A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction

Zhian Liu[1], Yongwei Nie[1*], Chengjiang Long[2], Qing Zhang[3] and Guiqing Li[1]

[1]School of Computer Science and Engineering, South China University of Technology, China
[2]JD Finance America Corporation, Mountain View, CA, USA
[3]School of Computer Science and Engineering, Sun Yat-sen University, China

Abstract

In this paper, we propose HF^2-VAD, a Hybrid framework that integrates Flow reconstruction and Frame prediction seamlessly to handle Video Anomaly Detection. Firstly, we design the network of ML-MemAE-SC (Multi-Level Memory modules in an Autoencoder with Skip Connections) to memorize normal patterns for optical flow reconstruction so that abnormal events can be sensitively identified with larger flow reconstruction errors. More importantly, conditioned on the reconstructed flows, we then employ a Conditional Variational Autoencoder (CVAE), which captures the high correlation between video frame and optical flow, to predict the next frame given several previous frames. By CVAE, the quality of flow reconstruction essentially influences that of frame prediction. Therefore, poorly reconstructed optical flows of abnormal events further deteriorate the quality of the final predicted future frame, making the anomalies more detectable. Experimental results demonstrate the effectiveness of the proposed method. Code is available at https://github.com/LiUzHiAn/hf2vad.

1. Introduction

Video Anomaly Detection (VAD) refers to the identification of events that do not conform to expected behaviors [3] in a video, with one example shown in Figure 1. This is an open and very challenging task as abnormal events usually much less happen than normal ones and the forms of abnormal events are unbounded in practical applications [25]. Obviously, it is impossible to collect all kinds of abnormal data in advance. Therefore, a typical solution to video anomaly detection is to train an unsupervised learning model on normal data, and those events or activities that are recognized by the trained model as outliers are then deemed as anomalies.
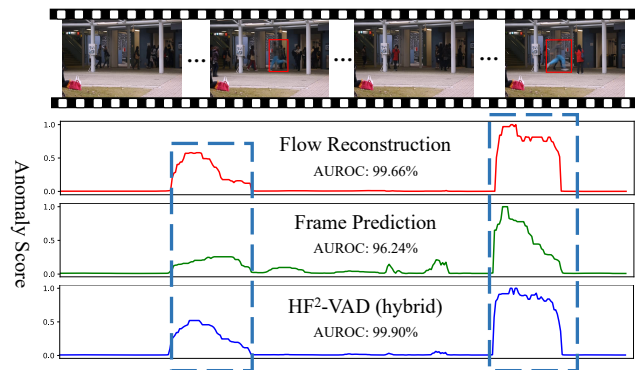
*Corresponding author: nieyongwei@scut.edu.cn

Figure 1. An anomaly example from CUHK Avenue [29] dataset. Here "*running*" is identified as an anomaly event on the walking avenue, and the blue dashed rectangles denote ground-truth abnormal sections where a person is running. The blue curve is the result of HF^2-VAD, the red curve is the result with flow reconstruction only, and the green curve is the result with a CVAE based frame prediction model conditioned on original flows. Area Under the Receiver Operation Characteristic (AUROC) is calculated. As can be seen, HF^2-VAD that combines flow reconstruction and reconstructed-flow guided future prediction performs the best.

Nowadays deep learning has shown great success in many real-world tasks such as visual recognition [13, 12, 14, 26], object detection [19, 28, 27, 15], shadow detection and removal [4, 44, 49, 48, 18], trajectory prediction [41], and image captioning [5]. Rather than traditional hand-crafted feature based methods [1, 20, 2, 34], a lot of modern deep neural network based methods [11, 45, 50, 31, 25, 30, 8, 36, 38, 47, 43, 24] have been proposed for VAD. In the era of deep learning, reconstruction and future prediction are two prevalent VAD paradigms. Reconstruction-based methods [11, 31, 8, 36, 7, 38] typically train autoencoders on normal data. At test time, abnormal data often incurs larger reconstruction errors, making them detectable from normal ones. Taking advantage of temporal characteristics of video frames, prediction-based methods [25, 30, 47] train a net-

work to predict the next frame based on the given previous frames and use prediction error for anomaly measuring. Recently, several works [50, 46, 35] are proposed to combine these two paradigms in a hybrid manner. Although these state-of-the-art methods have been able to detect anomalies in most cases, the results are still far from perfect.

In this paper, we propose a novel *hybrid* framework in a combination of *flow* reconstruction and flow-guided *frame* prediction, named as "HF^2-VAD", for *video anomaly detection*. As illustrated in Figure 2, the Conditional VAE (CVAE) based future frame prediction model accepts both previous video frames and optical flows as input. But instead of original flows, we reconstruct them in advance and then input the reconstructed flows into the CVAE model.

Inspired by [8, 38], we design a Multi-Level Memory-augmented Autoencoder with Skip Connections (ML-MemAE-SC) for optical flow reconstruction. Multiple memory modules are employed to memorize normal patterns at different feature levels, while skip connections are added between encoder and decoder to compensate for the strong information compression due to the memories. We observe that such a well-designed flow reconstruction network can reconstruct normal flows more clearly while producing larger reconstruction error for abnormal input.

We use the model of CVAE [42] for future frame prediction. On one hand, it takes the reconstructed flows by ML-MemAE-SC as condition, unifying the reconstruction module into the prediction pipeline naturally. On the other hand, by maximizing the quantity of evidence lower bound (ELBO) induced from the variables of observed video frames and reconstructed optical flows, the CVAE module essentially encodes the consistency between the input frames and flows when taking them for the future frame prediction.

