2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) | 978-1-6654-3396-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/AVSS52988.2021.9663742

Moving-Object-Aware Anomaly Detection in Surveillance Videos

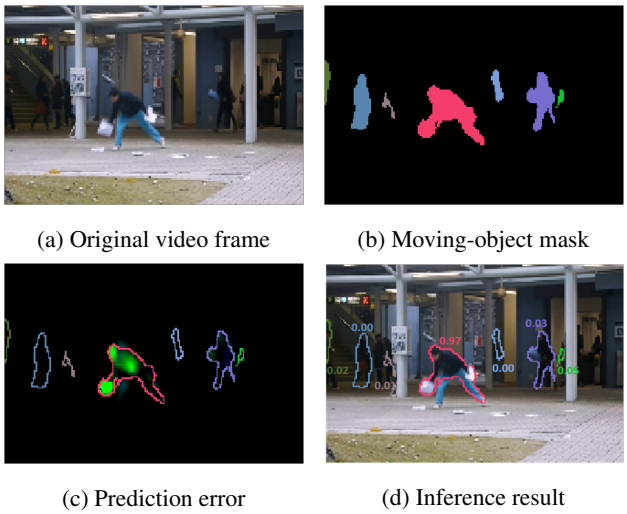
Chun-Lung Yang, Tsung-Hsuan Wu, and Shang-Hong Lai
Department of Computer Science, National Tsing Hua University, Taiwan
s108062583@m108.nthu.edu.tw, th.wu@mx.nthu.edu.tw, lai@cs.nthu.edu.tw

Abstract

Video anomaly detection plays a crucial role in automatically detecting abnormal actions or events from surveillance video, which can help to protect public safety. Deep learning techniques have been extensively employed and achieved excellent anomaly detection results recently. However, previous image-reconstruction-based models did not fully exploit foreground object regions for the video anomaly detection. Some recent works applied pre-trained object detectors to provide local context in the video surveillance scenario for anomaly detection. Nevertheless, these methods require prior knowledge of object types for the anomaly which is somewhat contradictory to the problem setting of unsupervised anomaly detection. In this paper, we propose a novel framework based on learning the moving-object feature prediction based on a convolutional autoencoder architecture. We train our anomaly detector to be aware of moving-object regions in a scene without using an object detector or requiring prior knowledge of specific object classes for the anomaly. The appearance and motion features in moving objects regions provide comprehensive information of moving foreground objects for unsupervised learning of video anomaly detector. Besides, the proposed latent representation learning scheme encourages the convolutional autoencoder model to learn a more convergent latent representation for normal training data, while anomalous data exhibits quite different representations. We also propose a novel anomaly scoring method based on the feature prediction errors of moving foreground object regions and the latent representation regularity. Our experimental results demonstrate that the proposed approach achieves competitive results compared with SOTA methods on three public datasets for video anomaly detection.

1. Introduction

Video surveillance has been commonly used for monitoring public safety, while manually inspecting abnormal events from surveillance videos is a tedious and exhaustive task. Automatic anomaly detection from surveillance video is expected to provide a fast and accurate solution, which has become highly demanded in practice. However, it is hard to give a clear definition for video anomalies or obtain a representative dataset for video anomalies for each scenario in practice since they are principally unknown events. Hence, this makes it very challenging to be solved by rule-based or supervised-learning-based approaches. Unsupervised anomaly detection approaches using a sample of the unlabeled data set as training data are most widely applicable for this task.



(a) Original video frame (b) Moving-object mask

(c) Prediction error (d) Inference result

Figure 1: Demonstration of the overall idea of our moving-object-aware video anomaly detection system. Our framework focuses on analysis of moving-object regions in a video, which provides object-level anomaly scoring in addition to the frame-level anomaly scoring for video anomaly detection. The inference result in (d) shows the associated anomaly scores for the moving objects.

