

ral context relationships limits their further performance improvement.

In this paper, we aim to explore better methods for mining complex spatio-temporal relationships in the video. Inspired by video codec theory [15, 18], we propose a brand-new VAD paradigm: video event restoration based on keyframes for VAD. In the video codec [18], three types of frames are utilized, namely I-frame, P-frame, and B-frame. I-frame contains the complete appearance information of the current frame, which is called a keyframe. P-frame contains the motion difference information from the previous frame, and B-frame contains the motion difference between the previous and next frames. Based on these three types of frames containing explicit appearance and motion relative relationships, the video can be decoded. Inspired by this process, our idea sprang up: It should be also theoretically feasible if we give keyframes that contain only implicit relative relations between appearance and motion, and then encourage DNN to explore and mine the potential spatio-temporal variation relationships therein to infer the missing multiple frames for video event restoration. To motivate DNN to explore and learn spatio-temporal evolutionary relationships in the video actively, we do not provide frames like P-frames or B-frames that contain explicit motion information as input cues. This task is extremely challenging compared to reconstruction and prediction-based tasks, because DNN must learn to infer the missing multiple frames based on keyframes only. This requires a strong regularity and temporal correlation of the event in the video for a better restoration. On the contrary, video restoration will be poor for anomalous events that are irregular and random. Under this assumption, our proposed keyframes-based video event restoration task can be exactly applicable to VAD. Fig. 1 compares the video codec and video event restoration task with an illustration.

To perform this challenging task, we propose a novel U-shaped Swin Transformer Network with Dual Skip Connections (USTN-DSC) for video event restoration based on keyframes. USTN-DSC follows the classic U-Net [36] architecture design, where its backbone consists of multiple layers of swin transformer (ST) [21] blocks. Furthermore, to cope with the complex motion patterns in the video so as to better restore the video event, we build dual skip connections in USTN-DSC. Specifically, we introduce a cross-attention and a temporal upsampling residual skip connection to further assist in restoring complex dynamic and static motion object features in the video. In addition, to ensure that the restored video sequence has the consistency of temporal variation with the real video sequence, we propose a simple and effective adjacent frame difference (AFD) loss. Compared with the commonly used optical flow constraint loss [19], AFD loss is simpler to be calculated while having comparable constraint effectiveness.

The main contributions are summarized as follows:

- We introduce a brand-new video anomaly detection paradigm that is to restore the video event based on keyframes, which can more effectively mine and learn higher-level visual features and comprehensive temporal context relationships in the video.
- We introduce a novel model called USTN-DSC for video event restoration, where a cross-attention and a temporal upsampling residual skip connection are introduced to further assist in restoring complex dynamic and static motion object features in the video.
- We propose a simple and effective AFD loss to constrain the motion consistency of the video sequence.
- USTN-DSC outperforms most existing methods on three benchmarks, and extensive ablation experiments demonstrate the effectiveness of USTN-DSC.

2. Related Work

2.1. Video Anomaly Detection

Over the past years, extensive works have been devoted to solving the VAD problem [4, 8, 10, 19, 25–27, 31, 33, 41–44, 46–48, 50, 51], which can be mainly categorized into two main groups based on traditional methods and deep neural network-based methods.

VAD based on traditional methods. Traditional VAD methods mainly utilize statistical models based on hand-extracted features or classical machine learning techniques. For example, Adam *et al.* in [1] characterized the normal local histograms of optical based on statistical monitoring of low-level observations at multiple spatial locations. Kim and Grauman [13] modeled the local optical flow pattern with a mixture of probabilistic principal component analyzers and trained a space-time markov random field to infer abnormalities. Cong *et al.* in [6] introduced a sparse reconstruction cost over the normal dictionary to measure the normality of testing samples. Although these traditional methods achieve better results in specific scenarios, their performance in some complex scenarios is severely constrained owing to poor feature representation capabilities.

VAD based on deep learning methods. With deep learning techniques flourishing in various fields [11, 14, 32, 34, 35, 38], anomaly detection methods based on deep learning have also been widely studied. The most prevalent of these methods are frame reconstruction and frame prediction. For example, Hasan *et al.* in [10] used the extracted features as input to a fully connected neural network-based autoencoder to learn the temporal regularity in the video. A regularity score was calculated according to the reconstruction error and used to determine whether an abnormality occurs. Xu *et al.* in [45] proposed a stacked denoising autoencoder to separately learn both the appearance and the motion features. Liu *et al.* in [20] presented a video anomaly detection

method that predicts the future frame with the U-Net architecture [36]. Yang *et al.* in [47] introduced a dynamic local aggregation network with adaptive clusterer for enhancing the representation capability of normal prototypes in the prediction paradigm. Although the approaches of frame reconstruction and frame prediction currently show promising results, the lack of mining and learning of higher-level visual features and comprehensive temporal context relationships hinder further performance improvement.

2.2. Video Restoration

Video restoration, such as video super-resolution [3, 7, 9], denoising [12], deblurring [39], and inpainting [52], has become a popular research topic in recent years. It aims to restore a clear and high quality video from a degraded low quality video. For example, Geng *et al.* in [7] proposed a real-time spatial temporal transformer to effectively incorporate all the spatial and temporal information for synthesizing high frame rate and high resolution videos. Kim *et al.* in [12] proposed a fast online video deblurring method by efficiently increasing the receptive field of the network without adding a computational overhead to handle large motion blurs. Sheth *et al.* in [39] proposed an unsupervised deep video denoiser, a convolutional neural network architecture designed to be trained exclusively with noisy data. Zou *et al.* in [52] proposed a progressive temporal feature alignment network for video inpainting, which fills the missing regions by making use of both temporal convolution and optical flow.

