Complete and temporally consistent video outpainting

Loïc Dehan, Wiebe Van Ranst, Patrick Vandewalle, and Toon Goedemé
EAVISE-PSI-ESAT, KU Leuven
Sint-Katelijne-Waver, Belgium
firstname.lastname@kuleuven.be

Abstract

We describe a novel method for video outpainting. The goal of outpainting is to fill in missing regions at the edges of video frames. Our focus lies on converting portrait (9:16) to landscape (16:9) video. In contrast, most video completion research is focused on inpainting: filling a masked section within the frame based on the remaining, known pixels.

Our proposed method consists of three main aspects: (1) We form a background estimation using video object segmentation and video inpainting methods, (2) we use optical flow to form temporal consistency, and (3) we propose image shifting to improve individual frame completions. Our method is able to successfully broaden the aspect ratio of a video. On most videos, we achieve realistic results. Only on videos with complex camera motion and foreground objects leaving the frame, the quality is less.

In contrast to other state-of-the-art methods, our method is able to reconstruct the full frame, including unseen image parts. Moreover, it is temporally consistent. We evaluate our method on the DAVIS and YouTube-VOS datasets. The code is publicly available[1].

Keywords*: video completion, video outpainting, background estimation, optical flow, image outpainting*

Figure 1. A frame from the DAVIS dataset [22] completed with the standard black bars, the video inpainting method of Gao *et al*. [7] and our method.

1. Introduction

As the popularity of mobile apps such as TikTok and Instagram increases, so does the amount of vertical video content. When this content is displayed on e.g. a television, the aspect ratio has to be changed to fit the screen. To do this, the regions around the video content have to be completed. Currently, these completions are either left blank, resulting in black bars around the video (seen in figure 1, top), or filled in with a blurred repetition of the original video. Our goal is to create more realistic completions around the given video content to improve the viewer's experience without diverting attention from the original video.

Video inpainting is closely related to video outpainting. Inpainting refers to the task of completing a marked region within a frame. When comparing inpainting to outpainting, the following additional complexities can be observed. First of all, in image inpainting, the missing region is somewhere within the given image. This provides omnidirectional information to an image inpainting network. In image outpainting, the missing region is at the side or around the image for which there are only known pixels on one side of the missing region. Secondly, when outpainting an image from portrait to landscape, the completion is twice as large as the input, whereas, in image inpainting, the missing region usually is a relatively small part of the image. Finally,

[1] https://github.com/Video-Outpainting/VideoOutpainting

due to the two previous points there is also a greater distance between the input and output pixels.

The contribution of our work is a full video outpainting method that provides visually pleasing, temporally consistent completions without damaging the original video content or introducing artifacts to the outpainted regions. The proposed image shifting technique allows to create realistic results even on larger outpainted regions. We base our method on an existing flow-based video inpainting method, from Gao *et al*. [7], in which optical flow is used to pass on information between neighboring frames to form temporal consistency and extend it to outpainting. Firstly, we reduce temporal artifacts by initially forming a background estimation before completing the frames. Secondly, we shift image contents to improve individual frame completions.

In the following sections, we will first discuss the related fields of image completion, video inpainting, and video outpainting. Next, we describe the datasets we used in section 3. In section 4 we describe our method for video outpainting. First, we look at the shortcomings of video inpainting when applied to outpainting and, next, we explain our improvements on image outpainting using image shifting and our video outpainting method. In section 5 we evaluate our method using several common image and video evaluation metrics. Next, we discuss the limitations of our method in sections 6. Finally, in section 7 we discuss our results.

2. Related Work

2.1. Image Completion

Image inpainting predates the use of deep learning techniques. Traditional approaches can be divided into two groups: diffusion-based and patch-based. Diffusion-based methods fill the gap by smoothly spreading the image content on the edges over the masked area. [20]. Patch-based methods [3, 23] fill in missing regions with patches from source images that maximize patch similarity. Both methods can complete smaller regions, but fail to realistically complete larger, more complex parts of an image.

