Progressive Temporal Feature Alignment Network for Video Inpainting

Xueyan Zou[1*] Linjie Yang[1] Ding Liu[1] Yong Jae Lee[2]
zxyzou,yongjaelee@ucdavis.edu, linjie.yang,Ding.Liu@bytedance.com
ByteDance AI Lab [1] University of California, Davis [2]

1. User Study

We conduct user study on two types of masks including object masks and curve masks. We apply object masks on the object removal task and curve masks on the corrupted video restoration task to simulate the real world applications. To conduct the user study, we use 20 videos from DAVIS dataset, in which 10 of them is for object removal and the rest 10 is for corrupted video inpainting. Further, we choose FGVC [1], STTN [2] as our baseline methods. To our best knowledge, these two methods are the current state-of-the-art approaches for video inpainting. We find 21 volunteers to participate in this user study. They are shown with four videos including ground truth, FGVC, STTN and our method at the same time. The interface is shown in Fig. 1, the ground truth video is placed in the first location and and the other three candidates are randomly placed at different locations anonymously.

Figure 1. User study sample video.

As shown in Fig. 2, our model outperforms the visual result of FGVC and STTN on all the 10 videos with curve masks for video restoration. On object removal, our approach ranks 1^{st} in 70% of the cases. The overall ranking of our model is 1.4 and 1 on object removal and corrupted video restoration as shown in Fig. 4.

*Work mostly done during an internship at ByteDance Inc.

2. Ablation Study

In addition to the ablation study on optical flow and ground truth flow, we further conduct an ablation study on validity mask to show the effectiveness of masking the inaccurate and unknown flow region during feature propagation. We train this model on DAVIS dataset that initialized from pretrained model on FVI dataset. Further, in order to exclude the effect of inaccurate optical flow, we use optical flow predicting using ground truth image. Each model is trained 200 epochs and we select the best model before 200 epochs. As shown in Table. 1, using validity mask will improve the overall performance.

3. Psudo Code

In this section we provide the pseudo code for our Temporal Shift-and-Align (TSAM) module.

Algorithm 1: Temporal Shift-and-Align Module

Input: $x, lf, rf, lfm, rfm, fd = 8$
out = **torch.zeros_like**(x)
lf = **flow_to_grid**(**resize_flow**(lf, (h,w)))
rf = **flow_to_grid**(**resize_flow**(rf, (h,w)))
lfm = **F.interpolate**(lfm, (h,w), mode='nearest')
rfm = **F.interpolate**(rfm, (h,w), mode='nearest')
lx = **F.grid_sample**(x[:, 1:, :fd], lf)
rx = **F.grid_sample**(x[:, :-1, fd: 2 * fd], rf)
out[:, :-1, :fd] = lf*(1-lfm) + lfm*x[:, :-1, :fd]
out[:, 1:, fd: 2 * fd] = rf*(1-rfm) + rfm*x[:, 1:, fd: 2 * fd]
out[:, :, 2 * fd:] = x[:, :, 2 * fd:]
where **x** is the feature from previous layer, **lf** denotes flow map that could warp feature from $t-1$ to t and **rf** denotes flow map that warp feature from t to $t+1$. **lfm** denotes the validity map for lf and **rfm** denotes the validity map of rf.

It takes the feature x from previous layer, forward/backward optical flow for each sample image which is denoted as lf, rf in Algorithm. 1. Further flow validity mask is provided as lfm and rfm.