Difference from these existing video restoration tasks, in this paper, we propose a new video restoration task that restores the video event based on keyframes. This task is more challenging compared to existing video restoration tasks because the missing multiple frames result in the discontinuity of temporal clues. This requires a strong regularity and temporal correlation of the events in the video for a better restoration. On the contrary, there will be a large restoration error for anomalous events, as it is irregular and random. Therefore, the task of restoring the video event based on keyframes can be well applied to VAD.

3. Method

In the unsupervised VAD framework, most approaches devote to designing models that characterize normal behavior patterns and consider deviations from them as anomaly classes. To explore a superior approach to modeling normal behavior, inspired by video codec theory [15], we propose a novel normal behavior modeling paradigm for VAD: Restore video event based on Keyframes to detect anomalies. Concretely, given a video sequence of length T , we take L keyframes in the video sequence as input, aiming to recover the missing $T - L$ frames according to these keyframes. Compared with reconstruction or prediction tasks, it is more

challenging but better to motivate DNN to mine and learn the higher-level visual features and comprehensive temporal context relationships in video sequences. To meet this challenging task, we propose a novel model called USTN-DSC for video event restoration. Next, we will describe the architecture and workflow of USTN-DSC in detail.

3.1. Network Overview

USTN-DSC follows the U-Net [36] architecture design, which consists of four parts: a feature extractor, an encoder, a decoder, and an output head. The feature extractor and output head mainly consist of 2D convolutional layers, and the encoder and decoder are a combination of swin transformer (ST) [21] and 2D convolutional layers.

Fig. 2 illustrates the overall framework of USTN-DSC. Given a video sequence $S = \{I_t | I_t \in \mathbb{R}^{H \times W \times C}\}_{t=1}^T$, where T denotes the length of the video sequence and H, W, C denote the height, width and number of channels of a video frame, respectively. Following the video codec theory [18], we first select the first and last frames of the video sequence as keyframes. To encourage USTN-DSC to automatically learn potential appearance and motion evolution relationships in the video sequence, instead of giving P-frames and B-frames with explicit motion information, we take the intermediate frame of video sequences as temporal transition frame for spatio-temporal relationship development. Therefore, we take $I_1, I_{(T-1)/2+1}$, and I_T of the video sequence S as the three keyframes of the input, and stack them up in chronological order as $X \in \mathbb{R}^{3 \times H \times W \times C}$. Then, the input X is first processed by the feature extraction module F_e for initial feature extraction and dimensionality reduction. Next is the encoding part, and the encoder consists of four stages, denoted by $E_n, n = 0, 1, 2, 3$, each of which is a stack of ST encoder block E_n^{ST} followed by a convolution layer φ_n , except for the final stage E_3 . Symmetrically with the encoder, the decoder is denoted by $D_n, n = 0, 1, 2, 3$, each of which is a stack of ST decoder block D_n^{ST} followed by a deconvolution layer φ_n^{-1} , except for the first stage D_0 . Finally, the output of the decoder is further transformed by the output head F_{out} to obtain the restored video sequence $\hat{S} = \{I_t | I_t \in \mathbb{R}^{H \times W \times C}\}_{t=1}^T$.

Specifically, in USTN-DSC, F_e first extracts the features $f_X \in \mathbb{R}^{3 \times H' \times W' \times C'}$ from X . Then, E_0^{ST} first takes f_X as input and obtains $f_0^e = E_0^{ST}(f_X) \in \mathbb{R}^{3 \times H' \times W' \times C'}$ and then follows a convolutional layer to obtain $e_0 = \varphi_0(f_0^e) \in \mathbb{R}^{3 \times H'/2 \times W'/2 \times C'}$. Subsequently, we have

$$\begin{cases} f_n^e = E_n^{ST}(e_{n-1}), & n = 1, 2, 3 \\ e_n = \varphi_n(f_n^e), & n = 1, 2 \\ e_3 = f_3^e \end{cases} \quad (1)$$

During the encoding phase, each e_n corresponds to the output features maps of the three keyframes, i.e. $e_n \equiv$

