SCIENCE CHINA
Information Sciences

• MOOP •

September 2022, Vol. 65 194101:1–194101:3
https://doi.org/10.1007/s11432-021-3444-9

Reverse erasure guided spatio-temporal autoencoder with compact feature representation for video anomaly detection

Yuanhong ZHONG[1*], Xia CHEN[1], Jinyang JIANG[2] & Fan REN[3]

[1] *School of Microelectronics and Communication Engineering, Chongqing University, Chongqing* 400044*, China;*
[2] *State Grid Chongqing Electric Power Research Institute, Chongqing* 401123*, China;*
[3] *Changan Software Technology Company, Changan Automobile Corp., Chongqing* 401120*, China*

Received 29 September 2021/Revised 8 December 2021/Accepted 11 February 2022/Published online 17 August 2022

Citation Zhong Y H, Chen X, Jiang J Y, et al. Reverse erasure guided spatio-temporal autoencoder with compact feature representation for video anomaly detection. Sci China Inf Sci, 2022, 65(9): 194101, https://doi.org/10.1007/s11432-021-3444-9

Video anomaly detection aims to learn normal patterns and identify the samples deviating from normal patterns as anomalies. In early research, methods based on handcrafted low-level features have been widely studied. However, the representation power of low-level features is insufficient for describing various patterns, causing a bottleneck in handcrafted-feature-based anomaly detection. Recent methods are typically used to build reconstruction or prediction models based on deep learning to represent normal frames and detect anomalies based on the representation error [1]. However, most existing detection methods based on deep learning adopt the loss function, such as l_1-norm and l_2-norm, to calculate the reconstruction or prediction error [2]. In these methods, all pixels in the frame are processed equally, that is, the model loses its focus and does not prioritize learning and reconstructing the complex regions that are difficult to reconstruct during training. Consequently, the model may not be able to obtain reconstructed image with high quality foreground, since the simple background pixels dominate the optimization of model. Unfortunately, such issue may reduce the performance of anomaly detection, because the foreground is more important than the stationary background in anomaly detection. Further, existing reconstruction methods attempt to minimize the difference between the reconstructed frame and its ground truth [3]. Although similarity is guaranteed in the pixel or even latent space, it is a one-to-one constraint, which ignores the similarity of different normal frames.

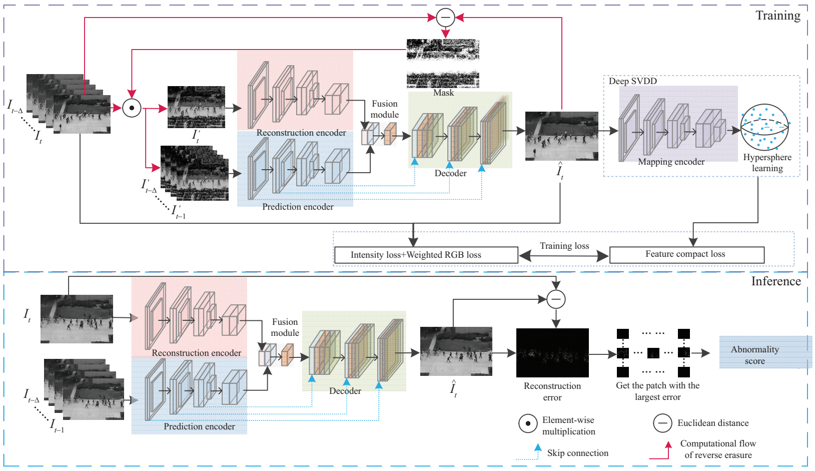
Framework. To solve these problems, we propose a dual-encoder single-decoder network, denoted by DESDnet, with a novel training strategy, including reverse erasure based on reconstruction error and deep support vector data description (SVDD) [4]. The workflow of the proposed method is shown in Figure 1(a). The DESDnet is designed to extract spatial and temporal features individually from the current and past frames. The decoder simultaneously utilizes spatial and temporal features to reconstruct the current frame and detect anomalies based on the reconstruction error. In the training strategy, reverse erasure is employed to guide the network to optimize in the expected direction by providing a prior information, further improving the reconstruction accuracy of the normal frame. Deep SVDD aims to enlarge the difference between the reconstructed images of normal and anomaly by controlling the features in the latent space.

