IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 2, 2022, accepted March 29, 2022, date of publication April 15, 2022, date of current version April 29, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3165977

Unsupervised Anomaly Video Detection via a Double-Flow ConvLSTM Variational Autoencoder

LIN WANG[1], **(Member, IEEE), HAISHU TAN**[2,3], **(Member, IEEE), FUQIANG ZHOU**[1], **WANGXIA ZUO**[4], **AND PENGFEI SUN**[1]

[1]School of Instrumentation Science and Opto-Electronics Engineering, Beihang University, Beijing 100191, China
[2]School of Physics and Optoelectronic Engineering, Foshan University, Foshan 528000, China
[3]Ji Hua Laboratory, Foshan 528000, China
[4]School of Electrical Engineering, University of South China, Hengyang 421001, China

Corresponding authors: Haishu Tan (tanhaishu@fosu.edu.cn) and Fuqiang Zhou (zfq@buaa.edu.cn)

This work was supported by the Ji Hua Laboratory of Guangdong Province, China, under Grant X200051UZ200.

ABSTRACT With the rapid increase of video surveillance points in the market in recent years, video anomaly detection has gained extensive attention in the security field. At present, the distribution of normal and anomalous data is unbalanced in unlabeled video data. Variational autoencoder (VAE), as one of the typical deep generative models, gets increasingly popular in unsupervised anomaly detection. However, this model is not good at processing time-series data, especially video data. In addition, the strong generalization ability which is over-reconstructing anomaly behavior of many autoencoder-based works leads to the missed anomaly detection. To solve these problems, in this paper, we present a double-flow convolutional long short-term memory variational autoencoder (DF-ConvLSTM-VAE) to model the probabilistic distribution of the normal video in an unsupervised learning scheme, and to reconstruct videos without anomaly objects for anomaly video detection. Experiments verify the effectiveness and competitiveness of our DF-ConvLSTM-VAE on multiple public benchmark datasets. In particular, our model achieves the state-of-the-art performance on anomalous event count.

INDEX TERMS Autoencoder, variational autoencoder, LSTM, ConvLSTM, anomaly detection.

I. INTRODUCTION

Anomaly detection has a wide range of practical applications in campus monitoring, intelligent transportation, banking transactions. Nowadays, in an era of data explosion, unlabeled data, especially unlabeled surveillance video data pervades every aspect of life. Compared to other algorithms [1], [2], unsupervised learning algorithms are becoming the future trend and are of great interest to scientists [3]–[6]. As an essential area of anomaly detection, anomaly video detection provides us with various pattern classification of normal and anomalous behaviors in respective domains [7]–[9]. In fact, anomaly video detection task suffers from several challenges. For the existing large amounts of video data, there is bound to be a large number of normal videos without event occurrence. Finding out the time period of major event occurrence is of great significance for storage and review of videos. Therefore, it is of great research value and practical significance to detect anomalous videos using unsupervised learning methods. In many cases, whether real-life events are normal or anomalous depends on their surrounding circumstances. For example, a person running in a sports field is perfectly normal, but in a court of law, it is clearly abnormal. Another example is the presence of a speeding truck on a campus sidewalk, which is clearly unusual and potentially dangerous. These cases show that identifying whether an event is an anomalous event is difficult. In addition, it is well known that video presentation learning is the most basic problem in video processing technology. Compared with the static images, video involves richer dynamic information about events. In addition, due to the diversity and variability of video, it becomes an urgent problem to study the algorithm which can find the internal spatio-temporal correlation and discriminating features of video.