More recent deep learning techniques can more realistically complete the masked regions. These methods are based around the advent of Generative Adversarial Networks (GAN) [8]. Pathak *et al*. [21] introduce an adversarial loss in addition to reconstruction loss to address that inpainting is multimodal. Iizuka *et al*. [10] formed improvements by introducing both global and local discriminators for deriving the adversarial losses. More recently, Yu *et al*. [31] presented a contextual attention mechanism in a generative inpainting framework, which further improves the inpainting quality. Nazeri *et al*. [18] observe that the structure of an image is represented in the edge map. They achieve photorealistic results by first completing the edges before completing the actual pixels. Zhao *et al*. [33] use aspects from style transfer research [6, 9] to introduce co-modulated GANs. Their method can generate realistic results on larger mask sizes. The method of Yang *et al*. [30] is limited to natural panoramas and only produces very low resolution results.

2.2. Video Inpainting

With moving from image inpainting to video inpainting comes the challenge of temporal consistency. The result must be consistent through time. Small variations between two consecutive frames may be evaluated equally as individual images. However, when the frames are replayed as video, the inconsistencies are noticeable to the human eye. In addition, the method must continue to work with complex movement of objects, movement of the camera itself, and variation in the background. Finally, a video provides more information about the scene than a single image.

Patch-based methods [13, 26] attempt to maintain temporal consistency by using segments of adjacent frames to form the completions. These methods cannot generate new image content, and reusing the content of other frames is not sufficient to complement the frames consistently and realistically.

Chang *et al*. [4] use 3D-gated convolution to create temporal consistency. This allows multiple frames to be used as input for each frame's completion. Deep learning-based methods are less effective here compared to image in/outpainting because of the high memory requirements. Each frame of the video contains as much information as the input in image inpainting. The amount of frames that can be used as input for the completion of a single frame is limited.

Finally, flow-based methods [7, 11, 28, 36] use optical flow. The per-pixel motion between frames can be used to propagate pixel values into the masked regions. The optical flow can be estimated using flow estimation networks [5, 24]. Gao *et al*. [7] propose to initially completing the edges within the optical flow. This extra step creates sharper results near the edges of moving objects. Very recently, Liu *et al*. [14] proposed a video inpainting method based on transformers.

2.3. Video Outpainting

There is comparatively less research focused on video outpainting itself. Earlier work in the video domain is the technique of Avraham and Schechner [2], who suggested a foveated method for video extrapolation. The resolution of their resulting video diminishes as the distance from the original content increases, which is similar to the behavior of the human fovea. The method proposed by Aides *et al*. [1] improved the visual details and general structure of such a peripheral scene.

Figure 2. Failures in the video outpainting method of Lee et al. [12]: black corners are not inpainted, and foreground object causing artifacts outside original frame.

Also, some other methods can be applied to video outpainting. Video stabilization methods [15, 29] are required to extrapolate a relatively small region outside the frame. Also, Gao et al. [7] have applied their video inpainting method to video outpainting, which we discuss in section 4.1.

Other methods in literature cover certain similar aspects of the problem but do not provide a complete and consistent result. The method from Lee *et al*. [12] warps and blends adjacent frames allowing them to complete parts of the outpainted region based on observed pixels. Any unobserved region of the scene is ignored. Their technique can exclusively form realistic completions based on information available in other frames. Regions that were never visible are left blank as illustrated in figure 2. In contrast, our method hallucinates video content in regions without visual evidence. This enables us to outpaint video from static cameras.

The work of Liu et al. [15] and Maggia et al. [17] can form full completions but creates temporal artifacts in the regions that were never visible to the camera because it does not impose temporal consistency like our technique does.

3. Dataset

Part of the completions is done based on an image completion network. This network must be able to realistically complete a large variety of images. The image completion network we use is trained on the Places [34] scene recognition dataset, which contains over 2.5 million images.

To evaluate our video outpainting method, we use the Densely Annotation Video Segmentation dataset (DAVIS) [22] and the YouTube-VOS [27] dataset. The DAVIS dataset is a high-quality, high-resolution densely annotated video segmentation dataset consisting of videos in two resolutions: 480p and 1080p. There are 50 video sequences with 3,455 annotated frames at pixel level. The YouTube-VOS dataset, similarly to the DAVIS dataset, is intended for Video Object Segmentation (VOS). The dataset is a large-scale benchmark that supports multiple VOS tasks, semi-supervised video object segmentation, and video object segmentation. The dataset consists of more than 4,000 high-resolution YouTube videos and totals more than 340 minutes of video content. For this study, we only use the frames of the videos (cropped on the left and right sides) as input, without using the annotated foreground masks.