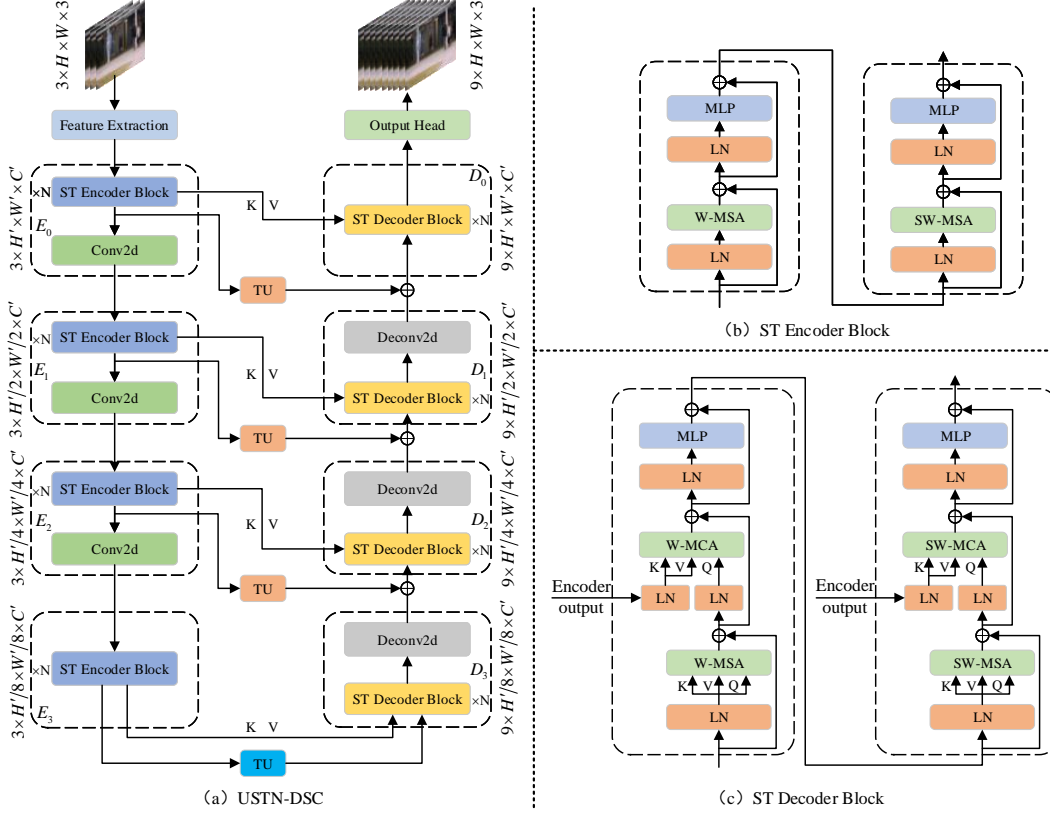


Figure 2. The overview architecture of USTN-DSC. (a) The complete network structure of USTN-DSC. TU represents the temporal upsampling module. (b) The structure of ST Encoder Block. LN represents Layer Normalization. W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations, respectively. (c) The structure of ST Decoder Block. W-MCA and SW-MCA are multi-head cross-attention modules with regular and shifted windowing configurations, respectively.

$\{e_n^1, e_n^{(T-1)/2+1}, e_n^T\}$. After obtaining the final output e_3 from the encoder, a temporal upsampling (TU) module located at the bottleneck generates the initial features $q_0 = TU(e_3)$ based on e_3 for subsequent restoration of missing frames. Then, q_0 is further divided into two parts q_0^f and q_0^b . q_0^f represents the initial features of the missing frames between I_1 and $I_{(T-1)/2+1}$, and q_0^b represents the one between $I_{(T-1)/2+1}$ and I_T . For the features corresponding to the timestamps of I_1 , $I_{(T-1)/2+1}$, and I_T , we directly keep following the ones on the corresponding time points in e_3 . Eventually, the prototype features used for decoder input are represented as $Q := (e_3^1, q_0^f, e_3^{(T-1)/2+1}, q_0^b, e_3^T)$.

Next, we move on to the decoding phase. First, D_3^{ST} in D_3 takes e_3 and Q as input and obtains $f_3^d = D_3^{ST}(e_3, Q)$ and then follows a deconvolution layer to obtain $d_3 = \varphi_3^{-1}(f_3^d)$. It is noted that D_n^{ST} here differs from the original ST block which only compute self-attention, we construct a skip connection from the encoder to compute cross-attention, which is the first channel of the dual skip connections. Next, we construct the second skip connection. The features e_2 from the encoder are temporally upsampled

into $e_2^u = TU(e_2)$ using the TU module. Then, similar to the synthesis process of the prototype features Q , e_2^u and e_2 are temporally dimensionally concatenated to obtain the combined features e_2^r . Then, e_2^r are added to d_3 in the form of residual and fed into D_2^{ST} followed by a deconvolution layer φ_2^{-1} . The operation of the subsequent layers is similar and we formulate them as follows:

$$\begin{cases} d_3 = \varphi_3^{-1}(D_3^{ST}(e_3, Q)), \\ e_n^r = TU(e_n) \cup e_n, & n = 0, 1, 2 \\ d_n = \varphi_n^{-1}(D_n^{ST}(e_n, d_{n+1} + e_n^r)) & n = 1, 2 \\ d_0 = D_0^{ST}(e_0, d_1 + e_0^r) \end{cases} \quad (2)$$

Finally, the output features d_0 are transformed into the restored video sequence $\hat{S} = \{I_t | I_t \in \mathbb{R}^{H \times W \times C}\}_{t=1}^T$ by the output head F_{out} . we show the specific structure of the F_e and F_{out} in the supplementary materials.

3.2. Encoder

The encoder of USTN-DSC mainly consists of stacked multiple ST encoder blocks followed by a convolutional layer. The detailed structure of the ST encoder block is

shown in Fig. 2(b). Although we deal with video data here, we do not use the division of the 3D shifted windows as in the video swin transformer [22] for capturing the temporal relationships, as this is a bit trivial for the input of only three frames. In order to use a more simple way to equip the ST with the ability to learn long-range spatio-temporal dependencies, we calculate the local window attention by combining all the windows on the space corresponding to the current window simultaneously. Specifically, given an input video sequence of size $3 \times H' \times W' \times C'$, a ST block partitions it into non-overlapping windows of size $3 \times \left\lceil \frac{H'}{M} \right\rceil \times \left\lceil \frac{W'}{M} \right\rceil \times C'$. Here we choose, $H' = W' = 64$, $M = 4$ and $C' = 96$. Then, we reshape the video frame features with divided windows into $\frac{H'W'}{M^2} \times 3M^2 \times C'$, where $\frac{H'W'}{M^2}$ is the total number of windows. Next, the reshaped features are first layer normalized (LN) and followed by a window-based multi-head self-attention (W-MSA) [21] to compute the local attention of each window. Immediately after, a multi-layer perceptron (MLP) follows another LN layer for further features transformations. Then, an additional ST block with shifted window-based multi-head self-attention (SW-MSA) [21] is applied to implement cross-window interactions to learn long-range dependency information. In this second ST block, every module is the same as the previous block except that input features are shifted by $\left\lfloor \frac{M}{2} \right\rfloor \times \left\lfloor \frac{M}{2} \right\rfloor$ before window partitioning. Using this alternating regular and shifting window partitioning way, it not only makes the ST block requires less computationally cost but also enables the ST block to have the cross-window interaction capability, thus capturing long-range dependencies in both spatial and temporal dimensions. Finally, the outputs of such ST blocks are downsampled by a convolutional layer with a stride of two, serving as the input of the next encoder stage.