Reverse erasure based on reconstruction error. We perform reverse erasure based on reconstruction error in the training phase. Specifically, after each training iteration, the pixel-level error between the target frame I_t and the reconstruction frame $\hat{I}_t$ is calculated. Based on whether the value in error map is larger than a given threshold, the mask is obtained by setting it to one or zero. Before the next epoch, the raw frames from $I_{t-\Delta}$ to I_t are multiplied pixel by pixel with the mask to create the input data for the network, denoted by $I'_{t-\Delta}$ to I'_t. Because the foreground contains many moving objects, the reconstruction error of the foreground is typically considerably larger than that of the background in the frame. Considering this, we set the threshold as the average reconstruction error of the corresponding frame. Thus, the erased frame retains most of the foreground pixels and discards most of the background pixels, helping the model to automatically form an attention mechanism to the foreground. In this case, a natural assumption is that both the simple background and the complex foreground will be reconstructed with high quality. The uncertain change in the input also makes the DESDnet more robust against noise.

Deep SVDD. To perform deep SVDD in the training phase, an encoding network behind decoder, called mapping encoder, is added; it has the same structure as the reconstruction encoder. The mapping encoder maps reconstructed frame into the low-dimensional feature representation. Deep SVDD expects these low-dimensional representations to be fitted into a hypersphere with a minimum

* Corresponding author (email: zhongyh@cqu.edu.cn)

(a)



(b)

	Avenue		Ped2		ShanghaiTech
	AUC	EER	AUC	EER	AUC
Deep-cascade (Sabokrou et al., 2017) [7]	–	–	–	8.2%	–
STAE-optflow (Zhao et al., 2017) [5]	80.9%	20.9%	88.6%	24.4%	–
Frame-Pred (Liu et al., 2018) [2]	84.9%	–	95.4%	–	72.8%
ALOCC (Sabokrou et al., 2018) [9]	–	–	–	13%	–
MemAE (Gong et al., 2019) [1]	83.3%	–	94.1%	–	71.2%
DeepOC (Wu et al., 2019) [4]	86.6%	**18.5%**	96.9%	8.8%	–
MNAD (Park et al., 2020) [3]	**88.5%**	–	97.0%	–	70.5%
ST-CaAE (Li et al., 2021) [8]	83.5%	23.5%	92.9%	12.7%	–
STCEN (Hao et al., 2022) [6]	86.6%	19.2%	96.9%	8.0%	73.8%
Proposed	87.7%	19.4%	**97.5%**	**6.7%**	**74.2%**

Figure 1 (Color online) (a) Workflow of the proposed model. (b) Frame-level AUC/EER comparison on Avenue, Ped2, and ShanghaiTech datasets with different methods. The best and second-best performances are represented in bold and underlined, respectively.

volume, forcing the network to learn the common factor of normal events. With this constraint, the reconstructed normal frames are as similar as possible, thus effectively increasing the distance between the reconstructed normal and abnormal frames. Wu et al. [4] employed deep SVDD to constrain the latent features of input frames, thereby essentially constraining the encoder used to map the input frame. However, owing to the generalization of the decoder constructed by convolutional neural network, we cannot guarantee that anomalies cannot be represented even if the encoder is constrained, that is, the difference between the reconstruction of the normal and anomaly may not be obvious. Consequently, the method in [4] is not suitable for our model, because the reconstruction error rather than the distance in the feature space, is utilized to detect anomaly in our method.