The above design facilitates to utilize the quality gap between the reconstructed normal and abnormal flows to improve the VAD accuracy of the CVAE-based prediction module. That is, the reconstructed normal flows usually have higher quality, with which the prediction module can successfully predict the future frame with smaller prediction error. In contrast, the reconstructed abnormal flows usually have lower quality, thus leading to future frame with larger prediction error. We use both the flow reconstruction and frame prediction errors as our final anomaly detection cues. In summary, the following aspects distinguish our work from the previous works [25, 8, 7, 47, 36, 30, 47]:

- First of all, multi-level memory modules are utilized in an encoder-decoder structure with skip connections, which guarantees normal patterns are well memorized so that the abnormal events or activities are sensitively identified.

- Second, we design the HF^2-VAD hybrid method that

predicts future frame from both previous video frames and the corresponding optical flows but with the flows being reconstructed beforehand. The reconstruction error enlarges the prediction error so that anomaly can be more easily detected.

- Finally, we conduct extensive experiments on three public datasets which show that our proposed HF^2-VAD achieves better anomaly detection performance than state-of-the-art methods.

2. Related work

Video anomaly detection. Feature extraction followed by a normality modeling was a popular paradigm in early works [40, 1, 20]. With the powerful feature representation capability of CNNs, many deep learning based VAD methods have been proposed in recent years. Reconstruction-based and prediction-based methods are the two mainstreams in the VAD community. In the reconstruction camp, autoencoders (AEs) are widely leveraged to reconstruct the training data, such as Conv-AE [11], ConvLSTM-AE [31] and GMFC-VAE [7]. They assume the anomalies are hard to be reconstructed well by the AE trained only on normal data. But this assumption not always holds since AE sometimes generalizes so well [8]. In the prediction camp, Liu *et al.* [25] first propose to predict the future frame and use prediction error as the anomaly indicator. Lu *et al.* [30] then introduce a variational Conv-LSTM for prediction. Yu *et al.* [47] borrow the cloze test concept in language study and apply it to video event completion, in which multiple models are built to predict each frame or flow in a clip separately. However, these methods simply take previous frames as input to predict the future, ignoring the correlation between optical flow and video frame. Further, some works joint these two paradigms together to develop hybrid approaches. For example, an AE comprising a shared encoder and two separate decoders is proposed in [36], to reconstruct frame and predict flow separately. Ye *et al.* [46] decompose reconstruction into prediction and refinement, proposing a predictive coding network. In our work, we integrate flow reconstruction and frame prediction seamlessly, making the quality of flow reconstruction essentially influence that of frame prediction.

Memory networks. Memory module in neural networks has attracted much attention recently. Graves *et al.* [10] introduce a differentiable neural computer model which consists of a neural network to extract features and an external memory module to store information explicitly. In order to suppress the generalization capability of AE, Gong *et al.* [8] propose a Memory-augmented Autoencoder (MemAE) for anomaly detection. MemAE receives information from the encoder and then uses it as a query to retrieve some similar memory slots which are then combined to yield new encoding features for the decoder to reconstruct. The MemAE
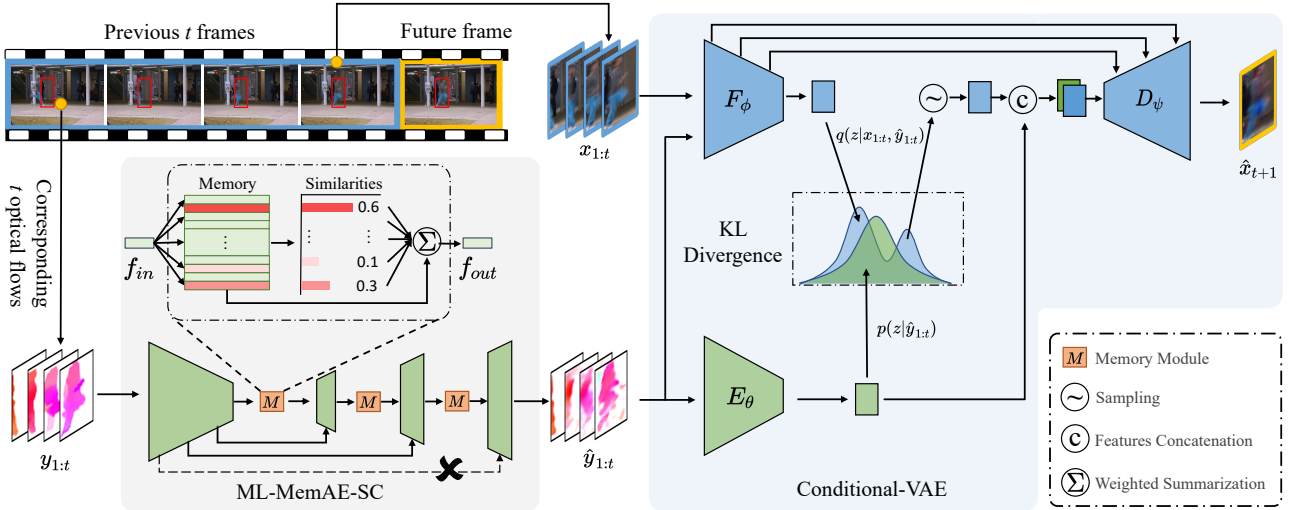
Figure 2. Overview of the proposed HF^2-VAD which integrates flow reconstruction and frame prediction into a unified framework. We first reconstruct the optical flows $y_{1:t}$ by an autoencoder with multi-level memory modules and skip connections to obtain $\hat{y}_{1:t}$. Then, the reconstructed optical flows together with the video frames $x_{1:t}$ are used by a CVAE model to predict the next future frame. If an abnormal event occurs, (1) the reconstructed optical flows $\hat{y}_{1:t}$ will show significant reconstruction error to its ground truth $y_{1:t}$, (2) using the $\hat{y}_{1:t}$ as a condition to guide future frame prediction, the prediction error (*i.e.*, difference between $\hat{x}_{t+1}$ and x_{t+1}) will be enlarged further.

is trained on normal data, thus encouraged to store normal patterns in the memory. Park *et al*. [38] follow this trend and present a more compact memory that can be updated during testing. In their works, the memory module is only placed at the bottleneck. We extend it and propose a multi-level memory-augmented autoencoder with skip connections which captures the normal patterns at different feature levels and train it on optical flows.