3.3. Decoder

Symmetrically with the encoder part, the decoder of the USTN-DSC also consists of four stages with $D_n, n = 0, 1, 2, 3$, each of which in turn is followed by ST decoder block with a deconvolution layer for upsampling. The detailed structure of the ST decoder block is shown in Fig. 2(c). We restore the missing frames in the decoding phase mainly by means of the conversion of the features from the keyframes extracted from the encoder part. For the missing video frames, they contain both slow moving objects, whose differences with keyframes are minimal, and objects with large motion, which need to be synthesized by inference of spatio-temporal relationship of the keyframes. In order to cope with these two different motion patterns for better restoration of missing video frames, we introduce the dual skip connections in the decoder section. First, we insert a corresponding multi-head cross-attention (MCA) af-

ter the regular and shifted windows-based multi-head self-attention. The MCA receives the features from the output of the previous decoding layer as query, and the features from the corresponding level of the encoder as key and value. By querying the features at different scales and distances in the encoder part of the corresponding level, MCA enables the decoder to assist in the generation of certain fast-motion object features of the missing frames. Second, we design a TU module consisting of a 3D deconvolution layer with a kernel size of $(T - 3) \times 3 \times 3$ and stride size of $1 \times 1 \times 1$ to upsample the features generated by the encoder to obtain the features of the intermediate missing frames. (Note that the TU module in the skip connection shares weights, except for the one in the bottleneck section.) Then, the features at the timestamps of the corresponding keyframes are filled with the original features from the encoder. Finally, the combined features are added to the output of the corresponding level of the decoder in the form of residual. This operation can compensate for the lack of original detail features query in the cross-attention connection and can further facilitate the decoder to better restore the detail information of background and slow objects in video sequence.

3.4. Loss Function

We mainly consider the loss function from both appearance and motion aspects. First, we use the charbonnier loss [16], which compensates for the shortcomings of the L_1 and L_2 losses, to compute the RGB differences between the corresponding output frame I_t and the real frame \hat{I}_t for appearance constraint:

$$L_{cb}(\hat{I}_t, I_t) = \sqrt{\|\hat{I}_t - I_t\|^2 + \epsilon^2}, \quad (3)$$

where ϵ is set to 10^{-3} in our experiments. For the motion constraint, we introduce a simple and effective AFD loss:

$$L_{fd}(\{\hat{I}_t\}_{t=1}^T, \{I_t\}_{t=1}^T) = \sum_{t=1}^{T-1} \sqrt{\|\|\hat{I}_t - \hat{I}_{t+1}\|^2 - \|I_t - I_{t+1}\|^2\|^2 + \epsilon^2}. \quad (4)$$

AFD loss directly promotes motion consistency by constraining the difference between the pixel of adjacent frames of the restored video sequence and the real video sequence. Compared with the computationally expensive optical flow constraint method, AFD loss is not only simple to compute but also has a comparable temporal constraint effect. Finally, the overall loss function is given as follows:

$$L_{all} = L_{cb} + L_{fd}. \quad (5)$$

3.5. Anomaly Detection on Testing Data

During the testing phase, we take T -frames as the processing unit, but we do not use the error between each re-

stored frame I_t and the real frame \hat{I}_t as its corresponding anomaly indicator. Because we find experimentally that the keyframes and frames adjacent to the keyframes have very slight errors with the real frames, even for anomalous events. Therefore, we take the $PSNR$ corresponding to the frame with the largest mean square error between the T -frames and the real frames as the anomaly detection indicator for this video sequence, formulated as follows:

$$\begin{cases} t^* = \max_{1 \leq t \leq T} \frac{1}{K} \sum_{i=0}^K (\hat{I}_{t,i} - I_{t,i})^2 \\ PSNR(\hat{I}_{t^*}, I_{t^*}) = 10 \log_{10} \frac{[max_{\hat{I}_{t^*}}]^2}{\frac{1}{K} \sum_{i=0}^K (\hat{I}_{t^*,i} - I_{t^*,i})^2} \end{cases}, \quad (6)$$

where K is the total number of image pixels and $max_{\hat{I}_{t^*}}$ is the maximum value of image pixels. We assign the same $PSNR$ value to each frame of a processing unit for anomaly metric calculation. To quantify the probability of anomalies occurring, we normalize each $PSNR$, following work [29], to obtain anomaly scores in the range $[0, 1]$:

$$S(t) = 1 - \frac{PSNR(\hat{I}_t, I_t) - \min_t PSNR(\hat{I}_t, I_t)}{max_t PSNR(\hat{I}_t, I_t) - \min_t PSNR(\hat{I}_t, I_t)}. \quad (7)$$

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate the performance of our method on three classic benchmark datasets widely used in the VAD community. (1) Ped2 [28]. It contains 16 training videos and 12 testing videos in fixed scenarios. Abnormal events include riding a bicycle, skateboarding, and driving a vehicle on the sidewalk. (2) Avenue [24]. It consists of 16 training videos and 21 testing videos with 47 abnormal events including throwing a bag, moving toward or away from the camera, and running on the sidewalk. (3) ShanghaiTech [26]. It contains 330 training videos and 107 testing videos with 130 abnormal events, such as affray, robbery, fighting, etc., distributed in 13 different scenes.