Figure 3. Top, a frame from the DAVIS dataset [22]. Bottom, the results of the video inpainting method from Lee *et al*. [13] applied to video outpainting.

4. Video Outpainting

Initially, we look at video inpainting methods to find the additional difficulties that arise with video outpainting. We then focus on one method and propose changes specific to video outpainting to generate more realistic completions.

4.1. Baseline: Inpainting For Video Outpainting

The first methods we tried, Onion-Peel network from Oh *et al*. [19] and Copy-and-Paste network from Lee *et al*. [13], are deep learning-based video inpainting methods. These methods use a set of neighboring frames as input for the completion of each frame. When applied to video outpainting, these methods produce blurry results without temporal consistency. Compared to inpainting, there is less surrounding information available to fill in the missing region. Also, the distance between the given image content and the regions to be filled in increases. Additionally, these methods also iteratively use the previous result for the next completion, which causes them to become incrementally blurrier. On top of that, the temporal consistency is not maintained when the foreground moves out of the frame. The result is illustrated in figure 3.

Next, for flow-based video inpainting methods, we looked at the method from Gao *et al*. [7]. In the regions where information can be extracted from other frames using

Figure 4. Video inpainting method from Gao *et al*. [7] applied to video outpainting. The center of the image was used as input, the left and right sides are completed.

the flow information, the network provides temporal consistency. In the remaining regions, this method fills in the missing pixels based on an image inpainting network. As mentioned previously, an image inpainting network cannot complete the image content. Additionally, this method generates temporal artifacts when a foreground object moves out of frame as illustrated in figure 4.

From these tests, we draw the following conclusions. For maintaining temporal consistency, flow-based methods are the most suitable. These methods can realistically complete large portions of the missing region based on optical flow. However, the image inpainting network used for the remaining completions provides a very unrealistic, blurry result. This is a consequence of the fact that these methods were not built to cope with the additional difficulties of outpainting.

4.2. Overview Of Our Method

An outline of our method is illustrated in figure 5. The input of our video outpainting method is an RGB video. Based on the desired final resolution, a masked region is created indicating which pixel values need to be completed. The extrapolation is done first on the right side of the original image content. To obtain the left side of the completion, the original image content is mirrored, and the method is repeated. Our method consists of 5 steps:

(1) Flow estimation: Using existing techniques, we determine the backward and forward optical flow. We estimate the per-pixel motion between adjacent frames and use it in step 4 to complete parts of the masked regions using temporal propagation. We use optical flow to propagate information between frames to form temporal consistency in the resulting video. More details can be found in section 4.3.

(2) Background estimation: We remove the foreground from the original video frames and optical flow by combining VOS and video inpainting methods. More details can be found in section 4.4.

(3) Flow completion: The optical flow is completed to the new aspect ratio. By completing the optical flow into the masked region, we can also propagate the result of any individual frame completion. More details can be found in section 4.3.

(4) Video completion: To extrapolate the video frames, we use the optical flow information and an image completion network. Based on the optical flow, parts of the masked regions can already be completed by propagating pixel values between adjacent frames. Next, the frame with the largest remaining masked region is selected and completed using an image completion network, as described in section 4.5. The result of this completion can then also be propagated to adjacent frames. This step is repeated until all frames are fully completed. More details can be found in section 4.3.

(5) Post-processing: The extrapolated regions are blurred and combined with the original video frames. More details can be found in section 4.6.

4.3. Video Completion

As mentioned in section 4.1, we base our method on the video inpainting method from Gao *et al*. [7]. We use their color propagation to form parts of the completion based on optical flow. To estimate the optical flow, we use the current state-of-the-art flow estimation method from Teed *et al*. [24].