VAE and CVAE. Along with the advances in discriminative models, generative models have also made great progress, such as GAN [9] and VAE [22]. In particular, VAE is a directed graphical model with latent variables, which includes a recognition process and a generative process. In the generative process (*i.e.*, decoding), the data x is generated by the distribution $p_\theta(x|z)$ when given latent variables z. Kingma and Welling [22] then introduce a recognition process $q_\phi(z|x)$ to approximate the intractable true posterior. Both the recognition and generative distributions can be learned by maximizing the log-likelihood of the $p(x)$ through the surrogate evidence lower bound (ELBO) objective function. To address the structured prediction problem, Sohn *et al*. [42] model the distribution of the output space conditioning on the input observation and propose CVAE. Let x, y and z represent the output data, observed condition and latent variables, CVAE is composed of a recognition network $q_\phi(z|x,y)$, a conditional prior network $p_\theta(z|y)$ and a generation network $p_\theta(x|y,z)$. Esser *et al*. [6] follow the CVAE framework and design a variational UNet that can disentangle the appearance and shape information of an image by which the image generating process can be well controlled. Different from them, we propose to predict future frame by CVAE with optical flow as the condition.

3. Methodology

As illustrated in Figure 2, our framework HF^2-VAD is composed of two components: Multi-Level Memory-augmented Autoencoder with Skip Connections (ML-MemAE-SC) for flow reconstruction followed by Conditional Variational Autoencoder (CVAE) for frame prediction. The whole framework is trained on normal data. At test time, both the reconstruction and prediction errors, *i.e.*, the difference between $y_{1:t}$ and $\hat{y}_{1:t}$, and x_{t+1} and $\hat{x}_{t+1}$, are used for anomaly detection.

In the following sections, we introduce ML-MemAE-SC first, and then the CVAE based future frame prediction model. Finally, we show how to use our model for anomaly detection.

3.1. Multi-Level Memory-augmented Autoencoder with Skip Connections

Placing a memory module at the bottleneck of AE is a recent development in VAD community [8, 38]. Figure 3 (a) shows this kind of Memory-augmented Autoencoder (MemAE). However, we observe that using just one memory can not guarantee all the normal patterns to be remembered, and abnormal input still has a certain chance to be reconstructed well. A natural extension to MemAE is placing more memory modules in other levels of AE as shown in Figure 3 (b), but too many memories lead to excessive information filtering, degrading the network to remember the most representative normal patterns rather than all needed ones. We solve this problem by adding skip connections between encoder and decoder, obtaining the Multi-Level Memory-augmented Autoencoder with Skip Connections (ML-MemAE-SC), as shown in Figure 3 (c). On one hand,
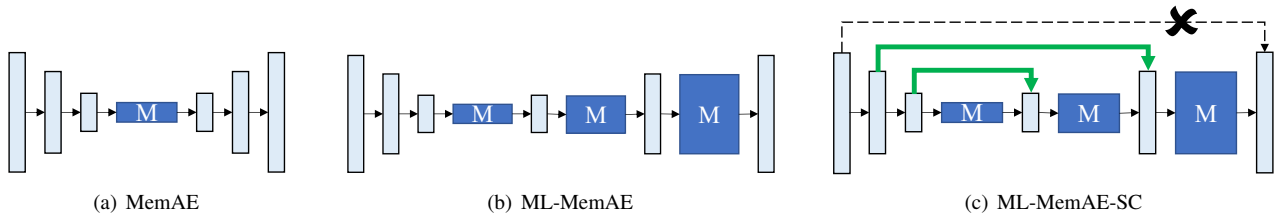
| (a) MemAE | (b) ML-MemAE | (c) ML-MemAE-SC |

Figure 3. (a) Structure of MemAE in which a memory module is placed at the bottleneck. (b) Extending MemAE with more memory modules in other levels of the decoder. (c) On the basis of (b), skip connections are added, but the black dotted skip connection should not be added, otherwise the functions of all the memory modules would be overridden.

the skip connections directly transfer encoding information to the decoder, providing more information for memories in different levels to discover normal patterns. On the other hand, with higher-level encoding features, though being filtered by the memories, the network can decode the input more easily. At test time, the proposed ML-MemAE-SC can reconstruct normal data clearly while performs poorly for abnormal data. To make it easy for readers to validate this, we conduct a toy example to explore many memory-augmented autoencoder variants and demonstrate the efficacy of the proposed ML-MemAE-SC. See Figure 4.

It is worth noting that the outermost skip connection, *i.e.*, the black dotted one in Figure 3 (c), should not be added. Otherwise the reconstruction might be fulfilled by the highest-level encoding-decoding information, making all other lower-level encoding, decoding and memory blocks not work.

We design a four-level ML-MemAE-SC, including three encoding-decoding levels and the bottleneck. In each level of the encoder, we stack two convolution blocks followed by a downsampling layer. In each level of the decoder, we first copy the feature map from the encoder and then concatenate it with the upsampled feature maps of the lower level. The concatenation then sequentially passes through two convolution blocks, a memory module, and an upsampling layer. In our implementation, a convolution block contains three layers: a convolution layer, a batch-normalization layer [17] and a ReLU activation layer [33]. The downsampling and upsampling layers are implemented by convolution and deconvolution [37].

For the memory modules, we adopt a similar implementation of [8]. Each memory module is actually a matrix $\mathbf{M} \in \mathbb{R}^{N \times C}$. Each row of the matrix is called a slot $\mathbf{m}_i$ with $i = 1, 2, 3, ..., N$. The role of a memory module is to represent the features fed into it by the weighted sum of similar memory slots, thus has the capability of remembering normal patterns when trained on normal data.