Evaluation Metric. Following the widely used evaluation metrics in the field of VAD, we use the frame-level area under the curve (AUC) of receiver operation characteristic to evaluate the performance of our proposed method.

Training Details. In the training phase, we first resize each frame to the size of 256×256 , while the values of the pixels in all frames are normalized to $[0, 1]$. Then, we use the Adam optimizer with L_2 and decoupled weight decay [23] by setting $\beta_1 = 0.9$ and $\beta_2 = 0.99$ to train the USTN-DSC. The initial learning rate is set to 2×10^{-4} and is gradually decayed following the scheme of cosine annealing. The length of the output video sequence T is set to 9. The ST block depth N of each stage in USTN-DSC is set to 6. Training epochs are set to 100, 150, 200 on Ped2, Av-

Methods		Ped2	Avenue	SHTech
others	SCL [24]	N/A	80.9	N/A
	Unmasking [41]	82.2	80.6	N/A
	AnomalyNet [51]	94.9	86.1	N/A
	DeepOC [43]	96.9	86.6	N/A
	MPED-RNN [30]	N/A	N/A	73.4
	Scene-Aware [40]	N/A	89.6	74.7
R.	Conv-AE [10]	90.0	70.2	60.9
	ConvLSTM-AE [25]	88.1	77.0	N/A
	Stacked RNN [26]	92.2	81.7	68.0
	AMC [31]	96.2	86.9	N/A
	MemAE [8]	94.1	83.3	71.2
	CDDA [5]	96.5	86.0	73.3
	MNAD [33]	90.2	82.8	69.8
	Zhong et al. [50]	97.7	88.9	70.7
P.	FFP [19]	95.4	84.9	72.8
	AnoPCN [48]	96.8	86.2	73.6
	MNAD [33]	97.0	88.5	70.5
	ROADMAP [42]	96.3	88.3	76.6
	MPN [27]	96.9	89.5	73.8
	AMMC-Net [4]	96.9	86.6	73.7
	DLAN-AC [47]	97.6	89.9	74.7
	USTN-DSC	98.1	89.9	73.8

Table 1. Quantitative comparison with the state of the art for anomaly detection. We measure the average AUC (%) on Ped2 [28], Avenue [24], and ShanghaiTech [26] datasets. The comparison methods are listed in chronological order. ('R.' and 'P.' indicate the reconstruction and prediction tasks, respectively.)

enue, and ShanghaiTech, respectively, with batch size set to 4. We train our model on a single NVIDIA RTX 3090 GPU.

4.2. Experiment Results

Comparison with Existing Methods. We compare the performance of USTN-DSC with various state-of-the-art methods under different paradigms in Tab. 1. It can be seen from Tab. 1 that the performance of our method on Ped2 and Avenue datasets achieve state-of-the-art compared to other methods and has a substantial improvement over the pioneer methods based on the deep learning reconstruction [10] and prediction [19] paradigms. This demonstrates that our method is a more effective modeling paradigm for learning normal behavior patterns to distinguish anomalies. For the ShanghaiTech dataset, the performance of our method does not achieve the optimum, but it is quite competitive compared with other methods. Because the ShanghaiTech dataset contains 13 different scenes, where the backgrounds and motion objects involved are quite complex and variable. This poses a higher demand on the ability of the model to learn the spatio-temporal relationships. However, we analyze the effect of different ST block depth N on the model performance in sec.4.3 and demonstrate that USTN-DSC is

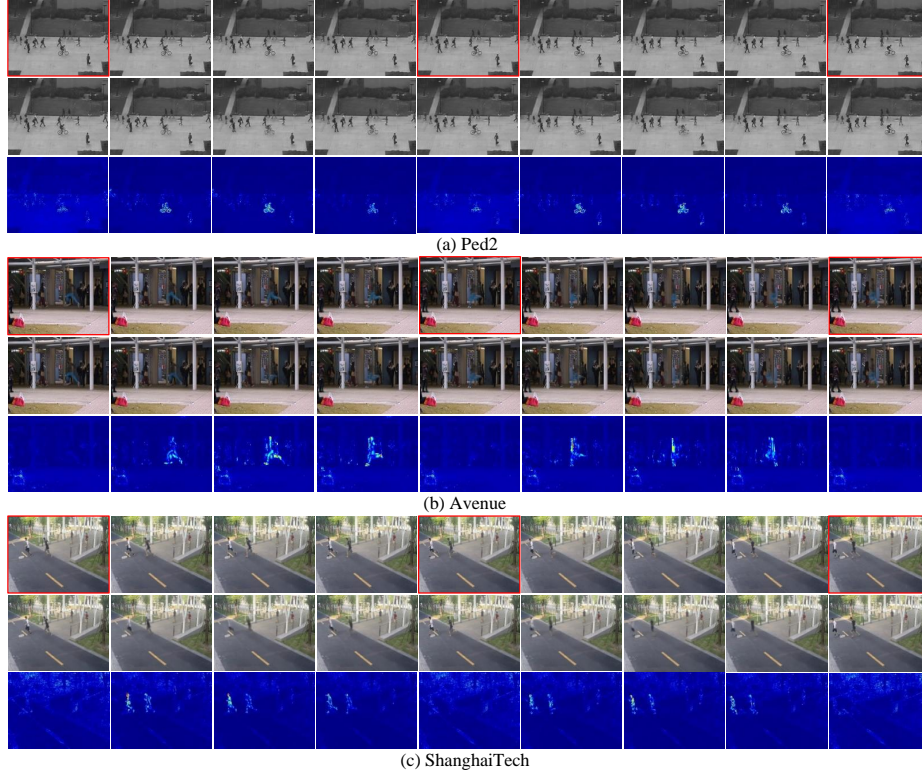


Figure 3. Examples of video events restoration on three datasets. For each dataset, the first row is the real video event sequence, the second row is the restoration results of our method based on keyframes, and the third row is the restoration error map. The frames with red borders are keyframes.