Reconstruction encoder and prediction encoder. In the DESDnet, the reconstruction encoder and prediction encoder are used to provide the spatial and temporal features for the decoder, respectively. Skip connections are employed between the prediction encoder and decoder to provide multiscale low-level features for image conversion. In the training phase, given the erased frames from $I'_{t-\Delta}$ to I'_t, I'_t is input into the reconstruction encoder to extract the appearance features, and the frames from $I'_{t-\Delta}$ to I'_{t-1} are input to the prediction encoder to extract the motion features. Compared with using optical flow to capture motion patterns, our method avoids the inaccuracy and high computational cost caused by optical flow calculation. In the testing phase, the raw frames from $I_{t-\Delta}$ to I_t are input into the reconstruction and prediction encoders to extract the spatial and temporal features of the video sequence, respectively.

Fusion module and decoder. To integrate appearance and motion features, we adopt a two-dimensional convolution layer followed by a Tanh activation layer as the fusion module. The convolution kernel is 1 × 1 with a channel size of 512. The appearance and motion features are cascaded and then input into the fusion module to obtain spatio-temporal features. Compared with the fusion method of concatenating features, our method reduces the computational cost and increases the representation ability of the network. Further, the spatio-temporal features are input into the decoder to reconstruct the frame.

Training loss. To constrain the reconstruction frame in pixel and latent space, the training loss function is defined as

$$L = \lambda_{\text{int}} L_{\text{int}} + \lambda_{\text{rgb}} L_{\text{rgb}} + \lambda_{\text{compact}} L_{\text{compact}}, \qquad (1)$$

where λ_{int}, λ_{rgb}, and λ_{compact} are the hyperparameters corresponding to each loss, and these determine their contribution to the total training loss.

The intensity loss L_{int} is to maximize the pixel-by-pixel similarity between the reconstruction frame $\hat{I}_t$ and its ground truth I_t, which is computed as

$$L_{\text{int}} = \|\hat{I}_t - I_t\|_2^2, \qquad (2)$$

where t is the t-th frame, and $\|.\|_2$ represents l_2-norm.

Inspired by [5], we adopt a weighted RGB loss L_{rgb} to improve the similarity of successive frames and constrain motion patterns.

$$L_{\text{rgb}} = \frac{1}{N} \sum_{i=1}^{N} \frac{N-i+1}{N} \|\hat{I}_t - I_{t-i}\|_2^2, \qquad (3)$$

where N denotes the number of previous frames. The weight of I_{t-i} is $\frac{N-i+1}{N}$, which decreases with an increase in i, because the larger the distance between frames is, the greater the difference between them is.

Based on deep SVDD, we define the constraint on the latent space as a feature compact loss:

$$L_{\text{compact}} = R^2 + \frac{1}{vn}\sum_{t=1}^{n} \text{argmax}\{0, \|\Phi(I_t;W) - c\|_2^2 - R^2\}, \tag{4}$$

where c and R represent the center and radius of the hypersphere, and n is the number of frames. $\Phi(I_t;W)$ is the feature representation of $\hat{I}_t$ output by the network with parameters W. argmax$\{\cdot\}$ is the function to take the maximum. In (4), the first term aims to minimize the volume of hypersphere and the second term is the penalty term of the samples lying outside the hypersphere. $v \in (0, 1]$ is used to weigh the volume and boundary losses of the hypersphere.

To constrain the reconstructions of all normal frames to a reachable range, the mean of feature vectors of the reconstruction frames extracted by the first epoch training model is regarded as the center c. During the subsequent training, the Euclidean distance between the feature representation of the reconstruction frame and center c is computed, and then L_{compact} is obtained. By minimizing the feature compact loss of normal frames, the reconstruction images of all normal frames will be more similar, whereas the reconstruction images of abnormal frames will be more different, so as to increase the distinguishability of the anomaly.