For the past few years, there has been a considerable amount of progress in applying deep learning to the unsupervised anomaly detection problem. One typical approach is to train a convolutional autoencoder (Conv-AE) to learn regular reconstruction from normal video frames (termed

978-1-6654-3396-9/21/$31.00 ©2021 IEEE

"image-reconstruction-based methods" in this paper). The trained Conv-AE is expected to well describe normal conditions while generating deviations for abnormal events. Nevertheless, these image-reconstruction-based methods failed to exploit the fact that foreground objects are the most critical information to analyze for detecting an abnormal event, thus making them susceptible to background noise. Some latest works take advantage of the generalized models pre-trained on large video datasets to provide semantic information in a video to their anomaly detection model. These works aim to lead their anomaly detector focusing on specific categories of objects in the video scene. Object-centric methods [10, 2, 3, 5, 4] guide their anomaly detector to check for specific object regions provided by pre-trained object detection model. These methods took advantage of the well-informed pre-trained models, meanwhile checking straight into the object-level feature to avoid interference from the background, leading to good results under specific premises. The one crucial premise for these semantic-information-based methods is that the anomaly categories must be included in their pre-trained models.

To fully exploit the idea of determining video anomaly based on foreground objects, meanwhile properly follow the unsupervised setting of anomaly detection, we propose a moving-object feature prediction framework for video anomaly detection. The proposed anomaly detector is trained to be aware of moving objects defined by optical flow feature instead of specific categories of objects. With the optical flow feature, we can further define foreground-object regions from the movement characteristics, computing anomaly scores for each foreground object from the corresponding prediction errors. The basic idea of the proposed video anomaly detection system is illustrated in Figure 1. Inspired by the recent anomaly detection works [7, 22, 1] that emphasize regularity on latent representation for normal conditions, we train an additional variational autoencoder (VAE) with the Conv-AE concurrently to enforce the normal latent representations to be well reconstructed by the VAE. Eventually, we develop a frame-level anomaly scoring strategy to provide quantified measure for anomaly detection and give quantitative comparisons with recent SOTA methods. Our proposed frame-level scoring strategy considering moving-object-based feature prediction and latent representation achieves very competitive results on the three public benchmark datasets for video anomaly detection.

Our main contributions are summarized as follows:

- We apply optical flow features as moving foreground object feature descriptors on image space. We train a convolutional autoencoder model to predict pertinent moving object information while ignoring meaningless background noise.
- We propose a training scheme that applies an additional variational autoencoder to model the normal latent representation of the convolutional autoencoder.
- We develop a novel anomaly scoring strategy. By considering moving-object feature prediction and latent representation simultaneously.
- The proposed method achieves competitive results compared with SOTA methods on three public benchmark datasets for video anomaly detection.

The rest of this paper is organized as follows. In section 2, we review some related works for video anomaly detection. Subsequently, we describe the details of the proposed moving-object feature prediction framework in section 3. Then, we give some experimental results and discussion in section 4. Finally, we conclude the paper in section 5.

2. Related Work

2.1. Image-reconstruction-based methods

Image-reconstruction-based methods usually apply Conv-AE to reconstruct or predict video frames. The fundamental concept is that only normal conditions are seen during the training process so that the Conv-AE model can only reconstruct or predict images well for normal samples. On the contrary, the model reconstructs or predicts poorly for abnormal data. Thus, the image reconstruction results can be used to determine if an abnormal event occurs. Some works consider time factors by using several consecutive frames to predict the characteristics of subsequent frames [9, 14]. Besides, [21, 23, 14] computed optical flow to represent the motion features. Some methods suggest the use of adversarial learning [8] to help the generator to produce more realistic images [23, 6, 14].

Frame generation in these methods usually reconstructs the entire image scene and evaluates anomaly score from the reconstruction error of the whole image. This approach often takes the unimportant background noise into consideration. To alleviate the problem due to background noise, our proposed method is to predict moving foreground features without influence by background.