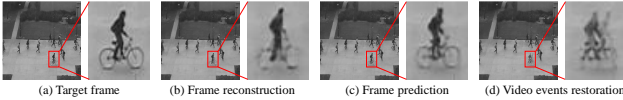


Figure 4. Compare the output results of USTN-DSC with reconstruction and prediction-based methods for anomalous samples.

able to further improve the model performance as the depth N continues to increase. In addition, USTN-DSC follows a simple U-Net architecture design. We do not employ other assists such as optical flow [4], adversarial training [19], extraction of the foreground objects [30], or memory enhancement [8, 33]. Nevertheless, compared to these well-equipped methods, USTN-DSC still surpassed them, which further validates the effectiveness of our method.

Qualitative Results. We present the results of our method to restore video sequences based on keyframes of abnormal video samples on the ped2, avenue, and ShanghaiTech datasets, respectively, in Fig. 3. It can be observed that for the normal regions in the video frames, our method is able to restore them well, while drastic errors occur for abnormal event regions. Then, Fig. 4 shows the results of video frames restored by our method compared with the output

of existing reconstruction-based [33] and prediction-based methods [19]. It can be seen that the frame (located in the middle of two keyframes) of the anomalous samples restored by our method have much large distortion and deformation errors in the anomalous region compared to the output of the previous two methods. This can further demonstrate that our approach, which considers the evolutionary relationship between appearance and motion over the long term, is able to effectively learn the more discriminative behavioral patterns in normal videos and thus be able to more accurately distinguish abnormalities. Fig. 5 shows the anomaly score curves of some video clips on the three datasets. Obviously, there is a sharp jump in the anomaly score with the occurrence of anomalous events, and the anomaly curve returns to flat when the anomalous events disappear. This further demonstrates that our method has excellent sensitivity to anomalies and can effectively detect anomalous events.

4.3. Ablation Study and Analysis

In this section, we analyze the impact of several key components, network parameters, and loss functions on the performance of our method. Due to limited space, the anal-

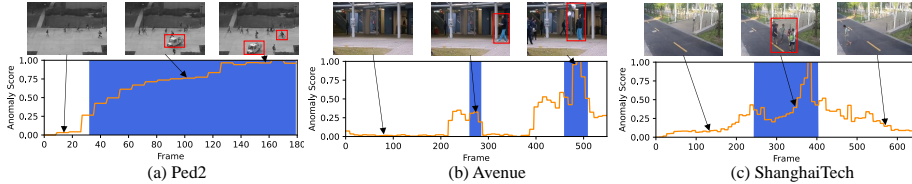


Figure 5. Anomaly score curves of several test samples of our method on three benchmark datasets.

	Ped2	Avenue	ShanghaiTech
w/o DSC	91.1	83.2	67.4
w CAC	96.4(+5.3)	87.9(+4.7)	71.2(+3.8)
w TUC	95.2(+4.1)	85.4(+2.2)	70.6(+3.2)
w DSC	98.1(+7.0)	89.9(+6.7)	73.8(+6.4)

Table 2. The AUC(%) obtained by USTN-DSC with different skip connection configurations on Ped2 [28], Avenue [24] and ShanghaiTech [26] datasets. (DSC: dual skip connections, CAC: cross attention connection, TUC: temporal upsampling connection)

	Ped2	Avenue	ShanghaiTech
w/o AFDL	97.2	88.5	71.5
w AFDL	98.1(+0.9)	89.9(+1.4)	73.8(+2.3)

Table 3. The AUC(%) obtained by USTN-DSC with or without adjacent frames difference loss (AFDL) on Ped2 [28], Avenue [24] and ShanghaiTech [26] datasets.

ysis of the selection of different video sequence lengths T is given in the supplementary materials.

Dual Skip Connections Analysis. In Tab. 2, we show the variation of the performance of our proposed network on the three datasets from without any skip connection to equipping two skip connections one by one. As we can see from Tab. 2, for the encoder-decoder network without any skip connection, its performance is quite unsatisfactory. When adding these two skip connections in turn, the performance of our method is significantly improved. Especially for the addition of cross-attention skip connection, it boosts the performance on Ped2, Avenue, and ShanghaiTech datasets by 5.3%, 4.7%, and 3.8%, respectively. This shows that the construction of dual skip connections plays a critical role in contributing to video event restoration. More detailed analysis can be referenced in the supplementary materials.

