Supplementary Material of "Learning 3D Photography Videos via Self-supervised Diffusion on Single Images"

Xiaodong Wang[1*], **Chenfei Wu**[2*], **Shengming Yin**[2], **Minheng Ni**[2], **Jianfeng Wang**[3], **Linjie Li**[3], **Zhengyuan Yang**[3], **Fan Yang**[2], **Lijuan Wang**[3], **Zicheng Liu**[3], **Yuejian Fang**[1†], **Nan Duan**[2†]

[1]Peking University [2]Microsoft Research Asia [3]Microsoft Azure AI
{wangxiaodong21s@stu, fangyj@ss}.pku.edu.cn, {chewu, v-sheyin, t-mni, jianfw, Lindsey.Li, zhengyang, fanyang, lijuanw, zliu, nanduan}@microsoft.com,

Appendix

A.1 Metric Introduction

For Novel View Synthesis:
LPIPS: Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018] measures image similarity based on deep features extracted by a learned network, and it better coincides with human judgment.
PSNR: Peak Signal-to-Noise Ratio (PSNR) measures image similarity with the pixel-wise independence assumption.
SSIM: Structural Similarity Index Measure (SSIM) is a perception-based method and measures image similarity by considering structural information.
For Image Outpainting:
FID: Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] is used to measure the quality of the generated image.
IS: Inception Score (IS) [Salimans *et al.*, 2016] is a common metric to measure the diversity of the generated image.
CLIP-SIM: CLIP Similarity Score (CLIP-SIM) [Radford *et al.*, 2021] is used to measure the semantic consistency between the generated image and text.

A.2 Image filtering on MSCOCO

MSCOCO-2017 [Caesar *et al.*, 2018] contains 172 classes: 80 thing classes, 91 stuff classes, and 1 unlabeled class. We focus on thing classes and utilize pixel-level annotations for training. We regard the regions belonging to thing classes as the input object regions, and models are supposed to outpaint the remaining regions. For each image, we utilize the pixel-level annotations to construct the binary mask to indicate the known objects and unknown regions, and captions are regarded as the text prompts.

A.3 Efficiency Analysis

We compare the inference time for generating one frame in Table. 1. Our method shows significant improvement compared to 3d-photo and is comparable to other GAN-based methods. Tab. 1 illustrates a 20-step diffusion process. However, it is possible to improve the speed by sacrificing performance with fewer steps.

Method	Backbone	Time (second)
3d-photo [Shih *et al.*, 2020]	CNN	120
SynSin [Wiles *et al.*, 2020]	GAN	1.8
SMPI [Tucker and Snavely, 2020]	GAN	1.8
AdaMPI [Han *et al.*, 2022]	GAN	0.6
Ours	Diffusion	4

Table 1: Inference time for 1 frame on NVIDIA's V100 GPU.

A.5 More Samples of Out-animation

Figure. 1 shows more samples of our proposed out-animation task. Our method can handle a wide variety of scenes from the open domain.

A.5 More Samples on MSCOCO

Out-animated Samples. Figure. 2 presents the out-animated results using the processed input objects from MSCOCO.

Outpainted Samples. Figure. 3 presents the outpainted results using the processed input objects from MSCOCO.

References

[Caesar *et al.*, 2018] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018.

[Han *et al.*, 2022] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. *arXiv preprint arXiv:2205.11733*, 2022.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

*Equal contribution.
†Corresponding author.

Figure 1: Illustration of our proposed out-animation task. We first outpaint the input into complete and suitable scenes based on its content and prompts, and then generate subsequent frames with 3D effects to form the out-animated videos. Our method can handle a wide variety of scenes from the open domain, such as people, objects, advertised goods, paintings, etc. (*This input comes from an artwork of a young artist Geng Jiahao, in 2022 ANOBO "A Drop of Water with One World" exhibition. The number indicates the frame number in a 3D video.)

Figure 2: Out-animation sample results from MSCOCO. The 3D cycle effect follows 3d-photo [Shih *et al.*, 2020]. Our generated 3D videos show good consistency and have realistic details.

Figure 3: Outpainted sample results from MSCOCO. From top to bottom: text prompt, input objects, ground truth, our prediction. Our method can outpaint the input objects to realistic and diverse images which show good text relevance.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[Shih *et al.*, 2020] 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*, pages 8028–8038, 2020.

[Tucker and Snavely, 2020] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020.

[Wiles *et al.*, 2020] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.