2.2. Semantic-information-based methods

Anomaly detection training is usually based on a short amount of normal data, yet surveillance videos contain highly complex and rich information. As deep learning models have demonstrated mature results in various computer vision applications, transfer-learning-based methods have been adopted. These works [18, 20, 10, 2, 3, 5, 4] proposed to use sophisticated models pre-trained on the large video datasets to provide semantic information of the videos. Georgescu *et al*. [4] exploited the pre-trained

YOLOv3 [24] to guide their anomaly detector to learn multiple proxy tasks in the bonding boxes. The regularity of human pose also provides an important cue for the video anomaly detection task [20, 18]. Morais *et al*. [20] first extracted human skeleton points from video frames by a pre-trained human pose estimation model. Therefore, the corresponding skeleton points of successive frames are connected into dynamic skeleton features, and abnormal human events are detected by checking the regularity of skeleton trajectories.

This type of approach takes advantage of some interpretable information provided by some pre-trained models; meanwhile, it filters out some redundant background information in the video scene for the anomaly detector model in advance. However, they require a crucial premise that anomaly categories must be included in their pre-trained models. In practical, which classes should be selected for the object detector to detect is tricky. Strong prior knowledge of anomaly categories makes them contradictory to unsupervised tasks. To make the anomaly categories keeping "unknown" in our unsupervised learning framework, we choose to define the foreground region by the movement feature in the scene.

2.3. Latent-representation-based methods

Although CNN autoencoder has been widely used in the image-reconstruction-based methods, some observed good reconstruction results for anomalous cases due to the powerful representation capacity of the CNN model. Since the model may not always induce large reconstruction errors for anomalous samples, this leads to missed anomaly detection. To prevent this, some recent works [7, 22, 1] proposed to restrict divergence of latent representations in autoencoder during the training process, encouraging close binding for latent representations of normal samples. Park *et al*. [22] applied a memory module to memorize normal patterns in the latent space for Conv-AE. For their update strategy of the memory, they record prototypical items of normal data. Counting similarity between prototypical items and encoded latent queries helps to identify anomalies. Chang *et al*. [1] exploited the deep k-means clustering to force normal latent representation to be clustered.

Even though the nonlinearity of neural network models makes them less interpretable, the highly complex models normally provide better solutions to many unsupervised learning problems. The traditional modeling approach may not necessarily provide the best solution to address the high-level coding embedding, such as latent queries in Conv-AE. We apply a similar concept in image-reconstruction-based anomaly detection methods onto the latent query domain. This is accomplished by applying an additional variational autoencoder to learn the reconstruction of normal latent queries of Conv-AE in the training process to restrict its latent representation for normal cases.

3. Proposed Method

We implement our moving-foreground-object-aware concept with a moving-foreground-object feature prediction framework. Our framework contains two regularity models trained from normal videos: a foreground-object-aware convolutional autoencoder (FOA-CAE) and a latent-query-restricting variational autoencoder (LQR-VAE). FOA-CAE encodes consecutive video frames, and then predicts foreground-object features in the future frame, while LQR-VAE restricts manifold divergence of FOA-CAE's normal latent representation. For detecting anomalous events in videos, we design an anomaly scoring strategy with focus on moving object region and latent representation. Figure 2 illustrates an overview of the proposed framework. The following subsections will describe details for the framework.

3.1. Moving-foreground-object feature prediction

To prevent the regularity learning in image space from being influenced by the color bias, we convert the original RGB frame images $I_1, I_2, ..., I_t$ to resized grayscale images $X_1, X_2, ..., X_t$, where t represents each time step in a video. After concatenating 3 consecutive pre-processed images X_{t-2}, X_{t-1}, X_t on temporal dimension, we obtain a 4D tensor $X_{t-2,t-1,t}$ with size $128\times192\times1\times3$ of $H\times W\times C\times T$. With temporal information implied tensor $X_{t-2,t-1,t}$ as the input of our FOA-CAE model, it is trained to predict two future foreground features A_{t+1}, M_{t+1}, which represent appearance and motion information for moving objects, respectively. For the motion feature M_{t+1}, we first feed I_t, I_{t+1}, and I_{t+2} into a pre-trained optical flow prediction model SelFlow [13] and resize the visualization result of optical flow output to obtain the motion feature M_{t+1}, which is an RGB image. The motion information is implied in the color of M_{t+1}. To be more specific, saturation and value imply the magnitude of pixel movement, and hue states the angle of movement direction. Regarding the appearance feature A_{t+1}, motion feature M_{t+1} is first adopted to provide moving object region. By applying thresholding on the motion feature M_{t+1}, we can obtain a binary mask that distinguishes foreground from background regions in an image by the movement between two consecutive frames. Then we multiply the binary motion mask with X_{t+1} pixel-wisely and copy three times of the multiplied result image on channel dimension to obtain the foreground appearance feature A_{t+1} with size $128 \times 192 \times 3$ of $H \times W \times C$. Eventually, we get the appearance feature A_{t+1} which preserves the appearance details for foreground objects while ignoring the meaningless background appearance in the scene.