Researchers usually extract handcrafted video features to detect anomalies over the past few years. Traditional methods are based on low-level features, such as histograms of optical flow(HOF) [10], spatio-temporal gradient [11], and mixture of dynamic textures(MDTs) [12], to complete anomaly classification tasks. These models based on manual feature

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

classification are inefficient, and their accuracy cannot meet actual requirements. With the development of the deep learning, many neural networks are proposed by researchers and used for detecting anomalies. More discriminating features of videos are learned by these networks through unsupervised learning. Hasan *et al.* [13] employ the Convolutional Autoencoder (ConvAE) to construct an anomalous event detection model. Although the network input is continuous multiple frames, 2D convolution is adopted, failing to fully utilize the temporal information between video frames. Considering the motion characteristics, Yiru *et al.* [14] use an autoencoder with 3D convolution for anomaly video detection.Although the model has the ability of reconstruction and prediction, it is not good at modeling long video. Xu *et al.* [15] leverage a stacked denoising autoencoders to learn both appearance and motion features, and based on the learned features, multiple one-class SVM models are used to predict the anomaly scores of each input. However, this method is time-consuming by dividing spatio-temporal features of video into optical flow and appearance features. Long Short-Term Memory(LSTM) network, a typical type of recurrent neural networks(RNN) architecture, is proposed by Hochreiter *et al.* [16] and widely used in many research tasks. Take video for example, this network has been applied in action recognition [17]–[19], video retrieval [20], [21], video segmentation [22], [23] and Video Captioning [24], [25], etc. LSTM-Autoencoder, a typical sequence-to-sequence [26] framework, is proposed by Srivastava *et al.* [17] and applied for learning video action recognition. LSTM also performs well in video anomaly detection task [27]. Medel and Savakis *et al.* [28] complete video anomaly detection by combining LSTM network and ConvLSTM unit. Base on ConvAE structure, Yong and Yong [29] add three ConvLSTM layers to learn the spatio-temporal information of video event and detect video anomalies. Lin *et al.* [30] explore a hybrid autoencoder architecture, composed of ConvAE and LSTM-Autoencoder with ConvLSTM unit, to improve the extrapolate capability of the corresponding decoder through the shortcut connection. The prediction branch of the hybrid autoencoder is used for anomalies detection. These models are designed by autoencoder structure and use the reconstruction error to detect anomalies. This reconstruction-based algorithm, as one of the common techniques for anomaly detection, calculates the maximum reconstruction error of test samples to determine whether it is anomalous or not. In fact, the anomalous object is generated in the reconstructed image by these methods and it is relatively fuzzy or low in pixels compared with the original image. It would be a better choice for detecting anomalous data if the reconstructed image contains only normal instead of any abnormal objects during the test phase.

In recent years, variational autoencoder (VAE) [31] has become increasingly popular. In particular, VAE cannot only generate the characteristic output close to the original input and reflect the similar information of similar data, but also learn the potential characteristic vector. An and Cho [32] propose an anomaly detection method using the reconstruction probability from the variational autoencoder. The reconstruction probability is a probabilistic measure that takes into account the variability of the distribution of variables. Experimental results of this paper show that this method outperforms autoencoder-based methods on MNIST dataset [33]. Compared with the reconstruction error used by the autoencoder and the principal component-based anomaly detection method, the reconstruction probability with a theoretical background is more principled and objective. However, VAE limits its applicability to time series, especially to video, for it does not take the temporal characters of video into account. For processing the time-series data, Sölc *et al.* [34] utilize RNNs and the variational inference to learn time-series data for anomaly detection. Park *et al.* [35] use a long short-term memory-based variational autoencoder(LSTM-VAE) for multimodal anomaly detection. These two papers demonstrate that the VAE-based models are better than the other approaches, and inspire us to apply a recurrent VAE for anomaly detection in video.

In this work, in order to solve above problems, our two models choose ConvLSTM units instead of LSTM units to learn the internal spatio-temporal relations of video. These two asymmetric models blend ConvLSTM with VAE architecture to reconstruct videos without anomaly objects for anomaly detection (see Figures 10 – 12). One is called ConvLSTM-VAE(Asymmetric); The other is named DF-ConvLSTM-VAE. More information about the structures of these two models is described in Section III. We use reconstruction error probability which is different from reconstruction probability to detect anomalies. Experiments verify the effectiveness and competitiveness of our DF-ConvLSTM-VAE on multiple public benchmark datasets. In particular, our model achieves the state-of-the-art performance on anomalous event count. The key contributions of our work can be summarized as follows:

