Video Event Restoration Based on Keyframes for Video Anomaly Detection
(Supplementary Materials)

Zhiwei Yang[1], Jing Liu[1†], Zhaoyang Wu[1], Peng Wu[2†], Xiaotao Liu[1]
[1]Guangzhou Institute of Technology, Xidian University, Guangzhou, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
{zwyang97, neouma, 15191737495}@163.com, xdwupeng@gmail.com, xtliu@xidian.edu.cn

1. Analysis of the video sequence length T

We show in Tab. 1 the variation of AUC on the Ped2 dataset for different values of T. We can infer that as T decrease, the difference between video frames decreases, which reduces the difficulty of network modeling and increases the probability of an anomalous event being restored. The increase in the false negative rate reduces the overall performance. Conversely, as T increases, the difference between video frames increases, which increases the difficulty of network modeling and decreases the probability of normal events being restored. The increase in the false positive rate also reduces the overall performance. When T is set to the middle value of 9, the best performance is obtained by USTN-DSC.

T	5	7	**9**	11	13
AUC	96.4	97.2	**98.1**	94.5	94.3

Table 1. The AUC(%) obtained by USTN-DSC for different values of T on the Ped2 dataset.

2. Feature Extraction Module

Tab. 2 shows the detailed architecture of the feature extraction module. Feature extraction module serves two main purposes. First, it takes advantage of the excellent local modeling ability of the convolutional neural network to capture the underlying local features, such as color, texture, edge, etc., which is beneficial to the restoration of detailed information of video frames in the decoding stage. Second, the feature extraction module can reduce the spatial resolution, thus effectively decreasing the computational effort of the network and accelerating the inference speed.

3. Output Head

The detailed architecture of the output head is shown in Tab. 3. The main role of the output head is to upsample the low resolution features map output from the decoder to the target resolution. To better enhance the quality of the restored video frames, following work [1], we use the PixelShuffle operation for upsampling.

4. More Analysis of DSC

Due to page limitations, we only quantitatively analyze the impact of DSC on model performance in the main paper. To demonstrate more intuitively the role of these two skip connections on the video event restoration task, we perform a qualitative analysis here. First, Fig. 1 visualizes the attention maps of cross attention connection on some samples of the Ped2, Avenue, and ShanghaiTech datasets. It can be observed from the Fig. 1 that the cross attention connection is mainly responsible for the transfer and transformation of dynamic objects features in the foreground. Further, we visualize the feature maps of the temporal upsampling residual connection in Fig. 2, and it is obvious that this skip connection mainly serves the feature transfer and transformation of the background static objects. In addition, we can find that the cross attention connection and the temporal upsampling residual connection in different decoding stages are responsible for different foreground and background parts, which complement each other well. As shown in the quantitative analysis in Tab.2 of the main paper, the performance of the model not equipped with the DSC performs very poorly. The qualitative analysis here illustrates more intuitively that the design of the DSC plays a crucial role in the recovery of static and dynamic objects in video events, facilitating the USTN-DSC to more accurately model normal behavior patterns to better distinguish anomalies.

[†] Corresponding authors.

5. Inference Speed

In the inference stage, our method is implemented on a single NVIDIA RTX 3090 GPU on a machine with CPU core of i7-10700K@3.80Ghz and 32G memory. Our method takes on average 5.3×10^{-3} seconds (188FPS) to process each image of size 256×256.

References

[1] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 1

Layer name	Structure	Output size
Layer 1	Conv2d (3 × 3) + BN + LeakyReLU	(3, 256, 256, 64)
Layer 2	Conv2d (3 × 3) + BN + LeakyReLU	(3, 256, 256, 64)
Layer 3	MaxPool2d (2 × 2)	(3, 128, 128, 64)
Layer 4	Conv2d (3 × 3) + BN + LeakyReLU	(3, 128, 128, 128)
Layer 5	Conv2d (3 × 3) + BN + LeakyReLU	(3, 128, 128, 128)
Layer 6	MaxPool2d (2 × 2)	(3, 64, 64, 128)
Layer 7	Conv2d (3 × 3) + LeakyReLU	(3, 64, 64, 96)

Table 2. Network architecture of the feature extraction module.

Layer name	Structure	Output size
Layer 1	Conv2d (3 × 3)	(9, 64, 64, 384)
Layer 2	PixelShuffle (2) + LeakyReLU	(9, 128, 128, 96)
Layer 3	Conv2d (3 × 3)	(9, 128, 128, 256)
Layer 4	PixelShuffle (2) + LeakyReLU	(9, 256, 256, 64)
Layer 5	Conv2d (3 × 3) + LeakyReLU	(9, 256, 256, 64)
Layer 6	Conv2d (3 × 3)	(9, 256, 256, 3)

Table 3. Network architecture of the output head.