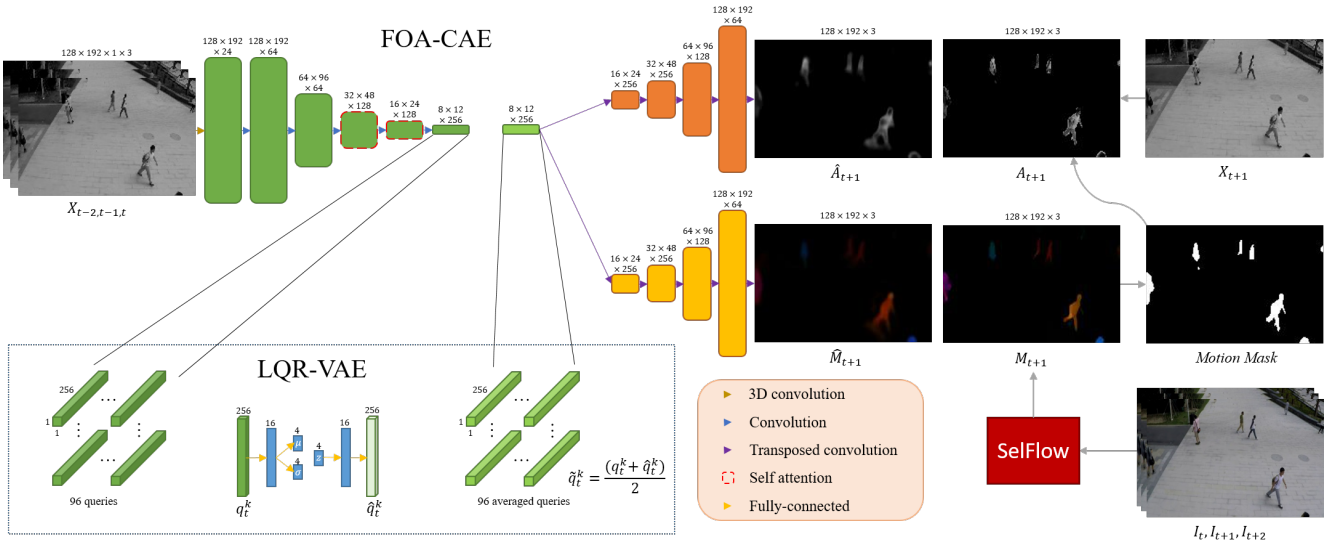


Figure 2: Proposed moving-foreground-object feature prediction framework.

3.2. FOA-CAE

U-Net [25] architecture has been extensively used on image-reconstruction-based methods [23, 14, 21]. Skip connections of U-Net provide the decoder with the feature maps from the encoder which contains low-level information of input and enhance generation ability by shortcuts. Our approach tends to take advantage of the encoding ability of the encoder rather than the generation capability of the model. To avoid weakening encoding ability caused by shortcut mechanism of skip connections [22], we exploit relatively simple Conv-AE as our main model architecture for FOA-CAE. There is one encoder for encoding $X_{t-2,t-1,t}$, two decoders for outputting A_{t+1} and M_{t+1}, respectively.

We follow some philosophy of designing the basic Conv-AE architecture from [21] yet lessening nearly half of the model capacity from their model. Moreover, we further ameliorate our model with self-attention mechanism to make it more suitable to be moving-object-aware.

