3D Cinemagraphy from a Single Image
Supplementary Material

Xingyi Li[1,3] Zhiguo Cao[1] Huiqiang Sun[1] Jianming Zhang[2] Ke Xian[3*] Guosheng Lin[3]
[1]Key Laboratory of Image Processing and Intelligent Control, Ministry of Education
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Adobe Research [3]S-Lab, Nanyang Technological University
{xingyi_li,zgcao,shq1031}@hust.edu.cn, jianmzha@adobe.com, {ke.xian,gslin}@ntu.edu.sg
https://xingyi-li.github.io/3d-cinemagraphy

This document includes the following contents:

1. Runtime for inference.
2. Visual examples of the ablation study.
3. Details of the user study.
4. Limitation discussion and failure cases.

A. Runtime for inference

We test our runtime on an NVIDIA GeForce RTX 3090 GPU. Given a still image with a resolution of 1280 × 720 as input, our method takes about 2 minutes to synthesize a 60-frame video clip.

B. Ablation Study

Here we provide a qualitative example of the ablation study in Fig. B.1. One can observe: 1) using features instead of RGB colors in 3D scene representation improves the rendering quality and reduces artifacts; 2) introducing inpainting can produce plausible structures around depth discontinuities; 3) 3D symmetric animation allows us to feasibly fill in missing regions.

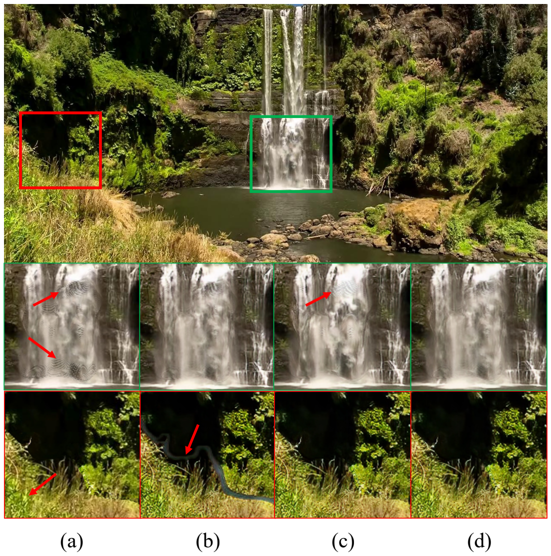
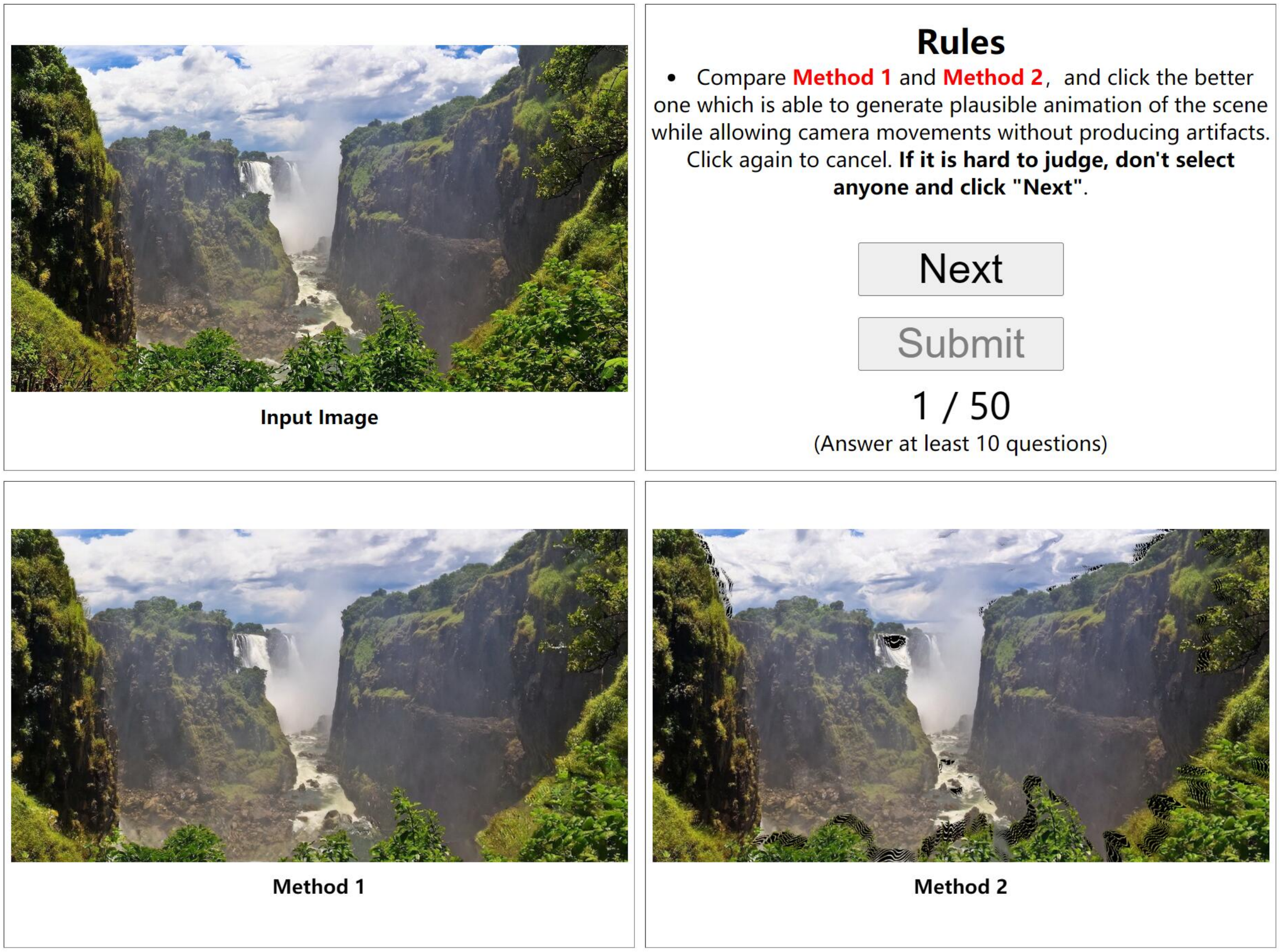


