Object-Ratio-Preserving Video Retargeting Framework based on Segmentation and Inpainting

Jun-gyu Jin† Jaehyun Bae† Han-gyul Baek Sang-hyo Park*
Kyungpook National University
Daegu, South Korea
eoqkr1025@knu.ac.kr jaybae@knu.ac.kr qorgksruf123@knu.ac.kr s.park@knu.ac.kr*

Abstract

The recent development of video-based content platforms led the easy access to videos decades ago. However, some past videos have a old screen ratio. If an image with this ratio is executed on a display with a wider screen ratio, the image is excessively stretched horizontally or creates a black box, which prevents efficient viewing of content. In this paper, we propose a method for retargeting the old ratio video frames to a wider ratio while maintaining the original ratio of important objects in content using deep learning-based semantic segmentation and inpainting techniques. Our research shows that proposed method can make a retargeted frames visually natural.

1. Introduction

Video retargeting is increasing in demand along the progression of the digital devices, displays and video-based content platforms, for instance, smart phones, in broadcasting and streaming services. Using those platforms, we can easily access to the videos from decades ago and also stream those videos on the latest displays. Herein, an optimization problem occurs that is the screen ratio of video frames between decades ago and recent frames could not the same, for instance, when the video has 4:3 of screen ratio and played on the 16:9 display, the video contents may have a black box on both side or the contents would be overstretched by forcing the video to fit to the display. Solving this problem is challenging using only super-resolution technology because most super-resolution techniques estimate both width and height pixels.

To handle the overstretched problem, many research were conducted on both image and video [2,8,12,15]. However, the initial works lean to the seam carving algorithm, which is based on a mathematical calculation, thus it may have a limitation with handling large object or complex images. The progression of deep-learning also led the retargeting tasks. Deep learning-based methods [4,6] are adopted to retargeting tasks to make more natural retargeted result. Although the performance of retargeting were good, the performance of [6] can highly depend on the accuracy of neural network-based object detection and [4] can process only image.

We propose a video retargeting method using deep learning-based instance segmentation and video inpainting. Our method focused to change the screen ratio from old ratio to wider ratio while preserving the ratio of important objects. We first segment the important objects and detach from each frame, and inpaint the missing location. After inpainting, we resize the frames to the target aspect ratio then relocate the objects. Our research shows that combining a state-of-the-arts (SOTA) of segmentation and inpainting deep learning models could solve the video retargeting tasks with a satisfactory results.

2. Related Work

2.1. Image and Video Retargeting

Most image retargeting research used seam carving [2] method. Seam carving method uses energy function to measure the importance of parts in image then expand the image to target aspect except the important region. However, applying seam carving to condensed image is hard because an appropriate seam could not be made on it. Setlur et al. [12] proposed a novel image retargeting method using segmentation with image inpainting method. Cho et al. [4] proposed a weakly and self-supervised deep convolutional neural network that to preserve the important content as much as possible using a shift map on image. This method can learn the important part implicitly by itself.

For video retargeting, Lee et al. [6] solve the video retargeting problem using deep learning-based object detection as a pre-process network. Cho et al. [1] proposed video retargeting model using recurrent neural network that could make many candidates of retargeted frames. However, the aforementioned works could lose some information of frames or

hard to preserve the quality of the object after retargeting the frame, also some method highly rely on pre-processing network. There are several video quality assessment (VQA) methods [10, 11] that use both the generated image and reference. In addition, they mainly deal with the videos which are suffered from the compressions or streaming distortions. However, our proposed framework does not have a reference frames because we resize the background and preserve the ratio of the object that can change the context of the original frame. Thus, these VQA methods could be hard to directly applied to our framework.

2.2. Segmentation and Inpainting

Segmentation and inpainting are closely related. To inpaint the image or video, masking is an essential that can guide an inpainting network about the location to be inpainted. These combined forms of research have recently been increasing. Song et al. [13] proposed a unified image segmentation and inpainting model that can predict segmentation labels. Katircioglu et al. [5] proposed image-based self-supervised segmentation method to overcome the limitation of supervised detection network, which is specialized in detecting a person. Many papers adopt segmentation network to solve the inpainting problem.