The basic architecture of the encoder consists of four convolution blocks (Conv-block). Each Conv-block contains three layers: 3×3 convolution with stride 2, batch-normalization, and leakyReLU activation. There are two primary modifications for the encoder. For the first modification, before four basic Conv-block, there is one 3D convolution layer to capture the spatiotemporal correlations. Then, we convert the 4D tensor to 3D feature maps by applying average pooling on the temporal dimension; this allows the 2D convolution layer to follow, saving the overall space and time cost of FOA-CAE. A Conv-block without striding is attached to enhance the diversity of the feature extraction in low-level space. The other modification is that we attach the self-attention module from [27] on the middle two Conv-blocks of the basic encoder structure. The information of dependencies at different positions provided by the adaptive attention map bridges long-range dependencies for any two positions of the feature maps with the non-linear transformation. This alleviates the limitation of convolutional layers' nodes only computing from a small local neighborhood of nodes, helping FOA-CAE better understand long-range information in a scene.

The decoders have a relatively simple structure with four deconvolution blocks (DeConv-block). Each DeConv-block contains four layers: 3×3 deconvolution with stride 2, batch-normalization, ReLU activation, and dropout. One deconvolution layer without striding with 3 channels is attached to the decoder's end, so that each decoder outputs one RGB image, $\hat{A}_{t+1}$ and $\hat{M}_{t+1}$. To guide the FOA-CAE to learn predicting normal future foreground feature images, we minimize the l_2 distance of pixel intensities between predicted feature images and the corresponding ground truth images:

$$\begin{cases} \mathcal{L}_{appearance} = \|A_{t+1} - \hat{A}_{t+1}\|_2^2 \\ \mathcal{L}_{motion} = \|M_{t+1} - \hat{M}_{t+1}\|_2^2 \end{cases} \quad (1)$$

The image prediction loss is defined as the weighted summation of two losses given above:

$$\mathcal{L}_{image} = \lambda_a \mathcal{L}_{appearance} + \lambda_m \mathcal{L}_{motion} \quad (2)$$

3.3. LQR-VAE

We apply VAE instead of AE to model the normal latent query of FOA-CAE because VAE encodes input as

distributions rather than vectors. This benefits our framework by applying fewer parameters while preventing overfitting during modeling plenty of high-level queries. A query map Q_t of size $H \times W \times C$ extracted by the encoder of FOA-CAE is disassembled to queries $q_t^1, q_t^2, ..., q_t^K$ with size $1 \times 1 \times C$, and the number of queries $K = H \times W$, where $K = 96$ in our implementation. Similar concept to image-reconstruction-based anomaly detection methods, VAE learns to reconstruct the input query without seeing abnormal samples during the training progress, thus it models normality. We apply l_2 loss with KL-divergence term provided by the original VAE paper [12] as our query loss:

$$\mathcal{L}_{query} = \frac{1}{K}\sum_{k=1}^{K}(\|q_t^k - \hat{q}_t^k\|_2^2 - \frac{1}{2}(1 + \log((\sigma_t^k)^2) - (\mu_t^k)^2 - (\sigma_t^k)^2)) \tag{3}$$

To restrict the encoded latent query divergence in FOA-CAE, we train the low capability LQR-VAE with FOA-CAE, simultaneously. Instead of simply inputting the original query map Q_t to the decoders of FOA-CAE, we input the query map $\tilde{Q}_t$ assembled by the averaged queries $\tilde{q}_t^1, \tilde{q}_t^2, ..., \tilde{q}_t^K$ of original queries $q_t^1, q_t^2, ..., q_t^K$ and the reconstructed queries $\hat{q}_t^1, \hat{q}_t^2, ..., \hat{q}_t^K$. The two models share one single optimizer for end-to-end training, and the objective function can be formulated as:

$$\mathcal{L} = \lambda_i \mathcal{L}_{image} + \lambda_q \mathcal{L}_{query} \tag{4}$$

3.4. Anomaly scoring strategy

Most previous image-reconstruction-based methods evaluate the model generated result by, *e.g.*, peak signal-to-noise ratio, structural similarity, or mean squared error of the whole image for frame-level anomaly scoring. Nguyen *et al.* [21] proposed a patch-based scheme of anomaly score estimation to reduce the effect of noise. To ignore the irrelevant background noise and to consider the importance of normalization for each moving object more precisely, we propose a novel anomaly scoring strategy highlighting moving object regions.