Figure B.1. **Visual examples of the ablation study.** (a) w/o features; (b) w/o inpainting; (c) w/o 3D symmetric animation; (d) Full model.

C. User Study

To investigate how our method performs in the view of humans, we conduct a user study on 50 photos from the test set of Holynski et al. [1] and the Internet. We build an online website for the user study. A screenshot of the website interface is shown in Fig. C.2. Our anonymous user study does not involve the collection of personally identifiable data. “Method 1” and “Method 2” exhibit the synthesized videos of two methods, where one is ours, and the other is the method randomly selected from 2D animation [1] → novel view synthesis [2], novel view synthesis [2] → 2D animation [1], novel view synthesis [2] → 2D animation [1] + moving average, naive point cloud animation, naive point cloud animation + 3D symmetric animation, 3D Photo [2], and Holynski et al. [1]. Note that the positions of the two methods are also random. Participants are required to choose the method that is able to generate plausible animation of the scene while allowing camera movements without producing artifacts, or none if it is hard to judge.

[*] Corresponding author.



- Compare Method 1 and Method 2, and click the better one which is able to generate plausible animation of the scene while allowing camera movements without producing artifacts. Click again to cancel. **If it is hard to judge, don't select anyone and click "Next".**

(Answer at least 10 questions)

Figure C.2. **Interface of the user study website.**

D. Failure Cases

Our approach inherits some limitations from single-shot novel view synthesis and single-image animation methods. Our method may not work well when the depth prediction module estimates erroneous geometry from the input image. For example, as shown in Fig. D.3, our method is usually prone to produce distortions on thin structures such as hands and light poles, when we move around the camera. Another limitation is that our motion estimator may sometimes fail to synthesize appropriate motion fields, e.g., some regions are mistakenly identified as frozen. This will lead to undesirable results. As a remedy, we have demonstrated the extensibility and flexibility of our framework: we can involve masks and flow hints as extra inputs to augment our motion estimator. This brings two advantages: (1) more accurate flow estimation; (2) interactive and controllable animation. Finally, as we take the first step towards 3D cinemagraphy, in this paper, we focus on handling common moving elements, i.e., fluids. In other words, our method may not apply to more complex motions, e.g., cyclic motion. We leave this for our future work.

References

[1] Aleksander Holynski, Brian L. Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5819, 2021. 1

[2] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

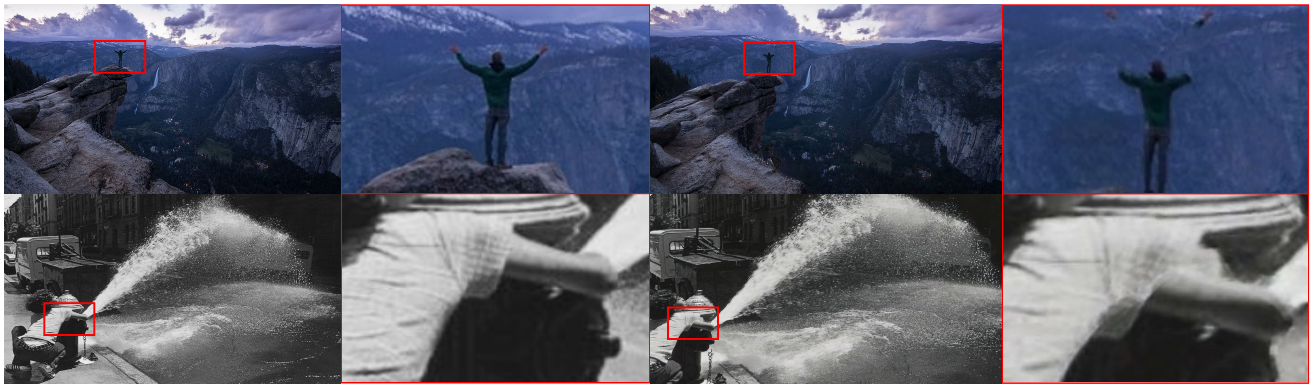


Figure D.3. **Failure cases.**