Exploring Discontinuity for Video Frame Interpolation - Supplementary Material -

Sangjin Lee*,1 Hyeongmin Lee*,1 Chajin Shin[1] Hanbin Son[1] Sangyoun Lee[1,2]
[1] Yonsei University
[2] Korea Institute of Science and Technology (KIST)
{sglee97, minimonia, chajin, hbson, syleee}@yonsei.ac.kr

(a) (b) (c) (d)

Figure 1. Overlaid examples of GDM dataset

1. Network Details

We employ three Video Frame Interpolation (VFI) models (AdaCoF [2], CAIN [1], VFIT [3]) as baselines to demonstrate the effectiveness of our methods. Since the network architectures are different, we apply our D-map estimators to them in suitable ways respectively. The module of estimating D-map is basically taking general structure of decoder.

AdaCoF uses a U-Net architecture and obtains parameters after passing to the network for interpolation. Therefore, we take a specific layer for the discontinuity map (D-map) from the decoded feature (the output of the 18th layer of its paper). The size of the feature map ($\mathcal{F}$) is $1/2$ the original size, and its channel size is 128.

VFIT is similar to AdaCoF, however, they use a hierarchical structure for the decoder. To estimate D-map without affecting the performance of interpolation, we select a specific layer from the highest-level feature (the output of the 17th layer in the paper). The size of the feature map ($\mathcal{F}$) is $1/16$ the size of the original, and its channel size is 2048.

CAIN takes simple architecture for interpolation with pixel-shuffle and attention. We obtain feature map from the output of the encoder (the output of the first convolution layer in the paper). The size of the feature map ($\mathcal{F}$) is $1/8$ the original size, and its channel size is 384. Consequently, we successfully employ various baseline algorithms for expanding motions by applying our methods in different ways.

*Both authors contributed equally to this work.

2. Network Complexity

Table 1 shows the additional complexity when our methods are applied. As we mentioned in Section 1, the numbers of specific layer for the discontinuity map are different respectively in each network, so the complexity for baseline networks are slightly different. The complexity can be decreased by adjusting the position where we extract the features for discontinuity map.

	AdaCoF	CAIN	VFIT
GFlops	49.7/ 53.4 (+6.9%)	88.5 / 114.4 (+22.6%)	456 / 521(+12.4%)

Table 1. The model complexity analysis

3. Graphical Discontinuous Motion dataset

We construct a new test set called Graphic Discontinuous Motion (GDM) dataset. The GDM dataset consists of high-resolution videos obtained from three types of games. It has 30 sequences and each sequence has 1920×1080 resolution. Its videos contain not only continuous motions but also discontinuous motions.

Figure 1 shows the part of the GDM dataset that contains various types of discontinuous motions. In Figure 1, (a) and (b) show the example of the user interface and the numbers. Case (c) represents when the scene suddenly switches. (d) is the chatting example when we usually do streaming or gaming. As shown in Figure 1, we find that the GDM dataset can be used for evaluating the performance of interpolation without bias.

4. Figure-Text Mixing Details

As general data augmentation, we randomly crop the 256 × 256 patches and flip them horizontally, vertically, and temporally for training. Then, we add Figure-Text Mixing (FTM) augmentation which consists of Figure Mixing

Figure 2. The examples of D-map in various videos.

and Text Mixing. For each mixing technique, we present implementation details in Section 4.1 and 4.2.

4.1. Figure Mixing

Figure Mixing (FM) contains two types of figure, square and circle. Each type can be added respectively. The probability of decision for adding FM is 0.5. Details of FM are shown below.

- **type:** square / circle
- **size:** $10 \leq$ *height, weight* ≤ 41
- **color:** random RGB colors
- **thickness:** $1 \leq$ *thickness* ≤ 4
- **position:** same position for entire videos

4.2. Text Mixing

Text Mixing (TM) presents various types of augmentation. Unlike FM, TM consists of four cases: 1) the position of the text is fixed in the entire video, 2) the text does not exist in the previous frame and appears in the future frame, 3) the existing text suddenly disappears, 4) the text moves up and down by its vertical size. The details of TM are shown below.

- **type:** random text
- **text length:** $5 \leq$ *length* ≤ 30
- **font size:** $10 \leq$ *size* ≤ 40
- **font type:** windows basic fonts
- **color:** random RGB colors
- **position:** same position for input frames

5. Additional Discontinuity Map Visualizations

The discontinuity map (D-map), which highlights discontinuously moving objects in the input image, is an important idea to deal with discontinuous motions. Figure 2 shows the visualizations of D-maps, where I_1 and I_2 are the previous and next adjacent frames. The first row of Figure 2 shows an example of an immediately changing scene. In this case, the D-map highlights almost the entire area that needs to be copied from the previous frame. The second and third rows of Figure 2 demonstrate the effectiveness of the D-map. D-map successfully separates the discontinuous regions from the entire frames containing both continuous and discontinuous motion. Therefore, these results prove that D-map plays a role to suit its purpose, which estimates the regions that should copy and paste from the previous frame. It is especially impressive that the highlighted objects are not in the training dataset and also cannot be learned even with FTM augmentation.