To complete the optical flow, we minimize the gradient within the masked region as described by Gao *et al*. [7]. They can form smooth completion when removing a moving object entirely or removing part of the background. However, their method causes temporal artifacts when applied to video outpainting with moving foreground objects, as discussed in the following section.

4.4. Background Estimation

When the object is moving on the edge of the frame, a potentially large portion of the object is not visible. In figure 6 we see half a person running on the edge of the camera's view. The completion of the optical flow is not trivial, as we would have to predict the object's movement. The flow completion method from Gao *et al*. [7] causes a mix of the foreground and background motion, since there is no completed outline of the foreground motion. This inaccurate flow completion causes temporal artifacts and deformation in the outpainted region, as illustrated in figure 6.

In simple cases, the outline of the foreground motion could be completed by adding an edge at the side of the frame. This solution removes the temporal deformation, but does not solve the temporal artifacts as these are the result of a mismatch between the extrapolated optical flow and the actual movement.

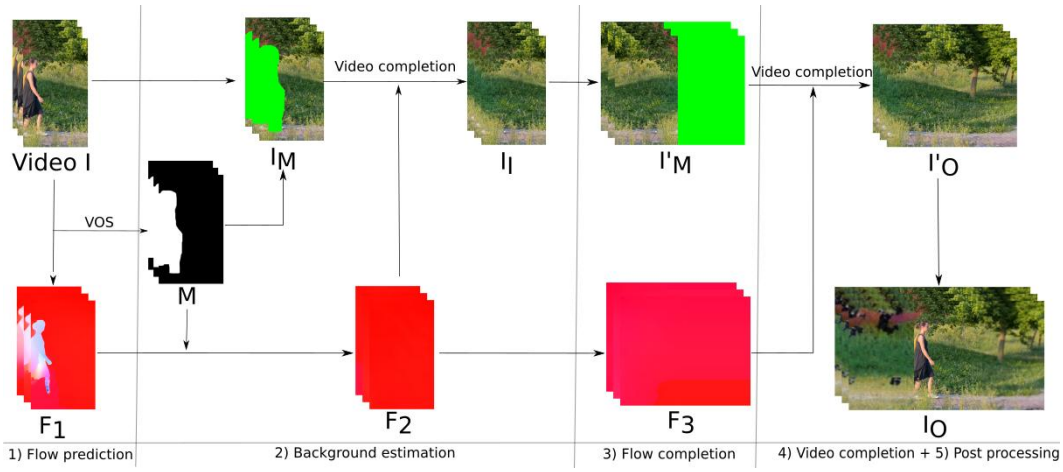

Figure 5. Overview of our video outpainting method consisting of the following five steps. (1) Flow Prediction: Estimating optical flow F_1 based on input frames I. (2) Background estimation: We segment the foreground M using VOS and remove the foreground using video inpainting methods. The masked region is completed in the optical flow to form F_2 and I_I. (3) Flow completion: The optical flow is extrapolated to the new width F_3 and the missing region is added to the frames I'_M. (4) Video completion: Complete the masked regions of I'_M to form the outpainted result I'_O. (5) Post processing: The extrapolated areas are blurred and appended to the original frames I to form the result I_O.

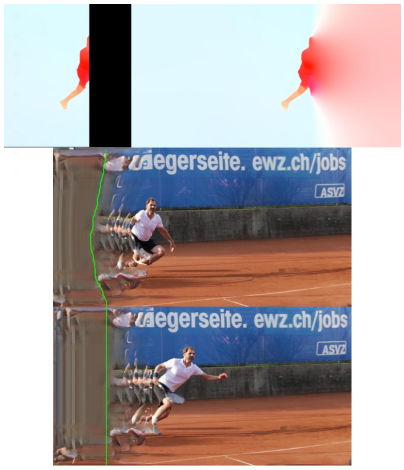

Figure 6. On top the estimated and completed optical flow of a frame with a moving object on the edge. In the middle and bottom two consecutive frames illustrate an example of temporal artifacts. The green line highlights the deformation.

We prevent these temporal artifacts by initially creating a background estimation. To form this background estimation, we combine the VOS method of Lu *et al*. [16] and the video inpainting method of Gao *et al*. [7].