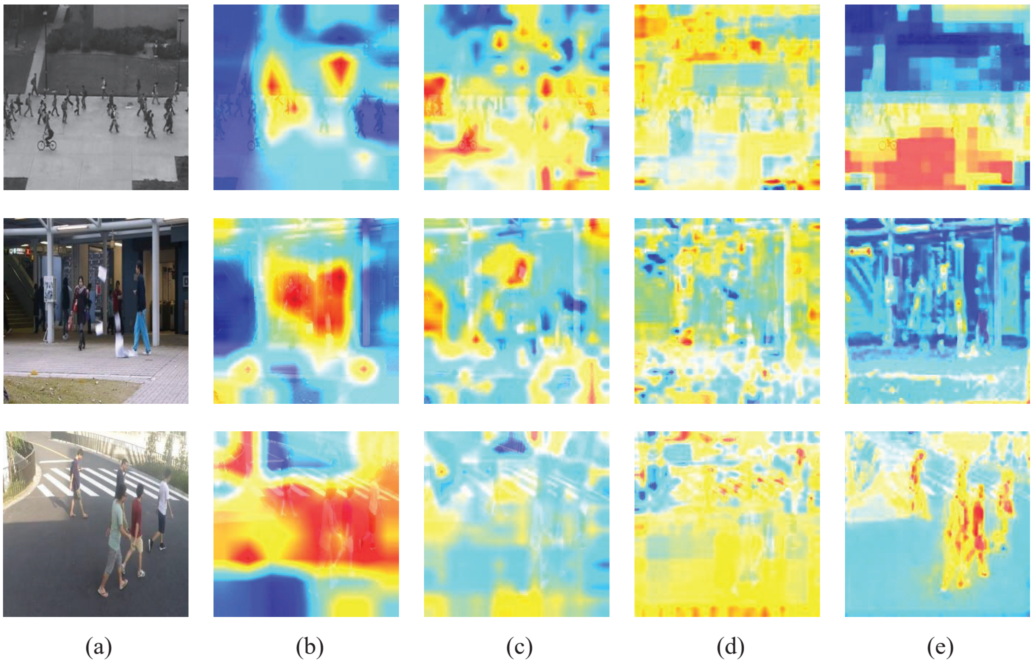

Figure 1. Visualization of attention maps of the across attention connection on some samples of the Ped2, Avenue, and ShanghaiTech datasets. Column (a) denotes the ground-truth frame, and (b) $\sim$ (e) denote the attention maps generated by cross attention connections corresponding to the decoder $D_3 \sim D_0$ stages, respectively.

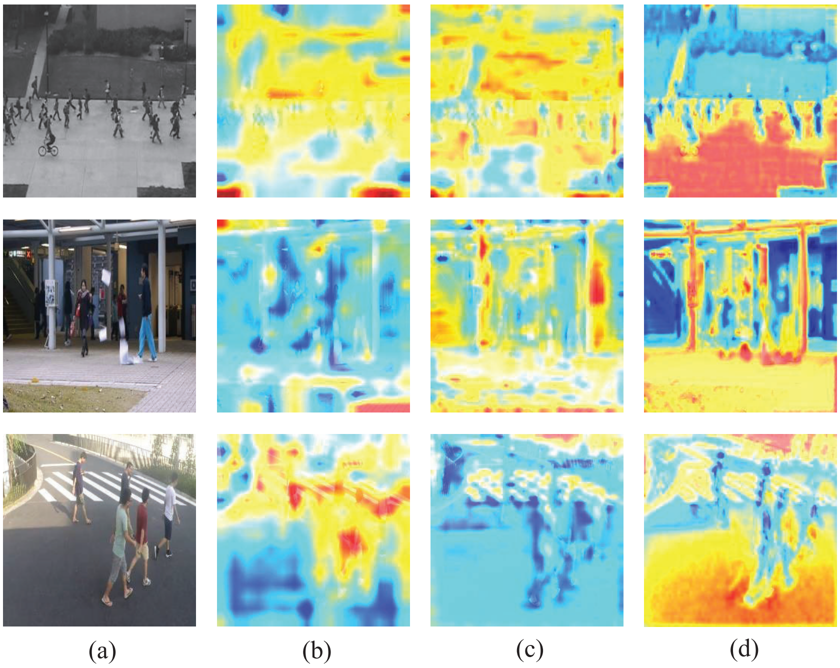

Figure 2. Visualization of the feature maps of temporal upsampling residual connection on some samples of the Ped2, Avenue, and ShanghaiTech datasets. Column (a) denotes the ground-truth frame, and (b) $\sim$ (d) denote the feature maps generated by temporal upsampling residual connections corresponding to the decoder $D_2 \sim D_0$ stages, respectively.