2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN) | 979-8-3503-3538-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICUFN57995.2023.10201216

Study on the Vulnerability of Video Retargeting Method for Generated Videos by Deep Learning Model

Aro Kim, Dong-hwi Kim, Sang-hyo Park
Kyungpook National University
School of Computer Science and Engineering
Daegu, South Korea
arokim37@knu.ac.kr, dhwi@knu.ac.kr, s.park@knu.ac.kr

***Abstract*—Text-to-video generation is getting attention and the generated videos can be used in many applications. However, it is uncertain whether existing deep learning techniques work well for generated videos. In this paper, we compose a study of how generated videos can be retargeted by deep learning models with the ratio of the main object preserved and looked for ways to improve the quality of the generated and retargeted video frames. Throughout the experiment, we discover the errors of video retargeting on the generated videos in the processes of segmentation, inpainting, and relocating.**

I. Introduction

When displaying a video on another device, adjusting the original screen ratio of a video to the other screen ratio well, called video retargeting, is necessary. The aspect ratio stands for the image ratio of its width to its height and can be expressed as 4:3, 16:9, and so on. When the aspect ratio value of a new screen increases and videos are displayed with the original ratio of themselves, the screen on both sides could be totally empty, which means we cannot use the advantage of wide-screened electronic devices. In addition, to match the increased aspect ratio of the screen, the object of the generated text-to-video needs to be overstretched, which harms the quality of experience (QoE). Thus, video retargeting research should be investigated thoroughly without harming QoE.

Furthermore, video retargeting [1], [2], [3] is more challenging when the source video is computer-generated video rather than natural video which can be easily obtained and utilized to train deep learning-based model. Particularly, generated video by text-to-video model, which has been in the spotlight and has become a big sensation recently [4], [5], [6] may increase the difficulty of maintaining QoE in video retargeting, which has not yet been extensively studied. As with generated images today [7],[8], generated video will have many more applications. It can be used as a dataset for other models and for content creation on various video streaming platforms. When retargeting such video, we found that it is also difficult to maintain the original size ratio of the objects in generated video.

To discover QoE of video retargeting on such new video dataset, in this paper we conducted a study of retargeting generated videos. In the study, a deep learning-based framework [9] has been applied to generated video sets obtained from demo videos of Make-A-Video [10]. In the generated videos, first, it detects the main objects, segments them, and detaches them from every single frame of the videos. Then, the framework includes inpainting the location where the objects are detached. After all, the inpainted frames are adjusted and resized to the new aspect ratio of the screen and segmented objects are relocated [9] to the frames. We conducted a study for the first time of retargeting generated videos by segmenting and detaching the main objects and relocating them. Throughout the experiment, we found several errors and are reporting them for the future work.

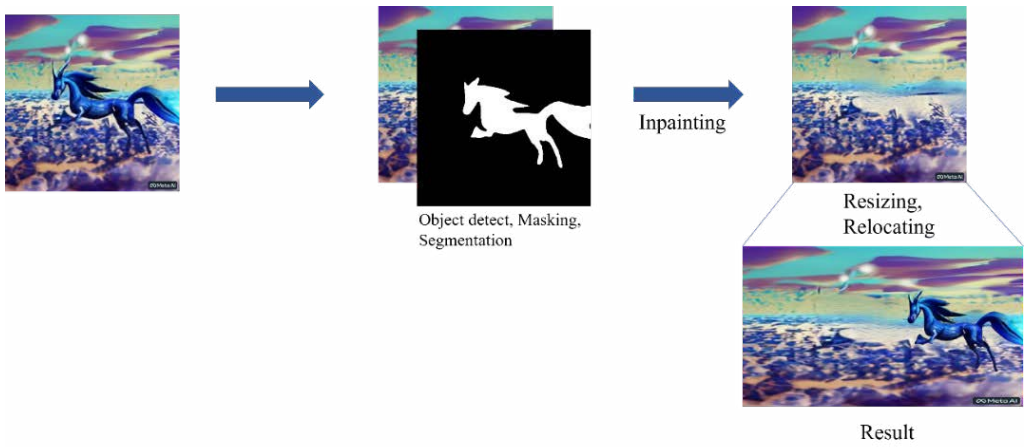

Fig. 1. Framework of the retargeting method of a generated video frame.

Fig. 2. Resized frames by hand compared to Relocated and Retargeted frames.

II. Dataset and Framework

A. Dataset

We downloaded text-to-video generation demo videos from Make-A-Video [10]. Fig. 1 shows the video frames we got online and used. All the videos were 512 × 512 resolution at 14 fps. One example of video was made upon given text, which is "A dog wearing a Superhero outfit with a red cape flying through the sky." In the video, A dog flying in the sky turns in its body and gets closer. One video has 3 seconds long and all other videos have 5 seconds. Overall, the frames showed low resolution.

B. Framework

Fig. 1 shows a framework of the retargeting method of generated videos. Our framework consists of three main processes: 1) detaching objects [11], 2) inpainting [12], 3) resizing and relocating that is adopted from [9]. We applied the retargeting method to generated videos. First, from a frame of generated video with its own aspect ratio, objects are detected by the segmentation method applied from [6]. If the object occupies the largest proportion of the frame, it was considered the main object. In addition, objects with the main object should be detected and detached together with the main object. We put them in the subclass to be detached altogether. After segmentation, the coordinate values of where the objects were located were stored and masked frames were made for inpainting. Second, for inpainting, we applied for work from [12]. The part where the detached objects were located was inpainted with the information of masked frames. Finally, the inpainted background frames were resized to the target aspect ratio. In this study, they were resized to the ratio of 16:9, width to height. After the background being resized, the detached objects were relocated.

