Supplementary Material

Paper ID 2723

1 Network Structure

As shown in the paper, our proposed network is built upon a recurrent encoder-decoder architecture, which consists of an encoding stage and a decoding stage. In the encoding stage, we have two sub-encoders: encoder a for the stream of the target frame and encoder b for the stream of the neighbor frames. At each encoding scale, we perform Boundary-aware Short-term Context Aggregation (BSCA) between feature maps of the target frame and the neighbor frames to fill the missing regions in the target feature map. Here we use the gated convolution proposed by Yu [3], which learns a dynamic feature selection mechanism and has shown impressive performance on image inpainting. The detailed network parameters of encoder a and encoder b are shown in Table 1.

Table 1. The detailed structure of the encoding stage. GatedConv: gated convolution [3]. FC, KS and SS refer to the feature channel, kernel size, and stride size, respectively.

Encoder a					
Index	Layer	Input	FC	KS	SS
1_a	GatedConv	Target frame	32	3	1
2_a	GatedConv	1_a	64	3	2
3_a	GatedConv	2_a	64	3	1
4_a	BSCA module	3_a, 3_b	64	-	-
5_a	GatedConv	4_a	96	3	2
6_a	GatedConv	5_a	96	3	1
7_a	BSCA module	6_a, 5_b	96	-	-
8_a	GatedConv	7_a	128	3	2
9_a	GatedConv	8_a	128	3	1
10_a	GatedConv	9_a	128	3	1
11_a	BSCA module	10_a, 8_b	128	-	-
Encoder b					
Index	Layer	Input	FC	KS	SS
1_b	GatedConv	Neighbor frame	32	3	1
2_b	GatedConv	1_b	64	3	2
3_b	GatedConv	2_b	64	3	1
4_b	GatedConv	3_b	96	3	2
5_b	GatedConv	4_b	96	3	1
6_b	GatedConv	5_b	128	3	2
7_b	GatedConv	6_b	128	3	1
8_b	GatedConv	7_b	128	3	1

In the decoding stage, the encoding-generated feature map is first refined by the Dynamic Long-term Context Aggregation (DLCA) module, and then a convolution LSTM (Conv-LSTM) layer is applied to increase temporal consistency. At last, the decoder takes the refined latent feature to generate the restored frame. The detailed network parameters of the decoding stage are shown in Table 2.

Table 2. The detailed structure of the decoding stage. The latent feature map refers to the encoding-generated feature map.

Index	Layer	Input	FC	KS	SS
1	DLCA module	latent feature map	128	-	-
2	Conv-LSTM	1	128	3	1
3	GatedConv	2	128	3	1
4	GatedConv	3	128	3	1
5	Upsample (×2)	4	128	-	-
6	GatedConv	5	96	3	1
7	GatedConv	6	96	3	1
8	Upsample (×2)	7	96	-	-
9	GatedConv	8	64	3	1
10	GatedConv	9	64	3	1
11	Upsample (×2)	10	64	-	-
12	GatedConv	11	32	3	1
13	GatedConv	12	3	3	1

2 Training Details

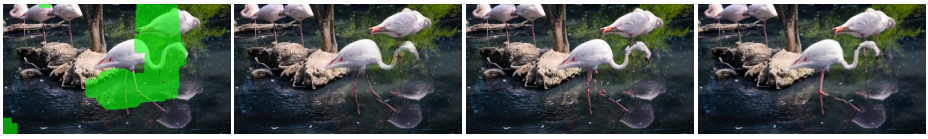
We adopt a two-stage training procedure which gradually learns the model modules: 1) In the first stage, we train the model without the ConvLSTM layer and the DLCA module in order to focus on learning short-term feature aggregation. Moreover, we only utilize the reconstruction loss as the loss function. 2) In the second stage, we add the ConvLSTM layer and the DLCA module in order to increase long-term temporal coherence. We fine-tune the model with the full loss function.

3 Implementation Details

Our model is implemented using Tensorflow. It runs on hardware with NVIDIA Tesla V100 GPUs. We use Adam optimizer with $\beta = (0.9, 0.999)$ and learning rate 10^{-4}. The training of our model takes 3 days using one NVIDIA Tesla V100 GPU.

4 Error Analysis

When dealing with large masks that have much overlapping with the objects, it is challenging for boundary context location and alignment, which may cause



(a) Input (b) CPNet (c) FGNet (d) Ours

Fig. 1. Error Analysis. Better viewed at zoom level 400%.

failures on our model. As shown in Figure 1, our model fails to correctly restore the feet of the bird, but other two models have even worse results. In the future, we will explore more advanced learning-based boundary alignment to mitigate the influence of large masks.

5 Video Inpainting Examples

We present video inpainting examples on the DAVIS test dataset. Here we compare our model against the state-of-the-art CPNet [1] and FGNet [2]. In the video file "examples.mp4", we provide side-by-side comparisons on challenging test videos.

References

1. Lee, S., Oh, S.W., Won, D., Kim, S.J.: Copy-and-paste networks for deep video inpainting. ICCV (2019)
2. Xu, R., Li, X., Zhou, B., Loy, C.C.: Deep flow-guided video inpainting. CVPR (2019)
3. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV (2019)