Anomaly detection on testing data. During the testing phase, the frames from $I_{t-\Delta}$ to I_t are input into the DESD-net, and reconstruction frame $\hat{I}_t$ is obtained. In our method, the patch with the largest reconstruction error in the test frame is utilized to calculate the abnormality score, which is conducive to highlighting the anomaly occurring within a small region in the scene. Unlike [4], we employ deep SVDD to constrain training rather than evaluate anomalies because high-level features in the latent space inevitably lose details. First, the partial score of each patch is defined as follows:

$$S(P) = \frac{1}{|P|}\sum_{i,j\in P} (I_{i,j} - \hat{I}_{i,j})^2, \tag{5}$$

where P represents a patch in frame I, and $|P|$ is the number of pixels in P. i, j indicates the spatial position of the pixels. The partial score of the patch with the largest $S(P)$ in the test frame is then selected as the abnormality score of the frame, denoted by Score. Finally, the Score of all frames in each video is normalized to the range of [0, 1].

$$\text{Score}^*(I_t) = \frac{\text{Score} - \min_{\text{Score}}}{\max_{\text{Score}} - \min_{\text{Score}}}, \tag{6}$$

where $\min_{\text{Score}}$ and $\max_{\text{Score}}$ represent the minimum and maximum values in the test video, respectively. Considering the temporal continuity of the events, a Gaussian filter is applied to smooth the frame-level abnormality scores in the temporal dimension. An anomaly can be detected based on the abnormality score, as the score of the abnormal frame is often higher than that of the normal frame.

Experiments and results. We compare the proposed method with state-of-the-art methods [1–9] on the Avenue, Ped2, and ShanghaiTech datasets, and area under curve (AUC) and equal error rate (EER) results are presented in Figure 1(b). A comparison of the AUC reveals that our method achieves good AUC performance on the three datasets, showing substantial competitiveness. On the Ped2 and Avenue datasets, the proposed method achieves an AUC of 97.5% and 87.7%, respectively, exhibiting the best and second-best detection performances among these methods. In particular, although the ShanghaiTech dataset is a challenging dataset for video anomaly detection, our method yields the highest AUC of 74.2%. As there are few methods for obtaining the EER on the ShanghaiTech dataset, we do not compare the EER results on this dataset. Compared with DeepOC [4] and STCEN [6], although the EER of our method is slightly worse on Avenue, the EER on Ped2 is better by 2.1% and 1.3%, respectively. For other methods, such as Deep-cascade [7] and STAE-optflow [5], we also obtain better EER results.

Conclusion. In conventional video anomaly detection based on deep learning, the deep network is optimized without focus and the similarity between different normal frames is ignored. To alleviate these issues, we designed a dual-encoder single-decoder network to reconstruct frames and proposed a training strategy involving reverse erasure based on the reconstruction error and deep SVDD to regularize the training of the network. With this training strategy, the proposed model achieved high performance in terms of both the AUC and EER. Future work will involve the application of our training strategy to more complex tasks.

Acknowledgements This work was partially supported by Special Project of Technological Innovation and Application Development of Chongqing (Grant No. cstc2019jscx-msxmX0167) and National Natural Science Foundation of China (Grant No. 61501069).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Gong D, Liu L, Le V, et al. Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 1705–1714

2 Liu W, Luo W, Lian D, et al. Future frame prediction for anomaly detection-a new baseline. In: Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 6536–6545

3 Park H, Noh J, Ham B. Learning memory-guided normality for anomaly detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, 2020. 14372–14381

4 Wu P, Liu J, Shen F. A deep one-class neural network for anomalous event detection in complex scenes. IEEE Trans Neural Netw Learn Syst, 2019, 31: 2609–2622

5 Zhao Y, Deng B, Shen C, et al. Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, 2017. 1933–1941

6 Hao Y, Li J, Wang N, et al. Spatiotemporal consistency-enhanced network for video anomaly detection. Pattern Recognition, 2022, 121: 108232

7 Sabokrou M, Fayyaz M, Fathy M, et al. Deep-cascade: cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans Image Process, 2017, 26: 1992–2004

8 Li N, Chang F, Liu C. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. IEEE Trans Multimedia, 2021, 23: 203–215

9 Sabokrou M, Khalooei M, Fathy M, et al. Adversarially learned one-class classifier for novelty detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 3379–3388