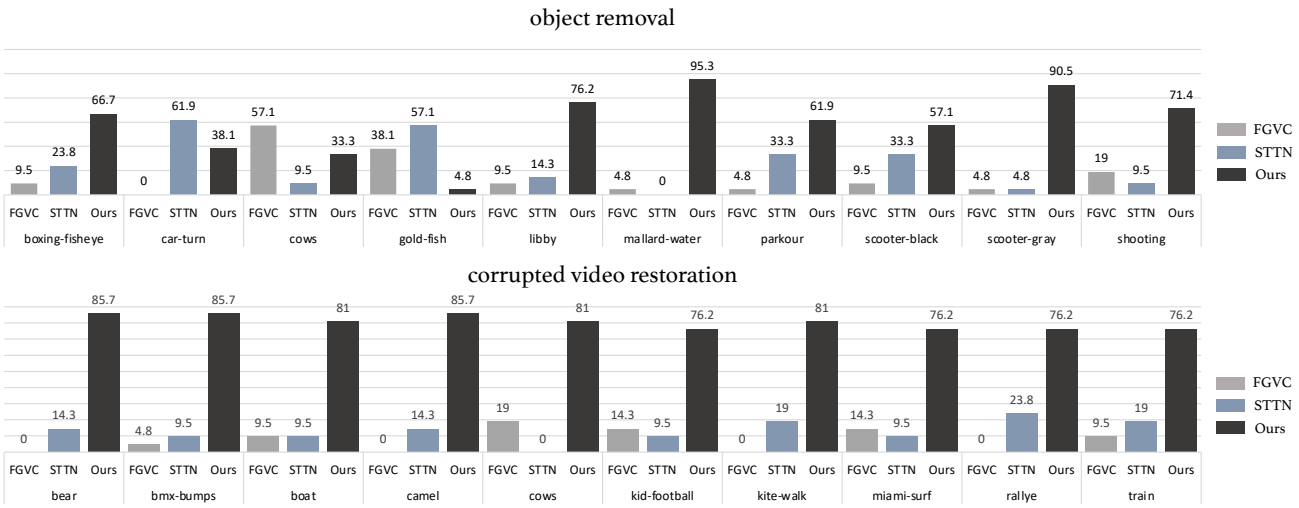


Figure 2. Bar chart of user study on DAVIS

Table 1. Ablation study on validity mask

	Object Mask			Curve Mask			Stationary Mask		
validity mask	PSNR	SSIM	VFID	PSNR	SSIM	VFID	PSNR	SSIM	VFID
	34.46	0.8928	0.3038	36.67	0.9552	0.1839	42.28	0.9757	0.1162
✓	34.5	0.8957	0.3061	36.79	0.9567	0.1789	42.22	0.9767	0.1145

4. Efficiency

We further explore the efficiency of our model. As the optical flow inputs have different computation time according to optical flow algorithms and methods. Thus, we only compute the forward time of our progressive temporal feature alignment network, including encoder and decoder shown in Fig.2 of main paper to compute the FPS. The total computation time is 37 FPS on a single V100 GPU that approximate real time excluding flow computation.

5. Network Architecture

Here, we give a more detailed view of our decoder architecture. It takes the intermediate outputs of the ResNet encoder as input. As shown in Fig. 3, given intermediate features of ResNet layers, we interpolate them into the same size of the corresponding decoder features and add them together to forward to the next layer.

6. Qualitative Result

We show more qualitative results on DAVIS dataset with four different types of masks. It shows that our method could generate higher resolution frames as well as detailed object structures compared with STTN, FGVC and FFVI.

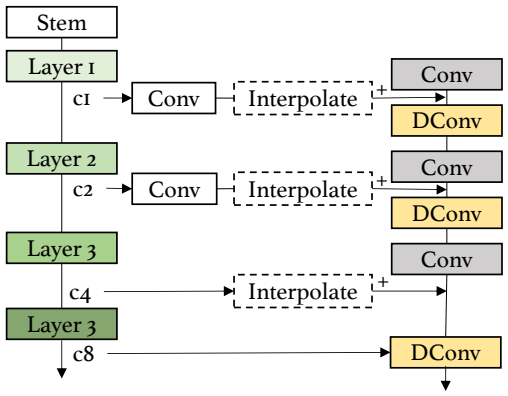


Figure 3. Network architecutre of encoder and decoder.