To train ML-MemAE-SC, we can feed normal video, image or optical flow into it, and try to reconstruct the input data. Let y be the input data, and $\hat{y}$ be the reconstructed result, we minimize the ℓ_2 distance between y and $\hat{y}$ as the reconstruction loss:

$$\mathcal{L}_{recon} = ||y - \hat{y}||_2^2. \tag{1}$$

Following [8], we add the entropy loss on the matching

probabilities $\hat{w}_i$ for each memory module as:

$$\mathcal{L}_{ent} = \sum_{i=1}^{M} \sum_{k=1}^{N} -\hat{w}_{i,k} \log \left(\hat{w}_{i,k} \right), \tag{2}$$

where M is the number of memory modules and $\hat{w}_{i,k}$ is the matching probabilities for the k-th slot in the i-th memory module. We balance the above two loss functions to obtain the following loss function to train ML-MemAE-SC:

$$\mathcal{L}_{ML-MemAE-SC} = \lambda_{recon}\mathcal{L}_{recon} + \lambda_{ent}\mathcal{L}_{ent}. \tag{3}$$

3.2. Conditional Variational Autoencoder for Future Frame Prediction

Future frame prediction is another prevalent VAD paradigm, often obtaining better anomaly detection accuracy than reconstruction-based methods [47, 38]. Future frame prediction tries to model $p(x_{t+1}|x_{1:t})$ such that the next frame x_{t+1} can be generated given $x_{1:t}$. Many works have explored the usage of optical flow as the auxiliary information to increase prediction accuracy [25, 36], but to our best knowledge there is no work that directly models $p(x_{t+1}|x_{1:t}, y_{1:t})$, where $y_{1:t}$ represents previous t flows.

Note that x_t can be warped to x_{t+1} given optical flow y_t, thus a vanilla network that directly maps $x_{1:t}$ and $y_{1:t}$ to x_{t+1} may learn a trial mapping. We observe that $x_{1:t}$ and x_{t+1} are from a very short duration in a video, and they are very similar to each other in content. It is reasonable to assume that $x_{1:t}$ and x_{t+1} are determined by the same hidden variables z that control the content information. We thus resort to Conditional Variational Autoencoder (CVAE) as the generative model for modeling $p(x_{t+1}|x_{1:t}, y_{1:t})$, in which we compute z from $x_{1:t}$, and then generate x_{t+1} from z, with $y_{1:t}$ as the conditions.

Formally, we have the following ELBO :

$$\log p(x_{t+1}|y_{1:t}) \geq \mathbb{E}_q \log \frac{p(x_{t+1}|z, y_{1:t})p(z|y_{1:t})}{q(z|x_{t+1}, y_{1:t})}. \tag{4}$$

Replacing $q(z|x_{t+1}, y_{1:t})$ in Eq. 4 by $q(z|x_{1:t}, y_{1:t})$ yields:

$$\begin{aligned} \log p(x_{t+1}|y_{1:t}) &\geq \mathbb{E}_q \log \frac{p(x_{t+1}|z, y_{1:t})p(z|y_{1:t})}{q(z|x_{t+1}, y_{1:t})} \\ &\approx \mathbb{E}_q \log \frac{p(x_{t+1}|z, y_{1:t})p(z|y_{1:t})}{q(z|x_{1:t}, y_{1:t})} \\ &= -KL[q(z|x_{1:t}, y_{1:t}) \,||\, p(z|y_{1:t})] \\ &\quad + \mathbb{E}_q[\log p(x_{t+1}|z, y_{1:t})], \tag{5} \end{aligned}$$

where KL denotes Kullback-Leibler divergence.

Guided by Eq. 5, we design our future prediction model as shown in Figure 2. We have two encoders E_θ and F_ϕ, and one decoder D_ψ. E_θ encodes optical flows $y_{1:t}$ to obtain $E_\theta(y_{1:t})$ from which the prior distribution $p(z|y_{1:t})$ can be obtained. F_ϕ admits the concatenation of $x_{1:t}$ and $y_{1:t}$ and outputs features $F_\phi(x_{1:t}, y_{1:t})$ from which the posterior distribution $q(z|x_{1:t}, y_{1:t})$ can be obtained. During training, we sample z from the posterior distribution, and concatenate z with the conditions $E_\theta(y_{1:t})$, which are finally sent to the decoder D_ψ to generate the future frame $\hat{x}_{t+1}$. Inspired by the Variational UNet proposed in [6], we add skip connections between F_ϕ and D_ψ to help generating x_{t+1}.

We assume $p(x_{t+1}|z, y_{1:t})$, $p(z|y_{1:t})$ and $q(z|x_{1:t}, y_{1:t})$ in Eq. 5 are all parametric Gaussian distributions. Hence, as the common practice in VAE [22], we arrive at the following loss function containing two parts:

$$\mathcal{L}_{CVAE} = KL[q(z|x_{1:t}, y_{1:t})||p(z|y_{1:t})] \\ + ||x_{t+1} - \hat{x}_{t+1}||_2^2, \quad (6)$$

where x_{t+1} is the ground truth future frame. Following [25], we also define a gradient loss:

$$\mathcal{L}_{gd}(X, \hat{X}) = \sum_{i,j} \left| |X_{i,j} - X_{i-1,j}| - |\hat{X}_{i,j} - \hat{X}_{i-1,j}| \right| \\ \left| |X_{i,j} - X_{i,j-1}| - |\hat{X}_{i,j} - \hat{X}_{i,j-1}| \right|, \quad (7)$$

where i, j indicate spatial pixel position in an image. Combining Eq. 6 and the gradient loss between the predicted future frame and its ground truth, we train our CVAE model by the following loss function:

$$\mathcal{L} = \lambda_{CVAE} L_{CVAE} + \lambda_{gd} \mathcal{L}_{gd}(\hat{x}_{t+1}, x_{t+1}), \quad (8)$$

where λ_{CVAE} and λ_{gd} are the balancing hyper-parameters.