Inpainting tasks are also improved along the progression of segmentation tasks. Liu et al. [9] proposed a partial convolutional layer for the problem that the model generates irregular images due to the initial hole problem. Yu et al. [16] proposed an encoder-decoder architecture consisting of gated convolutions that can train dynamic feature selection mechanisms for generating natural images. Zeng et al. [17] proposed AOT-GAN, a GAN-based model that aggregates both information and patterns to generate high-resolution images. Xu et al. [14] proposed to fill in the missing region through forward-backward propagation of flow-field by constructing a deep flow completion network (DFC-Net) model. Li et al. [7] proposed an end-to-end framework by propagating features of local neighboring frames and non-local reference frames rather than pixel propagation. One thing that is unfortunate is that the performance of inpainting nerwork is highly depend on the segmentation network because the inaccurate mask really ruins the inpainting process, however, the performance of segmentation network is enhancing, thus there is still room for improvement. Furthermore, in our best knowledge, no studies have applied a framework that combines segmentation and inpainting models applied to video domain to inpaint a masked background.

3. Proposed Framework

In this section, we show how our framework is processed from beginning to end with detailed descriptions.

Fig. 1 shows our proposed framework. Our framework can be divided to three main processes: 1) segmentation, 2) inpainting, and 3) resizing and relocating. First, old ratio of frame is feed to instance segmentation network to detect the important objects. After detecting the objects, detach the objects from input frame and make a binary masked frame for inpainting. Second, masked frame and input frame are feed to inpainting network to estimate the masked missing area. Note that we are processing video sequence, but we can also use image inpainting networks. Third, resize the inpainted frame to the target aspect ratio by stretching horizontally. Then, relocate the detached objects to appropriate location. Details of each process will be covered in subsequent subsections.

3.1. Segmentation

In the stage, we applied the work from [3], which is the SOTA model in instance segmentation tasks. The objects in the frame are detected by instance segmentation method to each frame. Using instance segmentation-based method, we could classify the class of the objects, and also the objects within the class themselves.

In this framework, our important object is person. Thus, our main objective is to segment the person in each frame. However, a person of significantly small size is not segmented. Because in the overall context, a significantly small person is less important. Thus, we included only when the proportion of person in the frame is larger than 1%, and this value was obtained by empirical experiments. Moreover, depending on the nature of the video clips, an important object can be added (i.e., person carry the bag). Therefore, we added a subclass object that can be carried by a person, for instance, bag and guitar. After segment the objects, we make a binary masked frames that can be used in inpainting stage, and also detach the important objects to prevent the objects from overstretched problem.

In addition, masking is one of the important factor that can directly affect to the whole context and result. Because most of the inpainting method uses the information of the neighboring local context and also the neighboring frames, masking should be process perfectly. Thus, to prevent the fault of missing pixel in segmentation network, we give an extra margin (i.e. morphology dilation to the mask 3 times using 3x3 cross filter) to the mask that can increase the performance of the inpainting network. Fig. 2 shows the comparison of the masked frame.

3.2. Inpainting

For inpainting, we adopted the work from [7], the SOTA model in this tasks. The input frames of inpainting are the video clip and binary masked frame. Here, we can use both of image and video-based inpainting model, but the difference between two models is that the image cannot use the spatio-temporal informations between the frames, thus