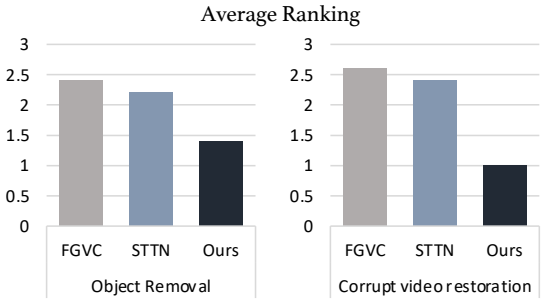
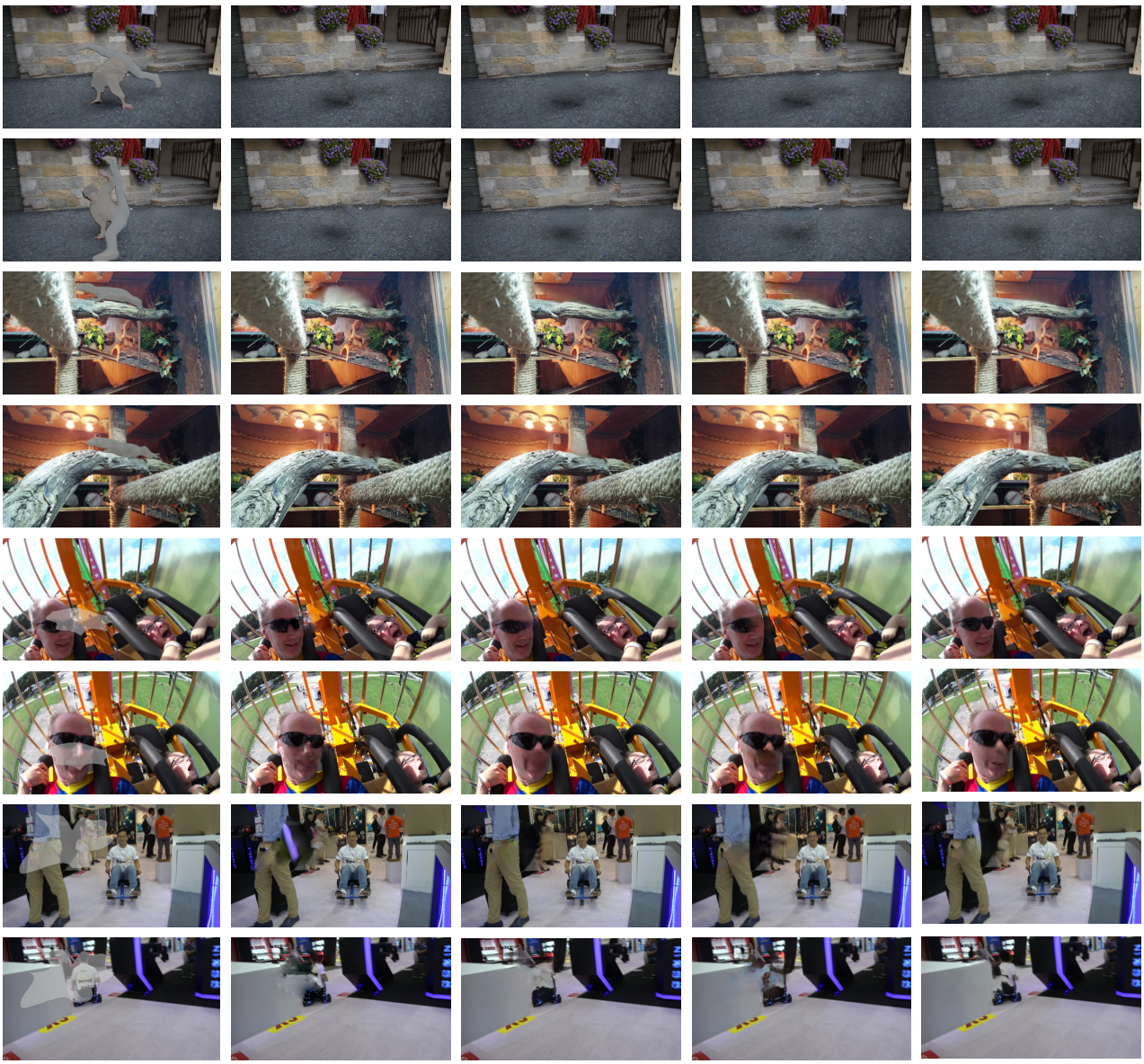


Figure 4. Average ranking of object removal and corrupted video restoration on DAVIS. Less value is better.

References

[1] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *ECCV*, 2020.

[2] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020.



(a) Input Frame (b) FGVC (c) STTN (d) FFVI (e) Ours

Figure 5. Best viewed with zoom in. Qualitative result on DAVIS with moving object/curve masks.

(a) Input Frame (b) FGVC (c) STTN (d) FFVI (e) Ours

Figure 6. Best viewed with zoom in. Qualitative result on DAVIS with stationary mask and object removal.