Deterministic *vs*. stochastic future prediction during testing. At test time, we can sample z stochastically in order to generate the future frame. But this would synthesize slightly different future frames at different time. In order to predict the future frame deterministically, we use the mean of the posterior distribution $q(z|x_{1:t}, y_{1:t})$ as the sampled z at test time, and all the experimental results of our method in this paper are obtained under this sampling strategy. But note that this two kinds of sampling strategies have similar anomaly detection performance. Please see the supplemental material for more details.

3.3. Anomaly detection

At test time, our anomaly score is composed of two parts: (1) the flow reconstruction error as $S_r = ||\hat{y}_{1:t} - y_{1:t}||_2^2$ and (2) the future frame prediction error as $S_p = ||\hat{x}_{t+1} - x_{t+1}||_2^2$. We obtain the anomaly score by fusing the two
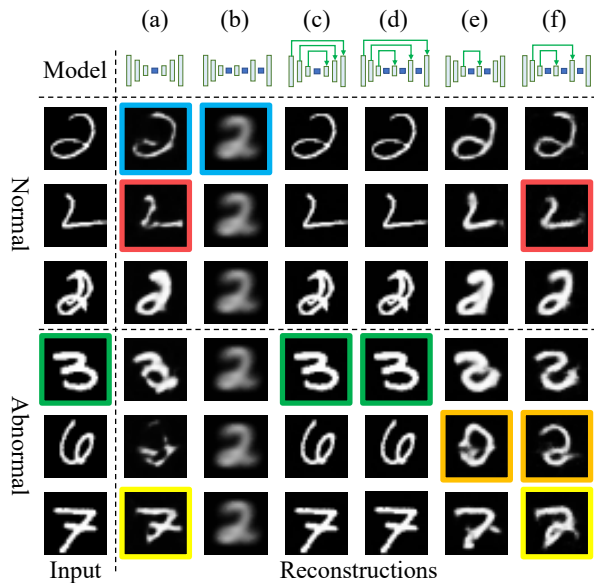

Figure 4. Visualization examples of memory-augmented autoencoder variants for reconstruction task on MNIST [23] dataset.

errors using a weighted sum strategy as:

$$S = w_r \cdot \frac{S_r - \mu_r}{\sigma_r} + w_p \cdot \frac{S_p - \mu_p}{\sigma_p}, \quad (9)$$

where $\mu_r, \sigma_r, \mu_p, \sigma_p$ are means and standard deviations of reconstruction errors and prediction errors of all the training samples, w_r and w_p are the weights of the two scores.

4. Experiments

4.1. Toy experiments on MNIST

We conduct toy anomaly detection experiments on the MNIST [23] dataset to explore the variations of memory-augmented autoencoders. We use "2" of the training set of MNIST as the normal data to train our models, and use the testing set of MNIST as our test data in which digits other than "2" are abnormal. We compare among 6 variants, the architectures of which are shown in the top row of Figure 4 from (a) to (f). (a) is the original MemAE. (b) is a variant of (a) by adding memory modules at all other levels of the decoder. (c) is another variant of (a) by adding skip connections between encoder and decoder at all levels. (d) is the combination of (b) and (c). In (e), we add one more memory to MemAE in the next level after bottleneck and also a skip connection for that level. In (f), we extend (e) by adding another pair of memory and skip connection in the next level further, which is the proposed ML-MemAE-SC.

As shown in Figure 4, MemAE reconstructs normal "2" very well, but it also reconstructs "3" and "7" successfully which however are anomalies. By simply adding more memory modules in (b), all tested inputs are reconstructed to a blurry "2", demonstrating that cascaded memories are

too aggressive to filter out all useful information. The models in (c) and (d) reconstruct both normal and abnormal inputs very clearly, due to the reason that the function of the memories is blocked by the outer skip connections. In (e) and (f), skip connection and memory are added in pairs, with skip connection providing additional features for the corresponding memory to learn more normal patterns. As can be seen, the reconstructed outputs of normal data in (e) and (f) are more clear than those in (a), while for the abnormal inputs other than "2", the models in (e) and (f) try to reconstruct them to "2" which cannot be accomplished by MemAE, indicating normality is better learned in (e) and (f). Since (f) has one more pair of skip connection and memory, the results of (f) are even better than those of (e).

4.2. Video Anomaly Detection Experiments

Datasets. To evaluate both qualitative and quantitative results of the proposed HF^2-VAD and compare it with state-of-the-art algorithms, we conduct experiments on three public video anomaly detection datasets, *i.e.*, UCSD Ped2 [34], CUHK Avenue [29] and ShanghaiTech [32]. (a) UCSD Ped2 consists of 16 training videos and 12 testing videos, acquired with a stationary camera. The training normal data contains only pedestrians walking, while abnormal events are due to either the circulation of non-pedestrian entities (*e.g.* car) or anomalous pedestrian motion patterns (*e.g.* skateboarding). (b) CUHK Avenue includes 16 training and 21 testing videos, collected from a fixed scene. There are 47 abnormal events in total such as running, throwing bag, etc. (c) ShanghaiTech is very challenging that contains videos from 13 scenes with complex light conditions and camera angles. The overall number of frames for training and testing reach 274K and 42K, respectively. There exist 130 abnormal events in the test set, scattering in 17K frames.

Evaluation criterion. Following the popular evaluation metric in VAD literature [34, 29, 25, 8, 47], we measure the Area Under the Receiver Operation Characteristic (AUROC) by varying the threshold over anomaly score. Higher AUROC indicates better VAD accuracy.

Implementation details. We do not train our model on the whole video frames but the foreground objects. Following [47], we extract all foreground objects for the training and testing videos. Each object is identified by an RoI bounding box. For each RoI, we build a spatial-temporal cube (STC) that contains not only the object in the current frame but also the contents in the same bounding box of previous t frames, where $t = 4$. The width and height of STCs are both resized to 32. Similarly, we extract the corresponding STCs for optical flows, which are estimated by FlowNet2.0 [16]. The extracted STCs for objects and optical flows are the buildingblocks used to train our model. During testing, the anomaly score of a frame is the maximum score of all objects in it. Considering the continuity

Table 1. Comparison of frame-level anomaly detection performance with state-of-the-art methods. We calculate AUROC ↑ (%) on UCSD Ped2 [34], CUHK Avenue [29] and ShanghaiTech [32]. Numbers in bold indicate the best results.