VOS is a binary labeling problem with the goal of separating foreground objects from the background of a video. It is possible to divide VOS into two categories: one shot and zero-shot. In one shot VOS, the first ground truth frame of the foreground mask is available, whereas in zero-shot VOS no ground truth frame is given. Within the context of this study, only the input frames are given and thus only zero-shot VOS is possible. We tried two zero-shot VOS methods. Zhou *et al*. [35] use optical flow for their Motion-Attentive Transition for Zero-Shot Video Object Segmentation (MATnet) method. They achieve the fastest results assuming the optical flow is provided. We found the results of Lu *et al*. [16] to be more accurate. Their method finds correlation between pairs of frames to form the segmentation.

We use the VOS method of Lu *et al*. [16] to detect the foreground, resulting in a foreground mask. This masked region is then completed using the video inpainting method. By removing the foreground entirely, we prevent the temporal artifacts and deformation described earlier.

The use of background estimation before extrapolating the video frames allows us to form more realistic results when there is movement near the end of the frame. This is common in portrait video due to the limited width of the frame. Additionally, we maintain the focus on the original video content by only extrapolating the background and not predicting the movement of the foreground outside the frame.

4.5. Image Completion

We replace the image inpainting network used by Gao *et al*. [7] by the large-scale image completion network from Zhao *et al*. [33]. This network is more suitable for outpainting as the masked regions are generally larger, and it can complete masks of variable shapes.

Currently, the network has to hallucinate what lays out-

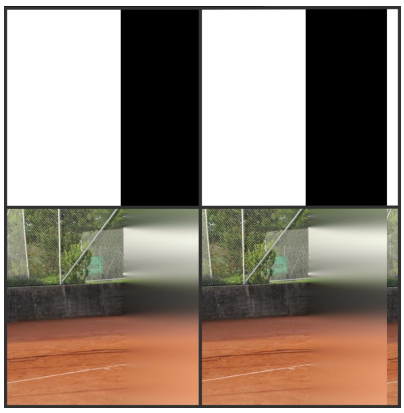

Figure 7. On the left the mask and associated frame, on the right the shifted images. The masked regions are indicated in black.

side the frame, whereas we want something coherent to the given image. In this paper, we propose to add information to the outside of the missing region by shifting the image content as illustrated in figure 7. This addition allows our method to create a more realistic result.

We noticed that, when using a circular shift, a stark transition could be generated in certain cases. Instead, we mirror the right-most known pixels to provide extra information to the image completion network.

By including this image shifting step, our method yields more realistic per-frame completions. In section 5 we evaluate our method with and without this step.

4.6. Post-Processing

Finally, we include an optional post-processing step. It is impossible to accurately predict the information outside the frame. By blurring the completion, there is a clear distinction between the original, real video content and the generated completions. This way, we improve the viewing experience without distracting from the original video or presenting generated footage as original, which might have ethical implications when applied to, for instance, news coverage. Moreover, this is in-line with the earlier proposed foveated approaches to video outpainting [1, 2], where it is proven that blurring the outer regions of the video also improves the viewing experience. We do not take this final blurring into account in during evaluation.

5. Experimental Results

In the previous sections, we described our method for video outpainting. Throughout the literature, there is currently still a lack of research focused on video outpainting. Only Gao *et al*. [7] apply their video inpainting method to video outpainting. We compare our method to this video inpainting method based on the following five evaluation metrics: 1. Mean squared error (MSE). 2. Peak signal-to-noise ratio (PSNR) 3. Structural similarity index measure (SSIM) 4. Learned perceptual image patch similarity (LPIPS) 5. Fréchet Video distance (FVD). We apply these metrics to the DAVIS dataset (480p) [22] and the YouTube VOS dataset [27].