AFD loss Analysis. Tab. 3 shows the performance differences on ped2, avenue, and ShanghaiTech datasets with and without AFD loss. It can be seen that the AFD loss has a performance boost on all three datasets, especially for the ShanghaiTech dataset, where it improves the AUC by 2.3%. This is because the ShanghaiTech dataset involves 13 different scenarios with complex motion patterns of foreground objects that have a higher dependence on motion constraints. This demonstrates the effectiveness of our pro-

	Ped2	Avenue	ShanghaiTech
$N=2$	97.1	87.8	71.5
$N=4$	97.7(+0.6)	89.2(+1.4)	72.6(+1.1)
$N=6$	98.1(+1.0)	89.9(+2.1)	73.8(+2.3)

Table 4. The AUC(%) obtained by USTN-DSC with different ST block depth N on Ped2 [28], Avenue [24], and ShanghaiTech [26].

posed AFD loss for motion constraints.

ST Block Depth N Analysis. As shown in Tab. 4, we show the performance variation on the Ped2, Avenue, and ShanghaiTech datasets by setting N to 2, 4, 6. From the Tab. 4, we can find that the performance of USTN-DSC on all three datasets gradually improves as N increases. Interestingly, for the Ped2 dataset, the performance improvement from increasing N is quite slight, while the improvement is very obvious for the Avenue and ShanghaiTech datasets. This can be explained by the fact that the scenes in the Ped2 dataset are fixed and the motion patterns are relatively simple, so a shallow network can meet the modeling requirements. For the Avenue and ShanghaiTech datasets, their scenes are more complex and diverse, and place higher demands on the modeling capabilities of the network. Due to hardware constraints, we are not attempting to set a larger N currently. However, it can be expected that the performance on Avenue and ShanghaiTech datasets can be further improved as N continues to increase.

5. Conclusions

In this paper, we introduced a brand-new video anomaly detection paradigm that is to restore a video event based on keyframes. To this end, we proposed a novel model called USTN-DSC for video events restoration, where a cross-attention and a temporal upsampling residual skip connection are introduced to further assist in restoring complex dynamic and static motion object features in the video. In addition, we introduced a temporal loss function based on the pixel difference of adjacent frames to constrain the motion consistency of the video sequence. Extensive experiments on three benchmark datasets show that our method outperforms most existing state-of-the-art methods, demonstrating the effectiveness of our method.

References

- [1] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560, 2008. 1, 2
- [2] Yannick Benezeth, P-M Jodoin, Venkatesh Saligrama, and Christophe Rosenberger. Abnormal events detection based on spatio-temporal co-occurrences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2458–2465, 2009. 1
- [3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4778–4787, 2017. 3
- [4] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 938–946, 2021. 1, 2, 6, 7
- [5] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *Proceedings of the European Conference on Computer Vision*, pages 329–345, 2020. 6
- [6] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3449–3456, 2011. 2
- [7] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17441–17451, 2022. 3
- [8] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1705–1714, 2019. 1, 2, 6, 7
- [9] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Space-time-aware multi-resolution video enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2020. 3
- [10] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–742, 2016. 1, 2, 6
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2
- [12] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Scholkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4038–4047, 2017. 3
- [13] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928, 2009. 2
- [14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 2
- [15] Théo Ladune and Pierrick Philippe. Aivc: Artificial intelligence based video codec. *arXiv preprint arXiv:2202.04365*, 2022. 2, 3
- [16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2599–2613, 2018. 5
- [17] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the International Conference on Machine Learning*, pages 1558–1566, 2016. 1
- [18] Guozhu Liu and Junming Zhao. Key frame extraction from mpeg video stream. In *Third International Symposium on Information Processing*, pages 423–427, 2010. 2, 3
- [19] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection—a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. 1, 2, 6, 7
- [20] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection—a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. 2
- [21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3, 5
- [22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 5
- [23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [24] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. 6, 8
- [25] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 439–444, 2017. 1, 2, 6
- [26] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 1, 2, 6, 8

- [27] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15425–15434, 2021. 1, 2, 6
- [28] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010. 6, 8
- [29] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 6
- [30] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019. 6, 7
- [31] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1283, 2019. 1, 2, 6
- [32] Xiushan Nie, Bowei Wang, Jiajia Li, Fanchang Hao, Muwei Jian, and Yilong Yin. Deep multiscale fusion hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):401–410, 2020. 2
- [33] Hyunjong Park, Jongyoun Noh, and Bumsuh Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020. 1, 2, 6, 7
- [34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 2
- [35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 2
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015. 2, 3
- [37] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. Real-time anomaly detection and localization in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–62, 2015. 1
- [38] Fangtao Shao, Jing Liu, Peng Wu, Zhiwei Yang, and Zhaoyang Wu. Exploiting foreground and background separation for prohibited item detection in overlapping x-ray images. *Pattern Recognition*, 122:108261, 2022. 2
- [39] Dev Yashpal Sheth, Sreyas Mohan, Joshua L Vincent, Ramon Manzorro, Peter A Crozier, Mitesh M Khapra, Eero P Simoncelli, and Carlos Fernandez-Granda. Unsupervised deep video denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1759–1768, 2021. 3
- [40] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 184–192, 2020. 6
- [41] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2895–2903, 2017. 1, 2, 6
- [42] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 1, 2, 6
- [43] Peng Wu, Jing Liu, and Fang Shen. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2609–2622, 2019. 1, 2, 6
- [44] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision*, pages 322–339, 2020. 2
- [45] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017. 2
- [46] Zhiwei Yang, Jing Liu, and Peng Wu. Bidirectional retrospective generation adversarial network for anomaly detection in videos. *IEEE Access*, 9:107842–107857, 2021. 2
- [47] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *Proceedings of the European Conference on Computer Vision*, pages 404–421, 2022. 1, 2, 3, 6
- [48] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcnn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1805–1813, 2019. 1, 2, 6
- [49] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *Proceedings of the International Conference on Machine Learning*, pages 1100–1109, 2016. 1
- [50] Yuanhong Zhong, Xia Chen, Jinyang Jiang, and Fan Ren. A cascade reconstruction model with generalization ability evaluation for anomaly detection in videos. *Pattern Recognition*, 122:108336, 2022. 1, 2, 6
- [51] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. AnomalyNet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10):2537–2550, 2019. 1, 2, 6
- [52] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16448–16457, 2021. 3

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

Zhiwei Yang¹, Jing Liu^{1†}, Zhaoyang Wu¹, Peng Wu^{2†}, Xiaotao Liu¹

¹Guangzhou Institute of Technology, Xidian University, Guangzhou, China

²School 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.

