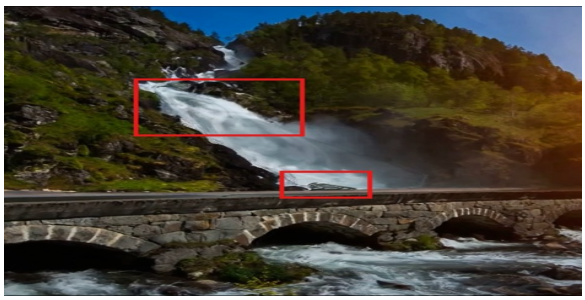


(a) Author's Results



(b) Our Results

Fig. 2



(a) Author's Results

(b) Our Results

Fig. 3

(a) Author's Results

(b) Our Results

Fig. 4

box, the water flow looks unrealistic, and the lower box displays inconsistencies in generation, failing to align smoothly with its environment. In comparison, our method on the right produces a more lifelike portrayal of water flow and effectively reconstructs the lower region of the water, maintaining harmony with the overall scene.

E. Ablation Study

This research thoroughly assesses how individual enhancements contribute to the original Eulerian flow-based motion estimation and depth inpainting system. We substituted the original Eulerian flow with a hybrid Runge-Kutta Eulerian flow approach, which enhances robustness when dealing with rapid movements and occlusions. To tackle depth discontinuities, we utilized a Gaussian filter in place of a median filter, leading to smoother transitions and less noise. For RGB inpainting, we implemented a bilateral filter to better maintain edge details and reduce artifacts in intricate textures. Lastly, we introduced a hybrid smoothing method with custom depth clipping ranges, providing increased flexibility and consistency in depth inpainting across various scene geometries. These changes collectively enhance both the performance and visual quality of the system.

V. Conclusion

In this study, we introduced an innovative technique for generating 3D cinemagraphs that significantly enhances current methods by tackling crucial issues like depth inaccuracies, motion field discrepancies, and visual artifacts.

By incorporating sophisticated depth estimation, a hybrid Runge-Kutta-Eulerian flow for creating motion fields, and improved inpainting methods, our approach results in smoother transitions, greater structural consistency, and enhanced perceptual realism.

Quantitative analyses show the superiority of our method, with marked improvements in PSNR, SSIM, and LPIPS metrics, indicating enhanced accuracy, structural cohesiveness, and visual appeal. Qualitative assessments further reinforce the robustness of our approach in depicting intricate scenes, such as fluid dynamics and smoke, with fewer artifacts and increased realism.

However, some challenges persist, including minor visual imperfections in certain textures. Future research will focus on refining these aspects and broadening the application of 3D cinemagraphs to encompass more dynamic and varied visual scenarios.

VI. References

[1] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7690–7699, 2020.

[2] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13653–13662, 2021.

[3] Chieh Hubert Lin, Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, and Ming-Hsuan Yang. InfinityGAN: Towards infinite-pixel image synthesis. In International Conference on Learning Representations (ICLR), 2022.

[4] Elizaveta Logacheva, Roman Suvorov, Oleg Khomenko, Anton Mashikhin, and Victor Lempitsky. DeepLandscape: Adversarial modeling of landscape videos. In Proceedings of the European Conference on Computer Vision (ECCV), pages 256–272. Springer, 2020.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[6] Siming Fan, Jingtan Piao, Chen Qian, Kwan-Yee Lin, and Hongsheng Li. Simulating fluids in real-world still images. arXiv preprint arXiv:2204.11335, 2022.

[7] 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 Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[8] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Singleview view synthesis in the wild with learned adaptive multiplane images. In ACM SIGGRAPH 2022 Conference Proceedings, pages 1–8, 2022.

[9] 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 (CVPR), pages 6498–6508, 2021.

[10] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 5865–5874, 2021.

[11] X. Li, Z. Cao, H. Sun, J. Zhang, K. Xian and G. Lin, "3D Cinemagraphy from a Single Image," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 4595-4605, doi: 10.1109/CVPR52729.2023.00446.

[12] J. Li, L. Cheng, Z. Wang, T. Mu, and J. He, "LoopGaussian: Creating 3D Cinemagraph with Multi-view Images via Eulerian Motion Field," arXiv.org, 2024.

[13] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," ACM Transactions on Graphics, vol. 31, no. 4, pp. 1–8, Jul. 2012.

[14] Gabi Luttwak and M. S. Cowler, "Advanced Eulerian Techniques for the Numerical Simulation of Impact and Penetration using AUTODYN-3D."

[15] A. Okeke, P. Tumba, O. Anorue, and A. Dauda, "Analysis and Comparative Study of Numerical Solutions of Initial Value Problems (IVP) in Ordinary Differential Equations (ODE) With Euler and Runge Kutta Methods," American Journal of Engineering Research, no. 8, pp. 40–53, 2019.

[16] E. Gedraite and M. Hadad, "Investigation on the effect of a Gaussian Blur in image filtering and segmentation,"

[17] Holynski, Aleksander & Curless, Brian & Seitz, Steven & Szeliski, Richard. (2020). Animating Pictures with Eulerian Motion Fields. 10.48550/arXiv.2011.15128.