III. Results

A. Comparison of increased frames by hand and retargeted frames by deep learning method

Fig. 2 shows the comparison of video frames that we stretched out the image size from 1:1 to 16:9 by hand and the final result of masking, inpainting, retargeting and relocating. When we applied the retargeting method [9], the objects are relocated while maintaining their original ratio of themselves. Even though the backgrounds are stretched, the objects are not overstretched since we segmented and relocated them.

B. Segmentation Error

Fig. 3 shows the results and errors for each process of the video frame. Fig. 3a indicates a case when the main object of the frame was not completely segmented. Only a portion of the main object of each frame was segmented. In Fig. 3a, one out of four people is not segmented and only three people can be seen in the masked frame. Fig. 3b shows a case when sub-objects are not segmented. In Fig. 3b, the red cape of a dog is not segmented, in turn, we cannot see the shape of the cape in the masked frame. It appears in the inpainted frame that is not intended to be there.

C. Inpainting Error

Fig. 3c and d show the case of inpainting errors. The inpainted frame of Fig. 3c shows some afterimages of the main objects. In Fig. 3d, even though the background is inpainted with the nearby color, it does not show the shape of trees of forests with a blurred image.

D. Relocating Error

Fig. 3e shows the case of relocating error. In the final frame, there are two main objects which are dogs together in one frame. Even though in the process of masking and detaching, the main object was detached successfully as we

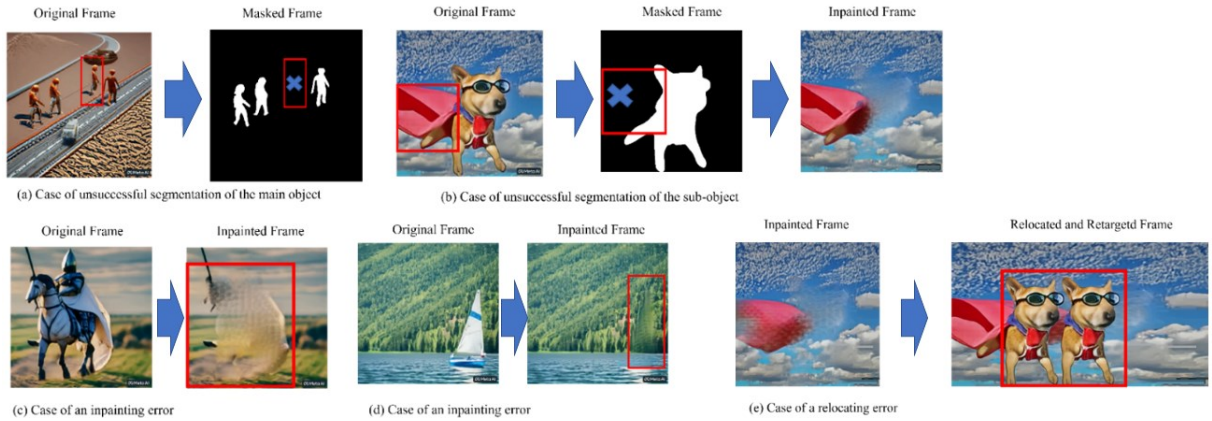

Fig. 3. Results using semantic segmentation and impainting methods for text-to-video generation data.

can see in the inpainted frame, in the relocated and retargeted frame, there exist two of them.

IV. Conclusion

The more text-to-video generation models come out, the more videos can be generated. They are being displayed on various electronic devices since they have versatile uses. In this paper, we applied a retargeting method to generated videos from make-a-video. Among the video frames, we could notice a white and thick border line between the object and the background. In addition, one of the objects had a blurred boundary between the background and the object. In the future study, we expect to improve the quality of the frames by using deep learning-based SR methods (such as [13]) rather than simple resizing. The results of applying the retargeted method to generated video frames show that the ratio of objects is preserved well when the background is stretched. By improving the inpainting method, we look forward to getting a higher-quality output without afterimages. Furthermore, in the future study, we need to find clues about several output images having two main objects even though all the stages conducted in the retargeting method are done successfully in this study.

Acknowledgment

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00167169, Development of Moving Robot-based Immersive Video Acquisition and Processing System in Metaverse) and in part by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4199990214394).

References

[1] Qi, M., Qin, J., Zhen, X., Huang, D., Yang, Y., & Luo, J. (2020, October). Few-shot ensemble learning for video classification with SlowFast memory networks. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 3007-3015).

[2] Rubinstein, M., Shamir, A., & Avidan, S. (2008). Improved seam carving for video retargeting. *ACM transactions on graphics (TOG)*, *27*(3), 1-9.

[3] Bansal, A., Ma, S., Ramanan, D., & Sheikh, Y. (2018). Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 119-135).

[4] Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. (2022). Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

[5] Esser, P., Chiu, J., Atighehchian, P., Granskog, J., & Germanidis, A. (2023). Structure and Content-Guided Video Synthesis with Diffusion Models. *arXiv preprint arXiv:2302.03011*.

[6] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi, H. (2023). Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. arXiv preprint arXiv:2303.13439.

[7] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In International Conference on Machine Learning (pp. 8821-8831). PMLR.

[8] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695).

[9] Jin, J. G., Bae, J., Baek, H. G., & Park, S. H. (2023). Object-Ratio-Preserving Video Retargeting Framework Based on Segmentation and Inpainting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 497-503).

[10] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., ... & Taigman, Y. (2022). Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792.

[11] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1290-1299).

[12] Li, Z., Lu, C. Z., Qin, J., Guo, C. L., & Cheng, M. M. (2022). Towards an end-to-end framework for flow-guided video inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17562-17571).

[13] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European conference on computer vision (ECCV) workshops (pp. 0-0).