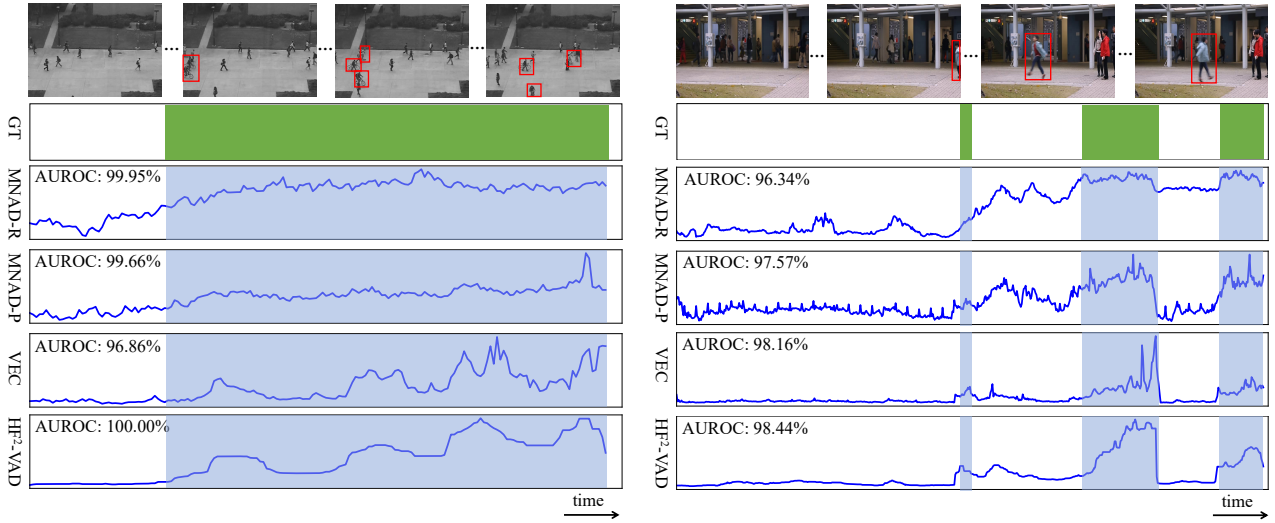| Method	UCSD Ped2	CUHK Avenue	SHTech
Conv-AE [11]	90.0	70.2	-
ConvLSTM-AE [31]	88.1	77.0	-
GMFC-VAE [7]	92.2	83.4	-
MemAE [8]	94.1	83.3	71.2
MNAD-R [38]	90.2	82.8	69.8
Frame-Pred. [25]	95.4	85.1	72.8
Conv-VRNN [30]	96.1	85.8	-
MNAD-P [38]	97.0	88.5	70.5
VEC [47]	97.3	90.2	74.8
ST-AE [50]	91.2	80.9	-
AMC [36]	96.2	86.9	-
AnoPCN [46]	96.8	86.2	73.6
HF^2-VAD w/o FP	98.8	86.8	73.1
HF^2-VAD w/o FR	94.5	90.2	76.0
HF^2-VAD	**99.3**	**91.1**	**76.2**

of activity, the anomaly scores of a video are smoothed by a median filter whose window size is 17.

We use PyTorch [39] to implement HF^2-VAD and adopt Adam optimizer [21] with $\beta_1 = 0.9, \beta_2 = 0.999$ to optimize it. The learning rate is initialized to $1e^{-4}$, decayed by 0.8 after every 50 epochs. The slot number N is set to 2K for all three datasets. The memory module numbers for Ped2, Avenue and ShanghaiTech are 3, 3 and 2, respectively. We train ML-MemAE-SC at first, then train CVAE model with the reconstructed flows, and finally fine-tune the whole framework. $\lambda_{recon}, \lambda_{ent}, \lambda_{CVAE}$ and λ_{gd} are $1.0, 2e^{-4}, 1.0$ and 1.0, respectively. The batch size and epoch number of Ped2, Avenue, and ShanghaiTech are set to (128, 80), (128, 80), and (256, 50), respectively. The error fusing weights (w_r, w_p) for Ped2, Avenue and ShanghaiTech are set to (1.0, 0.1), (0.05, 1.0) and (0.02, 1.0), respectively.

4.3. Results

Quantitative results. We compare our proposed HF^2-VAD with state-of-the-art methods, including (1) reconstruction-based methods: Conv-AE [11], ConvLSTM-AE [31], GMFC-VAE [7], MemAE [8] and MNAD-R [38]; (2) prediction-based methods: Frame-Pred. [25], MNAD-P [38], VEC [47] and Conv-VRNN [30]; and (3) hybrid methods including ST-AE [50], AMC [36] and AnoPCN [46]. In addition, we compare with two variants of our method, "HF^2-VAD w/o FP" and "HF^2-VAD w/o FR" which are formulated by removing frame prediction and flow reconstruction from HF^2-VAD, respectively. The results are summarized in Table 1, and the performances of other methods are obtained from their original papers.

(a) A Ped2 video with abnormal events: skateboarding and riding bicycle.

(b) An Avenue video with the abnormal event: kid running.

Figure 5. Two examples of anomaly detection comparison on USCD Ped2 [34] and CUHK Avenue [29]. From top to bottom, we show the sampled video frames, ground-truth abnormal sections (green regions are abnormal), result of MNAD-R [38], result of MNAD-P [38], result of VEC-VAD [47] and result of HF^2-VAD. Larger value in curves indicates more possible to be anomaly. Best viewed in color.