[†]Corresponding authors.

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.

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.

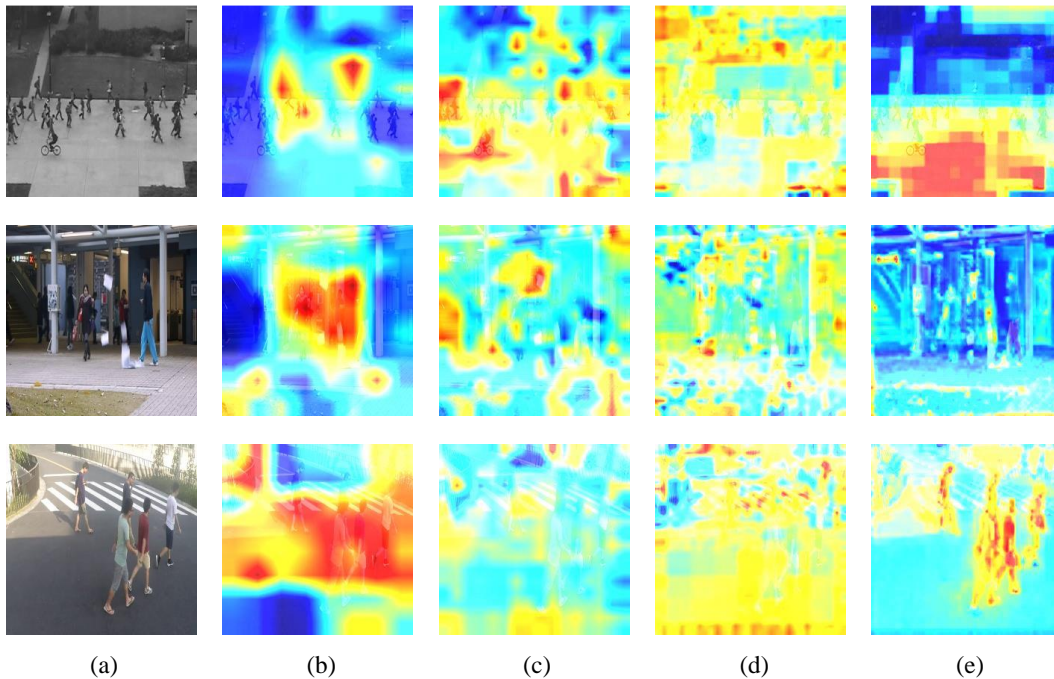


Figure 1. Visualization of attention maps of the across attention connection on some samples of the Ped2, Avenue, and ShanghaiTech datasets. Column (a) denotes the ground-truth frame, and (b) ~ (e) denote the attention maps generated by cross attention connections corresponding to the decoder $D_3 \sim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

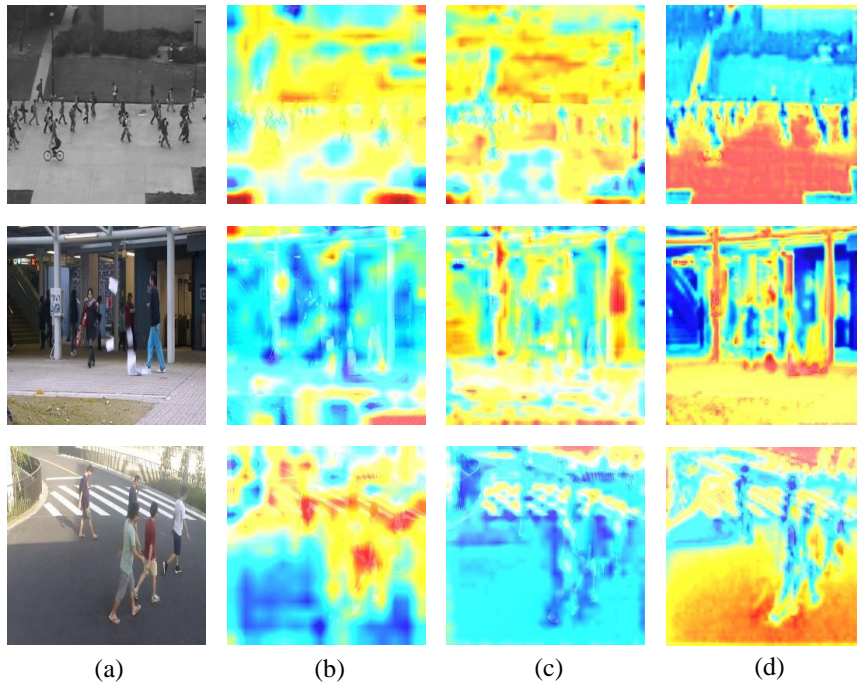


Figure 2. Visualization of the feature maps of temporal upsampling residual connection on some samples of the Ped2, Avenue, and ShanghaiTech datasets. Column (a) denotes the ground-truth frame, and (b) ~ (d) denote the feature maps generated by temporal upsampling residual connections corresponding to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