Unlike subordinating the specific object categories' location produced by the object detector model in object-centric methods, we define the moving foreground object regions by the reference of motion feature. We first apply thresholding on M_{t+1} to obtain the binary motion mask from the motion feature image. Then, we exploit the connected component algorithm [26] on the motion mask to separate different components and filter out the components with tiny areas to obtain the moving object mask O_{t+1} for the frame scene. With the moving object mask O_{t+1}, we can compute partial scores for each moving object $o \in O_{t+1}$ as follows:

$$\begin{cases} S_{appearance}(o) = \sqrt{\frac{1}{|o|}\sum_{(i,j)\in o}(A_{i,j} - \hat{A}_{i,j})^2} \\ S_{motion}(o) = \sqrt{\frac{1}{|o|}\sum_{(i,j)\in o}(M_{i,j} - \hat{M}_{i,j})^2} \end{cases} \tag{5}$$

where o denotes a moving object region and $|o|$ is its number of pixels. Without carefully selecting weights between partial scores, we multiply them directly as our object-level anomaly score:

$$S_{object}(o) = S_{appearance}(o) \times S_{motion}(o) \tag{6}$$

For frame-level image prediction anomaly score, we take the highest object-level anomaly score as the representative score for the frame, i.e.

$$S_{image} = \max_{o \in O_{t+1}} S_{object}(o) \tag{7}$$

For latent code reconstruction anomaly score, we consider root mean squared error and cosine-distance between Q_t and reconstructed $\hat{Q}_t$:

$$S_{query} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}({q_t}^k - {\hat{q}_t}^k)^2} + \frac{1}{K}\sum_{k=1}^{K}(1 - Cos({q_t}^k, {\hat{q}_t}^k)) \tag{8}$$

For final frame-level anomaly score, we multiply the normalized frame-level image prediction anomaly score with normalized latent code reconstruction anomaly score:

$$S_t = g(S_{image}) \times g(S_{query}) \tag{9}$$

Most previous video anomaly detection methods compute normalized scores for each video defined by the standard video datasets. We argue that it is ambiguous to define different videos in real application but feasible to define camera scenes, so we normalize the scores by camera scenes. Function $g(.)$ denotes the min-max normalization over whole frames in a camera scene: Considering the characteristics of continuity for events in the video, we apply Gaussian filtering for temporally smoothing the frame-level anomaly scores in the final stage. The scores assessed from frames containing abnormal events are expected to be higher than normal ones.

4. Experiments

4.1. Datasets

We perform our experiments on the three most widely used public benchmark datasets; namely, UCSD Ped2 dataset [17], CUHK Avenue dataset [15], and ShanghaiTech dataset [16]. These datasets are set for unsupervised anomaly detection problems with videos captured by static

surveillance cameras in the campus. Each dataset has its own characteristics.

UCSD Ped2 dataset includes 16 training and 12 testing videos. Grayscale and relatively low-resolution bird's-eye view videos of a crowded sidewalk are provided. The abnormal events are mostly about transportation-related violations on a sidewalk, such as cycling and skateboarding.

CUHK Avenue dataset contains 16 training and 21 testing videos. The dataset depicts an eye-level shot screen in front of a subway station. The irregular events are human behaviors like running, doing exercise, and throwing stuff.

ShanghaiTech dataset contains 330 training and 107 testing videos with 13 different scenes. Different camera angles and shots in different scenes, as well as diverse anomaly event types, make it considered one of the most challenging datasets among the three public datasets commonly used for video anomaly detection research.