As can be observed, the HF^2-VAD model achieves better results than state-of-the-art methods on all these three benchmarks, which demonstrates the effectiveness of our method. Especially, HF^2-VAD outperforms other hybrid methods by a large margin. For example, HF^2-VAD achieves the accuracy of 91.1% on the CUHK Avenue dataset, but the best accuracy of previous hybrid methods is 86.9% by AMC [36]. Very interestingly, "HF^2-VAD w/o FP" performs better on Ped2 than on CUHK-Avenue and ShanghaiTech. This is because optical flow provides more discriminative clues for the Ped2 dataset, which has also been observed by Yu *et al.* [47]. Oppositely, "HF^2-VAD w/o FR" performs better on CUHK-Avenue and ShanghaiTech than on Ped2. Our full model combines the advantages of both, achieving the best results.

Qualitative results. Examples in Figure 5 show anomaly curves of two testing videos compared among MNAD-R [38], MNAD-P [38], VEC [47] and HF^2-VAD. An anomaly curve shows the anomaly scores of all frames of a video sequentially, by which we can more intuitively compare the performance of different methods. As can be seen, VEC and HF^2-VAD perform much better than MNAD-R and MNAD-P in normal sections, producing lower and more stable anomaly scores. HF^2-VAD is even better than VEC as it can better recognize abnormal events as shown in the abnormal durations where the anomaly scores computed by HF^2-VAD are higher than those by VEC. The AUROC values in these figures coincide with these intuitions.

In Figure 6, we demonstrate several normal and abnormal images, showing the ground-truth in the first column and the predicted frames by HF^2-VAD in the second column. For saving space, we do not show predicted frames by VEC and MNAD-P, but instead show the difference maps
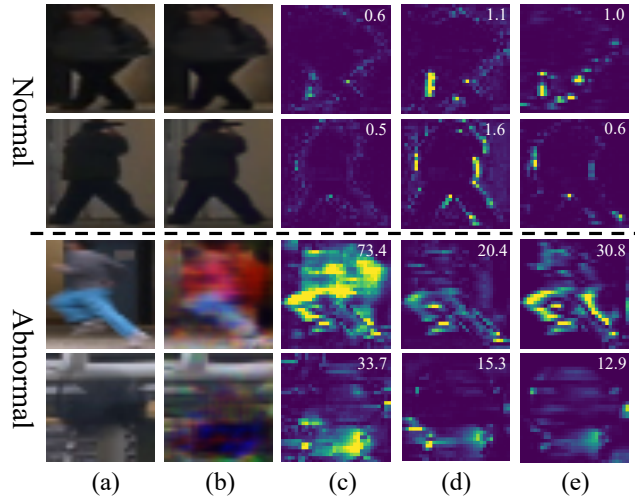

Figure 6. Visualization examples of frame prediction comparison with different methods. From top to bottom, we show normal and abnormal data. From left to right, we show (a) the ground-truth, (b) prediction results of HF^2-VAD, (c) difference maps of HF^2-VAD, (d) difference maps of VEC [47], and (e) difference maps of MNAD-P [38]. The numbers in each error map denote the corresponding sum-square-error between ground truth and the predicted frame. The lighter colors in error map denote larger prediction error. Best viewed in color.

between the ground-truth and the predicted frames by HF^2-VAD, VEC and MNAD-P in the last three columns respectively. The sum square errors of these difference maps are also shown. As we can see, HF^2-VAD produces less difference for normal images, while results in much larger differences for abnormal images. Taken the running person in the third row as an example, after the original abnormal optical flow being processed by ML-MemAE-SC, the flow fed into the CVAE is not consistent with the input images,

Table 2. Ablation study results on UCSD Ped2 [34] dataset. The anomaly detection performance is reported in terms of AUROC ↑ (%). Number in bold indicates the best result.

	Memory-augmented Reconstruction Models			Prediction Models		AUROC
				VAE	CVAE	
Flow	✓					96.27
		✓				97.75
			✓			98.81
Frame				✓		89.96
					✓	94.48
Hybrid	✓				✓	96.91
		✓			✓	98.28
			✓		✓	**99.31**

Figure 7. A failure case of the proposed HF^2-VAD. The red dashed circle denotes the video section where the VAD performance is not good. Best viewed in color.

yielding pixel shifting and color confusion as appeared in the predicted future frame.

4.4. Discussion

Ablation investigation. To analyze the role of different components of HF^2-VAD, we conduct ablation study on UCSD Ped2 dataset [34] and report the anomaly detection performance in terms of AUROC in Table 2. For the flow reconstruction part, we consider MemAE, ML-MemAE-SC with 2 memories and ML-MemAE-SC with 3 memories. For the prediction part, besides CVAE, we also investigate VAE for future frame prediction just based on previous frames. We then stitch the three reconstruction models with the CVAE to obtain three hybrid variants of our proposed method.

As can be seen from the table, just by reconstruction, the anomaly detection accuracy of ML-MemAE-SC with 3 memories is the highest, followed by ML-MemAE-SC with 2 memories, and finally MemAE, which are 98.81%, 97.75%, and 96.27%, respectively. Just by prediction, CVAE outperforms VAE by a large margin from 89.96% to 94.48%, which shows that optical flow, as additional information, is critically necessary to increase the prediction accuracy. By integrating reconstruction and prediction, all the hybrid models are better than their corresponding reconstruction-only or prediction-only models, in which the hybrid method comprising ML-MemAE-SC with 3 memories and CVAE achieves the highest score among all the compared variants, which is 99.31%.