5.1. Portrait To Landscape Conversion

We simulate the conversion from portrait (9:16) to landscape (16:9) video by removing a part on the left and right edges of the videos, as illustrated in figure 8. This way the center third can be used as input and the original full video as ground truth. The following sections discuss the results in more detail. Additionally, we also evaluated our method on the conversion from landscape to ultrawide (21:9) video with the same metrics. For illustration purposes, we generated a video playlist for the four compared methods, plus the results after the post-processing step mentioned in section 4.6. Standard completions entail leaving the completions blank as illustrated in the middle of figure 8. The results are publicly available[2] and can also be found in the supplementary material. There are videos for the following five completion methods:

1. Standard Completions
2. Gao *et al*. [7]
3. Ours without image shifting
4. Ours with image shifting
5. Ours with both image shifting and post processing

The results are shown in table 1. In the following sections, we discuss the results of the five above-mentioned evaluation metrics.

MSE and PSNR

The first and simplest evaluation metrics we used were the Mean Squared Error (MSE) and Peak Signal To Noise Ratio (PSNR). A lower MSE indicates a smaller deviation from the original images and thus a better result.

DAVIS dataset [22]	MSE↓	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
Standard	11293,18	7,95	0,330	0,5397	2009,12
Gao *et al*. [7]	1724,97	16,18	0,560	0,3049	1414,86
Video outpainting (ours)	1654,59	16,82	0,596	0,2635	1244,77
Video outpainting+image shift (ours)	**1513.49**	**17.33**	**0.600**	**0.2530**	**1099,11**
YouTube-VOS [27]	MSE↓	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
Standard	11271,97	8,177	0,354	0,470	2220,93
Gao *et al*. [7]	3008,74	14,37	0,500	0,385	1848,07
Video outpainting (ours)	2702,43	14,46	0,509	0,338	1642,46
Video outpainting+image shift (ours)	**2604,17**	**14,76**	**0,518**	**0,320**	**1374,85**

Table 1. Evaluation of our method for vertical to horizontal video conversion with and without image shifting compared to the video inpainting method of Gao *et al*. [7] and the standard completions on the DAVIS [22] and YouTube-VOS [27] datasets.

[2] github.com/Video-Outpainting/VideoOutpainting

A higher PSNR indicates a better result. Our algorithm achieves the best results here, but the deviation from the ideal value is still relatively large. MSE and PSNR are evaluations that compare pixel values. These pixel-based metrics evaluate to what extent the two images are identical to each other. Within the context of this research, it is not the goal, nor is it possible to recreate the image perfectly. Image and video outpainting are multimodal problems where multiple completions can form an equally realistic result. Therefore, these are not ideal evaluation metrics since the goal is not to recreate the original image content, but rather to generate realistic completions.

SSIM

The structural similarity index measure (SSIM) is a metric that compares two images based on three features: luminance, contrast, and structure. A value of 1 indicates the two images are identical.

Feature-based metrics form a more accurate evaluation of the results. Our result still deviates from the ideal value. This is because no information is available about what is visible outside the given video content. Thus, the complement is formed based on given content while it is possible that a drastic change occurs just outside the frame, as illustrated in figure 8. We can form a realistic completion, but there can always be a significant difference when these two images are compared. In figure 8 we see white buildings just outside the given video content and green forestry in our completion.

The results of Gao *et al*. [7] are relatively close to ours. Based on the optical flow, a large part of the video can be completed. The lower quality of the remaining completions may be more visible to the human eye. The improvements we proposed seem to have a positive but rather small impact on the three features of the SSIM.

LPIPS

Learned Perceptual Image Patch Similarity (LPIPS) [32] is a learning-based metric that approximates how people would evaluate images. Our method achieves the best results, but in this case the distinction is more significant. The result of the method of Gao *et al*. [7] is still high since part of the completion can be done based on the optical flow information. The remaining blurry completion and temporal artifacts seem not to influence this evaluation metric.

FVD

The previous metrics evaluated the frames individually as images. The Fréchet Video Distance (FVD) [25] is a metric specifically aimed at video. A distance of 0 between the two vectors indicates an identical image. Our method shows a significant improvement for this metric.