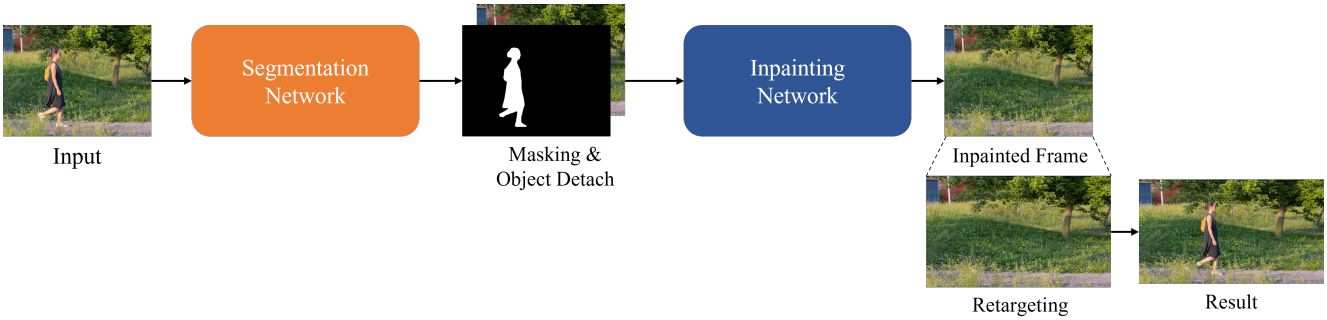


Figure 1: Framework of the proposed video retargeting method.

(a) Original input frame

(b) Masked frame

(c) Masked frame with margin

Figure 2: Comparison of masking

its inpainting performance is hard to outperform the video-based model. By feeding masks and frames into the inpainting model, an object-removed background frame is generated.

3.3. Resizing and Relocating

The object-removed background image is resized by increasing the width to a ratio of 16:9 to the height. We used bicubic interpolation for resize using *openCV*. The converting function can be show as

$$W_{new} = \frac{16}{9} * H, \tag{1}$$

where W_{new}, H denote a width after resizing is processed and the height of the original frame, respectively. In this stage, we can also replace using openCV to deep learning-based super-resolution models. For the versatility, the fractions in Eq. 1 depends on the target aspect ratio.

After resize the background of frame, we should place the detached objects. Since the background is stretched, but the size of the object retains its original size, thus, it is necessary to find the optimal location to create a smooth context when relocating the objects. We place the object that was located on the edge is equally relocated on the edge, and the object located in the center should be add a half of the increased width length to the original coordinates. The object relocating algorithm is explained in Algorithm. 1. This can hide the inpainted blurred location with the important objects.

Algorithm 1 Object Relocation Algorithm

if $object$ is in left edge **then**
    $offset \leftarrow 0$
else if $object$ is in right edge **then**
    $offset \leftarrow W_{new} - W$
else
    $offset \leftarrow (W_{new} - W)/2$
end if
for $x, y \in object$ **do**
    $x_{new}, y_{new} \leftarrow x + offset, y$
end for

4. Experiment and Results

4.1. Dataset and Environment

The dataset used in this study is an video clip with no more than three human objects in the DAVIS 2017 trainval dataset. This dataset is a collection of video clips of 854 x 480p sizes, which have a wider screen ratio, but we preprocess the dataset as a 4:3 old screen ratio by cropping the center of frames to size of 640 x 480p. Thus, our objective is to retarget to the screen ratio of 854 x 480p. Our framework was tested on NVIDIA RTX 3060 with CUDA 11.3 and CUDNN 8.2.1. Using pytorch 1.9.0 version with python 3.8.8. In a video inpainting experiment using the E^{2}FGVI model, the experiment was performed by removing the far-

thest reference frame because of the limited resources. Our code is available at https://github.com/realJun9u/ORPVR.

4.2. Results

In this subsection, we compare between our framework, seam carving and just horizontally stretching method (resizing). Fig. 3 shows the retargeted result from 4:3 of screen ratio to 16:9. We can easily see that our framework generates more natural frame by preserving the original ratio of the person compared to resizing and seam carving method, which seem overstretched. Seam carving was one of the most popular methods that can easily retarget images, but when it comes to preserving object ratios, it is not possible to properly preserve the ratio of the objects. For instance, in Fig. 3c, seam carving does not preserve the object, and also the pixles of tree were crushed. However, in Fig. 3l, we can see the subset of backpack in front of the person, which came from the imperfect segmentation. Note that when the segmentation work is well preceded, the remaining work can be carried out easily.