4.2. Implementation details

We implement our work by using Tensorflow 2.0 framework and Python with an NVIDIA RTX 2080 Ti GPU. All images are normalized to the range from 0 to 1. We tend to find a more universal solution for all datasets, so most of the hyperparameter settings are the same for 3 datasets. The FOA-CAE model and LQR-VAE model share one single Adam optimizer [11] with learning rate set 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and batch size set 32 for end-to-end training. For training loss, the weights are set to $\lambda_a = 0.5$, $\lambda_m = 0.5$, and $\lambda_i = 1$ for three datasets. For the weighting of query loss λ_q are set 0.5, 0.005, 0.0005 for UCSD Ped2 dataset, CUHK Avenue dataset, and ShanghaiTech dataset, respectively since the diversity difference between these three datasets. Training epochs are set differently for each dataset since different abundance for the datasets: 200 for UCSD Ped2 dataset, 80 for CUHK Avenue dataset, and 8 for ShanghaiTech dataset. For the pre-trained optical flow prediction model, we apply the SelFlow model [13], which was pre-trained on the KITTI 2015 dataset [19]. The system's inference time is 16 fps (detailed in the supplementary materials).

4.3. Comparison with the state-of-the-art

To evaluate our method, we measure frame-level area under curve (AUC) of the receiver operating characteristic (ROC) curve, a standard metric for video anomaly detection (detailed in the supplementary materials). Quantitative results of our method and the state-of-the-art methods for video anomaly detection are shown in Table 1. We use different scopes to separate different prior knowledge condition of anomaly type been utilized in the method. Methods in the "Object" scope outperform other methods via strong local information. By being aware of moving object features, our proposed method achieves the best result in the 'Image' scope on UCSD Ped2 and ShanghaiTech datasets, gaining 0.7% and 2.3% on UCSD Ped2 and ShanghaiTech, respectively. And 87.8% AUC score on Avenue dataset is also a competitive performance. Worth mention that ShanghaiTech is the most challenging dataset that contains videos with variant camera scenes. Instead of training independent FOA-CAE and LQR-VAE for each scene, we train the general model for all 13 scenes and achieve the best performance comparing with other state-of-the-art methods. The adaptability to different scenes demonstrates that our method is robust against background appearance variations.

	Method	UCSD	CUHK	Shanghai
Pose	Morais *et al*. [20]	-	**86.3**	73.3
	Markovitz *et al*. [18]	-	-	**76.1**
Object	Ionescu *et al*. [10]	94.3	87.4	78.7
	Doshi *et al*. [2], [3]	97.8	86.4	71.6
	Georgescu *et al*. [4]	97.5	91.5	82.4
	Georgescu *et al*. [5]	**98.7**	**92.3**	**82.7**
Image	Conv2D-AE [9]	85.0	80.0	60.9
	TSC [16]	91.0	80.6	67.9
	StackRNN [16]	92.2	81.7	68.0
	Liu *et al*. [14]	95.4	85.1	72.8
	Nguyen *et al*. [21]	96.2	86.9	-
	MemAE [7]	94.1	83.3	71.2
	Mem-guide [22]	**97.0**	**88.5**	70.5
	Cluster-drive [1]	96.5	86.0	**73.3**
	Ours	**97.7**	**87.8**	**75.6**

Table 1: Quantitative comparison with SOTA methods for video anomaly detection. Frame-level AUC scores (%) on UCSD Ped2, CUHK Avenue, and ShanghaiTech dataset.

4.4. Qualitative results

Figure 3 depicts some qualitative experimental results on the three datasets. We can see from the predicted $\hat{M}$ that our FOA-CAE can well predict the normal objects while leading to large predicting errors for the abnormal objects. The proposed model can detect a variety of anomaly event types. In the figure, the prediction errors associated with the moving object contours demonstrate that the proposed model can accurately detect the anomalous objects from the videos.