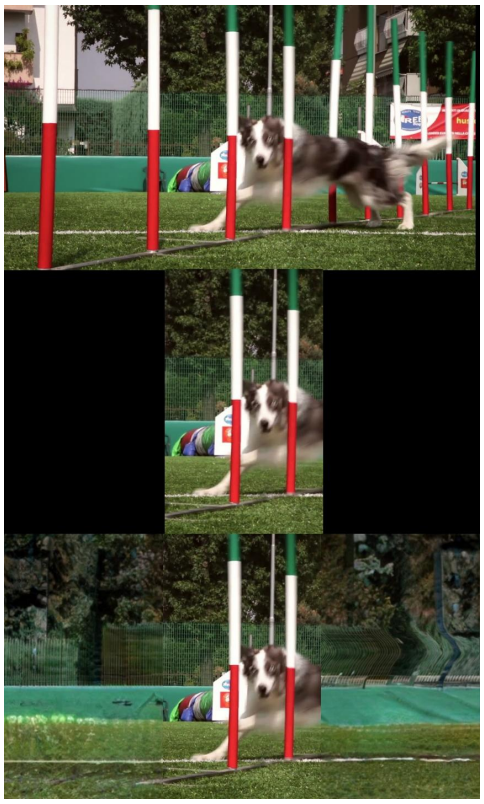

Figure 8. Top, a single frame from the DAVIS dataset [22]. Middle, the input segment of the frame for our method. Bottom, the completed frame.

5.2. Ultrawide Aspect Ratio

In the previous section, we evaluated our results on portrait (9:16) to landscape (16:9) video conversion. This conversion results in a masked region twice the size of the given frame. We also evaluated our method on the conversion of horizontal to ultrawide (21:9) video. The videos in the dataset are all in landscape format. To simulate this conversion, we remove and re-complete the left- and right-most 1/8 of the width. The results are shown in table 2. Visual results are publicly available on our webpage[3] and can also be found in the supplementary material. There are videos for the following five completion methods:

- Standard Completions
- Gao *et al*. [7]
- Ours without image shifting
- Ours with image shifting
- Ours with both image shifting and post processing

Method	MSE↓	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
Standaard	11657,35	7,80	0,329	0,546	346,62
Gao et al. [7]	301,33	23,00	0,809	0,074	254,55
Video outpainting (ours)	277,60	23,82	0,852	0,065	224,77
Video outpainting+image shift (ours)	**239,18**	**24,34**	**0,890**	**0,062**	**207,26**

Table 2. Evaluation of our method for horizontal to ultra-wide video conversion with and without image shifting compared to the video inpainting method of Gao *et al*. [7] and the standard completions on the DAVIS [22] dataset.

Extrapolating a smaller masked region provides a better result. There is less information to complete. Additionally, foreground objects fill a smaller portion of the frame, which improves our background estimation. These factors cause our method to generate significantly more realistic results for this conversion. Our method achieves the best results and can form realistic completions on the videos from the dataset.

6. Limitations

As seen in our results, we chose not to complete the foreground object(s) into the outpainted area. This deliberate choice is to avoid that moving objects near the edge of frames cause artifacts as illustrated in figure 2. Related works do not address these artifacts. We chose to focus on forming more realistic, visually pleasing completions and address these artifacts by first forming a background estimation instead of trying to predict what may or may not have happened outside of frame. This way the viewing experience is enhanced and we do not distract the viewer from the original video content, we avoid drawing the attention to erroneous completions outside of the original video frame.

Our method initially forms a background estimation before forming the completions. This prevents visual artifacts and maintains the focus on the original video content. But, this approach has two drawbacks. Firstly, when the video consists of a close-up foreground object, there is very little information available about the background. This causes the completion to be less realistic.

Secondly, only completing the background causes the foreground objects to disappear into the completed region when they extend outside the original image frame. This unnatural effect is frequently visible in our evaluation dataset, because the input is a cropped wide video. In real use-cases of our algorithm, e.g. a vertical video captured by a mobile phone, this problem is less present because the foreground object is then most likely kept within the frame by the user. However, in order to resolve this, one could separately try to predict the foreground motion outside the frame. We have chosen not to do this, as it would boil down to generating fake video evidence, and go beyond a mere visual enhancement of the original video.

[3] http://github.com/Video-Outpainting/VideoOutpainting

When the video contains faster or more complex camera motion, the completions based on optical flow become less realistic. We notice deformations in the background in those videos.

7. Conclusion

We described a method for video outpainting. Our method expands the aspect ratio of a given video by completing the space at the sides of the given video frames in a realistic and temporally consistent manner.