6. Additional Qualitative Results

We show additional results for discontinuous motion in Figure 3. The 1st row in Figure 3 is an example of immediate scene transition. The previous algorithms force interpolation of the two unrelated input frames, resulting in

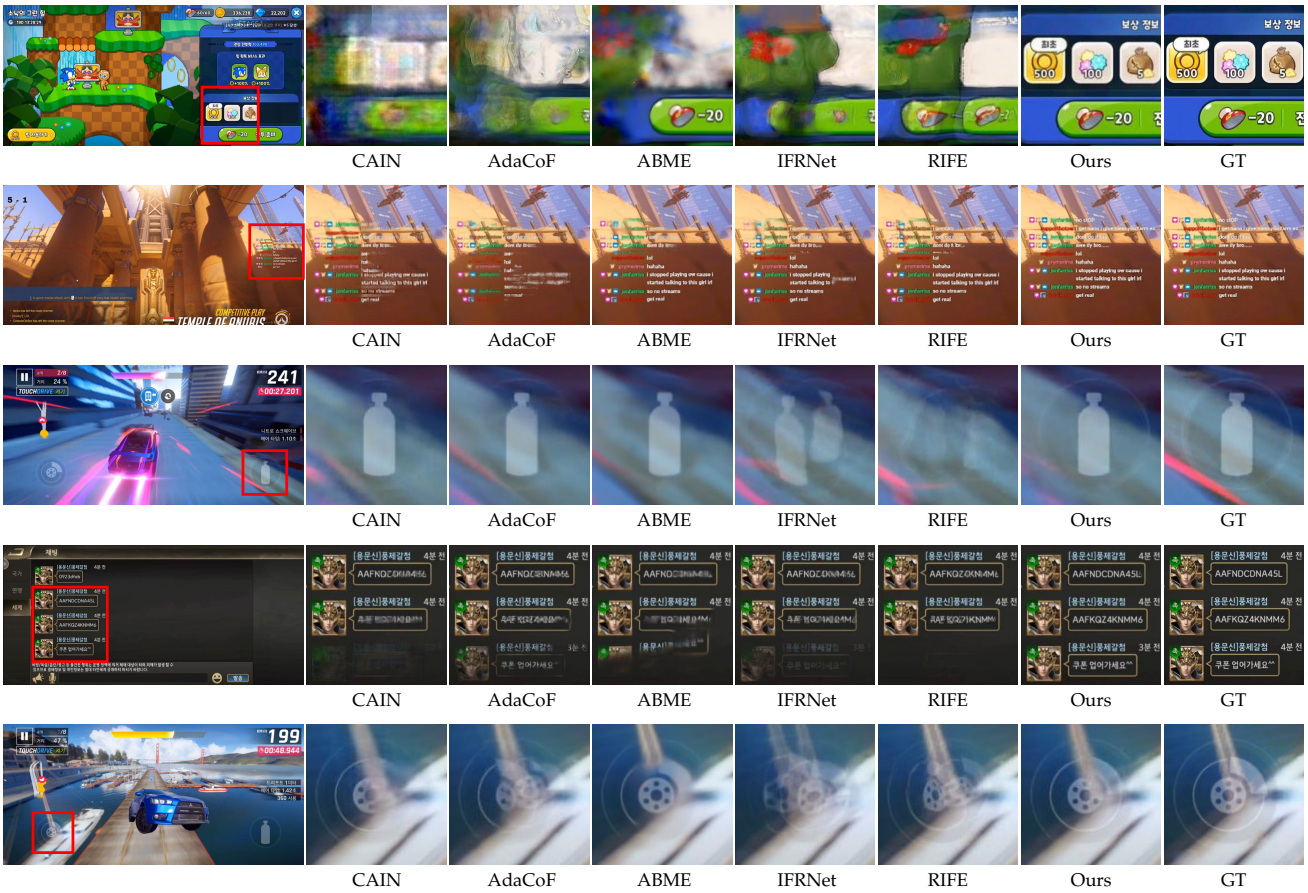

Figure 3. Visual comparison of with discontinuous motion.

damaged frames. However, our method shows a clear result, almost the same as the ground truth image. The 2nd and 4th rows represent the typical discontinuous motion, the scenario of a chatting window. The text moves up discontinuously. The previous methods produce overlapped or distorted frames. However, our method catches the discontinuous motion and shows clear results. 3rd and 5th rows are examples of static user interfaces that can be frequently found in many games. The previous results usually fail, especially when there are some large motions near them. However, our method robustly maintains the structure of the user interfaces better than previous algorithms.

References

[1] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10663–10671, 2020. 1

[2] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5316–5325, 2020. 1

[3] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17482–17491, June 2022. 1