5. Ablation Study

5.1. Segmentation Performance

Table 1 shows the performance of segmentation network evaluated by intersection over union (IoU) that predicted the person from the clips on the DAVIS dataset what we have used. Although it has a high value, it is hard to say that it has been precisely predicted (such as backpack straps, human shoes). To compensate the error, we added a margin to the mask before we feed it into inpainting network.

Clip	IoU(person)
breakdance-flare	0.9
crossing	0.92
hike	0.94
lucia	0.88
judo	0.8
parkour	0.93
schoolgirls	0.92
average	0.9

Table 1: IoU results on clips from DAVIS Dataset.

5.2. Inpainting Method

We attempted image inpainting for a single frame and video inpainting using consecutive frames to remove objects from the image. In this subsection, we compared the performance of inpainting which method can generate more real and smooth results that fits to our framework. To solve the inpainting problem with a single image-based method, we applied the work from Zeng *et al.* [17] named AOT-GAN, and two video-based method, DFCNet [14], and E^{2}FGVI [7].

On a single image, the performance is highly compliant, but it can be seen that when AOT-GAN is applied to the video, resulting in poor performance. Fig. 4 compares the inpainting result between image method and video method. In Fig. 4b and 4f, it is hard to find the front of the truck and tire, but in 4d and 4h, it almost perfectly estimate the hidden area which was hidden by the person. We can see that video inpainting method shows more recover the details compared to the image-based method. Because our framework is to convert from old screen ratio to wider ratio in video, video inpainting seems more appropriate.

In Fig. 4c and 4g, DFCNet was not able to handle the spatio-temporal information naturally—objects seem not removed, and the front wheel of truck was seen as two. Here also shows that still the frames are degraded, but estimated the missing area better than image-based method. On the other hand, in 4d and 4h, E^{2}FGVI handled object removal and spatio-temporal domain naturally. Therefore, we can confirm that using E^{2}FGVI would be best to our framework to generate more natural inpainted results.

5.3. Optimal Relocation Method

We tried several methods to determine the proper location when relocate the objects to the object-removed retargeted background frame.

1. **Original coordinates.** Use the coordinates which is identical to the original frame of the objects. In this case, the objects that were located on the right edge is located in an unnatural position.

2. **Offset coordinates.** Use the coordinates by adding offset to original coordinates, and offset is the increase in width. In this case, the object that was located on the left edge is located in an unnatural position. Offset is determined in proportion to the width increased during the Resize Background process.

3. **Dynamic coordinates.** Use the original coordinates if the object was located on the left edge, and use offset coordinates if the object was located on the right edge. If the object is not on any edge, use the coordinates that adds half of the increase in width to the original coordinates. Please refer the Algorithm 1.

Fig. 5 shows the result of diverse relocation method. When consider the object with shadow, the original coordinates method and Offset coordinates seems unnatural. However, dynamic coordinates method almost preserve the context of the original input frame. Therefore, we adopt dynamic coordinates method in our framework.

(a) lucia, 21st (b) Bicubic interpolation (c) Seam carivng (d) Ours

(e) breakdance-flare, 34th (f) Bicubic interpolation (g) Seam carivng (h) Ours

(i) hike, 38th (j) Bicubic interpolation (k) Seam carivng (l) Ours

(m) parkour, 29th (n) Bicubic interpolation (o) Seam carivng (p) Ours

Figure 3: Retargeting result from our framework

6. Discussion

From our quantitative experiments, our retargeting framework can change the old ratio to wider ratio while preserve the original ratio of interest objects. Based on segmentation and inpainting, our framework really fits to the video that the object moves near the center of the frame, and the object does not go out from the frame. However, when the part of the object is lose in the frame or the object moves near the edge of frame, our framework cannot relocate accurately because to relocate the object appropriately, we need all parts of the object (i.e., from head to feet), but the objects often locate on the edge of the frame, thus the information of the object is already lose. To solve this problem, other method should be added.