Our method forms a background estimation to reduce temporal artifacts formed in the outpainting stage. We do this using existing video object segmentation and video inpainting methods. Temporal consistency is achieved using optical flow. Regions that cannot be completed based on the flow information are completed using an image completion network. We propose to shift some image content to the edge to create more realistic results.

We evaluate our method on two datasets and several evaluation metrics. We determine that pixel-based (MSE and PSNR) and feature-based (SSIM) methods are less suitable to evaluate video outpainting. Learning-based metrics (LPIPS and FVD) form a more accurate evaluation. Our method achieves the better results as compared to other state of the art methods in all of these metrics.

Acknowledgements

This work has been supported by the Flemish Government under the AI Research Program.

References

[1] Amit Aides, Tamar Avraham, and Yoav Y. Schechner. Multiscale ultrawide foveated video extrapolation. In *2011 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, 2011. 2, 6

[2] Tamar Avraham and Yoav Y. Schechner. Ultrawide foveated video extrapolation. *IEEE Journal of Selected Topics in Signal Processing*, 5(2):321–334, 2011. 2, 6

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), Aug. 2009. 2

[4] Y. Chang, Z. Liu, K. Lee, and W. Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[5] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 2

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *ICLR*, 2017. 2

[7] C. Gao, A. Saraf, J. Huang, and J. Kopf. Flow-edge guided video completion. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, oct 2020. 2

[9] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[10] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):107:1–107:14, 2017. 2

[11] Huang J., Kang S., Ahuja N., and Kopf J. Temporally coherent completion of dynamic video. *ACM Trans. Graph.*, 35(6), Nov. 2016. 2

[12] Sangwoo Lee, Jungjin Lee, Bumki Kim, Kyehyun Kim, and Junyong Noh. Video extrapolation using neighboring frames. *ACM Transactions on Graphics (TOG)*, 38(3):1–13, 2019. 3

[13] S. Lee, S. Oh, D. Won, and S. Kim. Copy-and-paste networks for deep video inpainting. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[14] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[15] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Hybrid neural fusion for full-frame video stabilization. pages 2279–2288, 10 2021. 3

[16] X. Lu, W. Wang, J. Shen, D. Crandall, and J. Luo. Zero-shot video object segmentation with co-attention siamese networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5

[17] B. Maggia. Video outpainting using conditional generative adversarial networks. Available at https://shareok.org/handle/11244/330234 accessed april 19, 2022. 3

[18] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2

[19] S. Oh, S. Lee, J. Lee, and S. Kim. Onion-peel networks for deep video completion. pages 4402–4411, 10 2019. 3

[20] M. Oliveira, B. Bowen, R. McKenna, and Y. Chang. Fast digital image inpainting. In *Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP)*, pages 261–266, 2001. 2

[21] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. pages 2536–2544, 06 2016. 2

[22] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 1, 3, 6, 7, 8

[23] T. Ruzic and A. Pizurica. Context-aware patch-based image inpainting using markov random field modeling. *IEEE Transactions on Image Processing*, 24(1):444–456, 2015. 2

[24] Z. Teed and J. Deng. *RAFT: Recurrent All-Pairs Field Transforms for Optical Flow*, pages 402–419. 11 2020. 2, 4

[25] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[26] Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):463–476, 2007. 2

[27] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, , and T. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv:1809.03327*, 2018. 3, 6

[28] R. Xu, X. Li, B. Zhou, and C. Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[29] Yufei Xu, Jing Zhang, and Dacheng Tao. Out-of-boundary view synthesis towards full-frame video stabilization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4822–4831, 2021. 3

[30] Z. Yang, J. Dong, P. Liu, Y. Yang, and S. Yan. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10561–10570, 2019. 2

[31] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4470–4479, 2019. 2

[32] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[33] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. Chang, and Y. Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 5

[34] B. Zhou, A. Lapedriza, A. Khosla, and A. Oliva. Places: A 10 million imagedatabase for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452 – 1464, 2017. 3

[35] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Transactions on Image Processing*, 29:8326–8338, 2020. 5

[36] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16448–16457, 2021. 2