Computation time. The experiments are performed on an NVIDIA RTX 3090 GPU and an Intel Core(TM) i9-7920X CPU @ 2.90GHz. As [47], we need to preprocess the input video to extract all foreground objects and construct STCs for objects on the frame and optical flow. For our current implementation, the preprocessing phase averagely takes about 0.092s per frame. As for the anomaly detection phase, the model inference and anomaly score cal-
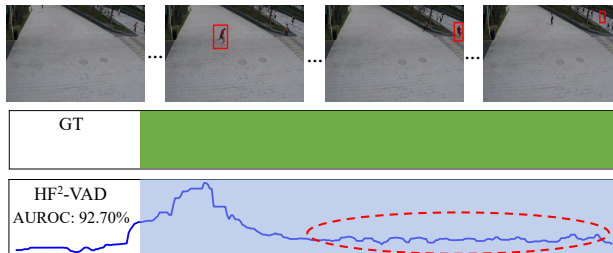
culation together costs about 0.015s per frame. Overall, the running speed of HF^2-VAD is about 10fps.

Failure case. Figure 7 shows a failure case of our method on a ShanghaiTech testing video in which a person is running. HF^2-VAD can easily detect it after the person comes into the view of the camera. But as the object runs farther away from the camera, the anomaly score becomes smaller and smaller. Our method fails on anomaly object that is far away, because a very far object, even though it is running, the absolute optical flow values are similar or even smaller than those of the normal objects near the camera. We conjecture that the scene depth is a very important variable that should be considered to further improve the VAD accuracy for the ShanghaiTech dataset, which would be investigated in future work.

5. Conclusion

In this paper, we have explored the possibility of combining reconstruction and prediction, which are the two most prevalent VAD paradigms nowadays, to obtain a hybrid VAD method for high-accuracy video anomaly detection. Experiments show that the proposed method outperforms previous reconstruction-only or prediction-only methods, and also is the most superior in the hybrid approach camp. Our integration strategy is novel, by utilizing a CVAE to predict future frame taking both previous video frames and optical flows as input, and proposing an effective reconstruction method, *i.e.* the ML-MemAE-SC, to preprocess the flows in advance. That means our hybrid method is not a simple combination of reconstruction and prediction, but the reconstruction can effectively influence the prediction quality. It is the strong and inherent entanglement between the reconstruction and prediction that makes our method performs better than state-of-the-art approaches.

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-occurences. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2458–2465. IEEE, 2009. 1

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys*, 41(3):1–58, 2009. 1

[4] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[5] Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. Dual graph convolutional networks with transformer and curriculum learning for image captioning. In *Proceedings of the ACM International Conference on Multimedia*, 2021. 1

[6] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 3, 5

[7] Yaxiang Fan, Gongjian Wen, Deren Li, Shaohua Qiu, Martin D Levine, and Fei Xiao. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. *Computer Vision and Image Understanding*, page 102920, 2020. 1, 2, 6

[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, 3, 4, 6

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[10] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016. 2

[11] 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

[12] Tao Hu, Chengjiang Long, and Chunxia Xiao. Crd-cgan: Category-consistent and relativistic constraints for diverse text-to-image generation. *arXiv preprint arXiv:2107.13516*, 2021. 1

[13] Tao Hu, Chengjiang Long, and Chunxia Xiao. A novel visual representation on text using diverse conditional gan for visual recognition. *IEEE Transactions on Image Processing*, 30:3499–3512, 2021. 1

[14] Gang Hua, Chengjiang Long, Ming Yang, and Yan Gao. Collaborative active learning of a kernel machine ensemble for recognition. In *IEEE International Conference on Computer Vision*. IEEE, 2013. 1

[15] Gang Hua, Chengjiang Long, Ming Yang, and Yan Gao. Collaborative active visual recognition from crowds: A distributed ensemble approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):582–594, 2018. 1

[16] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 6

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[18] Ashraful Islam, Chengjiang Long, Arslan Basharat, and Anthony Hoogs. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[19] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *AAAI Conference on Artificial Intelligence*, 2021. 1

[20] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928. IEEE, 2009. 1, 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3, 5

[23] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 5

[24] Bo Li, Sam Leroux, and Pieter Simoens. Decoupled appearance and motion learning for efficient anomaly detection in surveillance video. *arXiv preprint arXiv:2011.05054*, 2020. 1

[25] 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, 4, 5, 6

[26] Chengjiang Long and Gang Hua. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 1

[27] Chengjiang Long and Gang Hua. Correlational gaussian processes for cross-domain visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[28] Chengjiang Long, Gang Hua, and Ashish Kapoor. A joint gaussian process model for active visual recognition with expertise estimation in crowdsourcing. *International Journal of Computer Vision*, 116(2):136–160, 2016. 1

[29] 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. 1, 6, 7

[30] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–8. IEEE, 2019. 1, 2, 6

[31] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo*, pages 439–444. IEEE, 2017. 1, 2, 6

[32] 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. 6

[33] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *In Proceedings of the International Conference on Machine Learning*, volume 30, page 3, 2013. 4

[34] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010. 1, 6, 7, 8

[35] 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. 2

[36] 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, 4, 6, 7

[37] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 4

[38] Hyunjong Park, Jongyoun Noh, and Bumsub 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, 3, 4, 6, 7

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[40] Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for video Technology*, 18(11):1544–1554, 2008. 2

[41] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn: Sparse graph convolution for pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[42] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 3483–3491. Curran Associates, Inc., 2015. 2, 3

[43] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with contrastive learning of long and short-range temporal features. *arXiv preprint arXiv:2101.10030*, 2021. 1

[44] Jinjiang Wei, Chengjiang Long, Hua Zou, and Chunxia Xiao. Shadow inpainting and removal using generative adversarial networks with slice convolutions. *Computer Graphics Forum*, 38(7):381–392, 2019. 1

[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. 1

[46] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1805–1813, 2019. 2, 6

[47] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020. 1, 2, 4, 6, 7, 8

[48] Ling Zhang, Chengjiang Long, Qingan Yan, Xiaolong Zhang, and Chunxia Xiao. Cla-gan: A context and lightness aware generative adversarial network for shadow removal. *Computer Graphics Forum*, 39(7), 2020. 1

[49] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In *AAAI Conference on Artificial Intelligence*, 2020. 1

[50] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017. 1, 2, 6