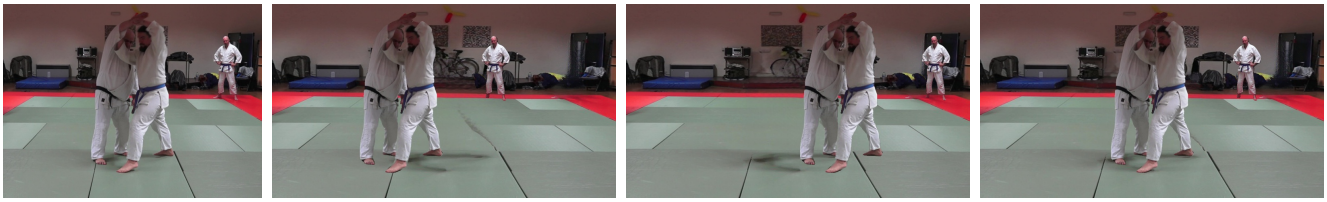
7. Conclusion

Due to the development of electronic devices, many electronic products with various screen ratios have emerged, thus the demand for video retargeting is increasing. However, research on converting the screen ratio while maintaining the original ratio of the object still insufficient. Moreover, a unified segmentation and inpainting framework in video has not been fully investigated. In this paper, we proposed a novel unified framework that can retarget the old video screen ratio to the wider target aspect ratio horizontally while preserving the quality of the objects using segmentation and inpainting networks. Our results show that the quality of retargeted frame seems more natural than just stretching the frame horizontally. For future work, further research on how to relocate the objects is needed. If the object is located at the edge of the frame, and the information on a part of the body is lost by repeatedly entering and

(a) Original 22nd frame (b) AOT-GAN model (c) DFCNet model (d) E^{2}FGVI model

(e) Original 23rd frame (f) AOT-GAN model (g) DFCNet model (h) E^{2}FGVI model

Figure 4: Comparison of inpainting method on Crossing from DAVIS.



(a) Original input frame (b) Original coordinates (c) Offset coordinates (d) Dynamic coordinates

(e) Original input frame (f) Original coordinates (g) Offset coordinates (h) Dynamic coordinates

(i) Original input frame (j) Original coordinates (k) Offset coordinates (l) Dynamic coordinates

Figure 5: Object relocation results of each methods

leaving the screen, retargeting may be unnatural.

Acknowledgment

†: These authors equally contributed to this paper.
This study was supported in part by Basic Science Re-

search Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1I1A3072227) 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] Video retargeting: Trade-off between content preservation and spatio-temporal consistency. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 882–889, 2019.

[2] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. In *ACM SIGGRAPH 2007 papers*, pages 10–es. 2007.

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[4] Donghyeon Cho, Jinsun Park, Tae-Hyun Oh, Yu-Wing Tai, and In So Kweon. Weakly-and self-supervised learning for content-aware deep image retargeting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4558–4567, 2017.

[5] Isinsu Katircioglu, Helge Rhodin, Victor Constantin, Jörg Spörri, Mathieu Salzmann, and Pascal Fua. Self-supervised segmentation via background inpainting. *arXiv preprint arXiv:2011.05626*, 2020.

[6] Seung Joon Lee, Siyeong Lee, Sung In Cho, and Suk-Ju Kang. Object detection-based video retargeting with spatial–temporal consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4434–4439, 2020.

[7] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17562–17571, 2022.

[8] Feng Liu and Michael Gleicher. Video retargeting: automating pan and scan. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 241–250, 2006.

[9] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 82–100, 2018.

[10] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. St-greed: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Transactions on Image Processing*, 30:7446–7457, 2021.

[11] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441, 2010.

[12] Vidya Setlur, Saeko Takagi, Ramesh Raskar, Michael Gleicher, and Bruce Gooch. Automatic image retargeting. In *Proceedings of the 4th international conference on Mobile and ubiquitous multimedia*, pages 59–68, 2005.

[13] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. *arXiv preprint arXiv:1805.03356*, 2018.

[14] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2019.

[15] Bo Yan, Kairan Sun, and Liu Liu. Matching-area-based seam carving for video retargeting. *IEEE Transactions on circuits and systems for video technology*, 23(2):302–310, 2012.

[16] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019.

[17] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 2022.