4.5. Ablation study

We evaluate different components in our framework on CUHK Avenue dataset to observe each component's contribution to the model performance. Table 2 shows the AUC score for the quantitative ablation study. The first two rows show the results for one-stream autoencoder trained to predict appearance and motion, respectively, that promis-

(a) (b) (c) (d) (e) (f)

Figure 3: Qualitative results on UCSD Ped2 (a), CUHK Avenue (b)(c), and ShanghaiTech (d)-(f). Each example displays 4 images: the original video frame, pseudo ground truth motion feature M, predicted $\hat{M}$, and the corresponding prediction error (from top to bottom).The anomaly events are (a) unexpected bicycle and truck on the sidewalk, (b) throwing papers, (c) jumping, (d) chasing, (e) falling down, and (f) exhibition drill.

ing results are produced by both settings. The third row shows that the two-stream autoencoder predicting both appearance and motion feature provides better performance since more information is used to describe the moving foreground objects. The fourth to sixth rows demonstrate improved accuracy after including LQR-VAE to the settings for the first three rows. Although the result without LQR-VAE is already outstanding on the two-stream framework, it gains 0.7% AUC improvement after applying LQR-VAE, which indicates the effectiveness of the LQR-VAE module. When training the complete framework but only considering S_{query} for frame-level anomaly score (the seventh row), it can achieve an 80.2% AUC score, which again proves the reliability of the proposed LQR-VAE for anomaly detection. We also test to exchange the latent-query-restricting variational autoencoder (LQR-VAE) to an autoencoder model (named LQR-AE on the last two rows). The experimental result shows better performance for the VAE model in our latent representation learning task.

Appearance	Motion	LQR-VAE	LQR-AE	AUC
✓	✗	✗	✗	83.1%
✗	✓	✗	✗	84.9%
✓	✓	✗	✗	87.0%
✓	✗	✓	✗	85.2%
✗	✓	✓	✗	85.9%
✓	✓	✓	✗	87.7%
▲	▲	✓	✗	80.2%
✓	✓	✗	✓	86.4%
▲	▲	✗	✓	78.4%

Table 2: The evaluation of different components in our framework on the CUHK Avenue dataset. ▲ denotes the corresponding score is not used for the final scoring.

5. Conclusions

Since foreground-object information plays an important role in detecting video anomaly events, we focus on how to incorporate moving-foreground-object information into the video anomaly detection framework in this paper. With optical flow information of moving foreground objects provided for our proposed FOA-CEA's training, it learns better descriptions of normal events from the videos. We also present an effective unsupervised training scheme that leverages the proposed LQR-VAE module to learn the normal representations for the normal latent queries of a Conv-AE model. In addition, we developed a novel frame-level scoring strategy that considers both latent representation regularity and moving foreground objects based on an object-level scoring scheme. Experimental results demonstrate that the proposed method can achieve SOTA performance on the three primary benchmark public datasets for unsupervised video anomaly detection. The inference time of our proposed framework is about 16 frames per second, which is reasonable for practical application.

References

[1] Y. Chang, Z. Tu, W. Xie, and J. Yuan. Clustering driven deep autoencoder for video anomaly detection. In *European Conference on Computer Vision*, pages 329–345. Springer, 2020.

[2] K. Doshi and Y. Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 934–935, 2020.

[3] K. Doshi and Y. Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 254–255, 2020.

[4] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12742–12752, 2021.

[5] M. I. Georgescu, R. Ionescu, F. S. Khan, M. Popescu, and M. Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[6] T. Golda, N. Murzyn, C. Qu, and K. Kroschel. What goes around comes around: Cycle-consistency-based short-term motion prediction for anomaly detection using generative adversarial networks. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[7] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. 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.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[9] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[10] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] P. Liu, M. Lyu, I. King, and J. Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019.

[14] W. Liu, W. Luo, D. Lian, and S. 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.

[15] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[16] W. Luo, W. Liu, and S. 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.

[17] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010.

[18] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10539–10547, 2020.

[19] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015.

[20] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. 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.

[21] T.-N. Nguyen and J. Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1283, 2019.

[22] H. Park, J. Noh, and B. Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020.

[23] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[24] J. Redmon, S. Divvala, R. Girshick, and A. 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.

[25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[26] K. Wu, E. Otoo, and K. Suzuki. Optimizing two-pass connected-component labeling algorithms. *Pattern Analysis and Applications*, 12(2):117–135, 2009.

[27] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.