- For the disadvantage of strong generalization ability of many autoencoder-based models, and the VAE does not take the temporal dependence in data into account, which limits its applicability to time series, especially video sequence. We present two models-ConvLSTM-VAE(Asymmetric) and DF-ConvLSTM-VAE to solve this disadvantage. These two models are consisting of ConvLSTM and VAE, to model the probability distribution of video sequence by capturing the crowd spatial-temporal features. The experimental results verify the validity of these two asymmetric models.
- Based on the analysis and verification of the ConvLSTM-VAE(Asymmetric) model, we propose an improved network, namely DF-ConvLSTM-VAE to detect anomalies. The DF-ConvLSTM-VAE model adopts the idea of asymmetric structure and increase the width of network structure to achieve high training efficiency and short test time.
- The DF-ConvLSTM-VAE model is successfully utilized for anomaly detection in videos. The experimental results demonstrate that the DF-ConvLSTM-VAE model

has a certain competitiveness compared with current leading methods on benchmark datasets.

The remainder of this paper is organized as follows. In Section II, we briefly review many related works. Section III, describes the proposed approach. Experiments are conducted for analysis in Section IV. We discuss the limitation of our work in Section V. Finally, we draw conclusions and present future research directions in Section VI.

II. RELATED WORKS

A. CONVOLUTIONAL LSTM UNIT

Convolutional Long Short-term Memory (Conv-LSTM) unit, as a variant of the LSTM unit, is firstly proposed by Shi *et al.* [36]. Compared to the usual fully connected LSTM (FC-LSTM) [17], spatial information is encoded by ConvLSTM when dealing with spatio-temporal data in input-to-state and state-to-state transition. With respect to predicting future video sequences for a synthetic Moving-MNIST Dataset [37], ConvLSTM exhibits superior performance than FC-LSTM.

The formulation of the ConvLSTM unit can be summarized with Equation (1), where the symbol '$*$'denotes a convolution operation, and '$\circ$'denotes the Hadamard product. The input, forget, cell, output and hidden state of each timestep are denoted by i, f, C, o and H respectively, the activation is denoted by σ, and the weighted connection between states by a set of weights, W. The input is fed in as images, while the set of weights for every connection is replaced by convolutional filters.

$$\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i)\\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f)\\
C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c)\\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_{t-1} + b_o)\\
H_t &= o_t \circ \tanh(C_t) \qquad (1)
\end{aligned}$$

This operation prompts ConvLSTM to work better with images than the FC-LSTM, for the model has the ability to propagate spatial characteristics temporally through each ConvLSTM state. Inspired by this, our two models apply the Conv-LSTM as a basic block for recurrent connections inside the VAE model.

B. AUTOENCODER

An autoencoder (AE), composed of an encoder and a decoder, aims to reconstruct input data x from a learned hidden representation z. The objective function of an AE is represented in Equation (2) below, where ϕ and θ denote the hidden parameters of the encoder E and the decoder G, and $\mathcal{L}_{AE}$ denotes the loss of AE. We use the reconstruction error of each test data to calculate the anomaly score, and we consider that the data with high anomaly score is anomalies. The AE can behave well in reconstructing normal data, while failing to do so with anomaly data that the autoencoder has not encountered.

$$\begin{aligned}
z &= E_\phi(x)\\
\mathcal{L}_{AE}(x, \phi, \theta) &= \|x - G_\theta(z)\|^2 \qquad (2)
\end{aligned}$$

C. VARIATIONAL AUTOENCODER

The Variational Autoencoder (VAE) is proposed by [31]. The structure of VAE is similar to that of AE. But essentially, a difference between them is that the encoder of VAE forces the representation z to obey some kind of prior probability distribution $p(z)$ (e.g. $\mathcal{N}(0, I)$). Then the decoder generates new realistic data with code z sampled from $p(z)$. $p_\theta(z)$ is the prior distribution of the latent variable z. By inheriting the architecture of an AE, a VAE consists of the following three parts.

(1) Recognition network (encoder network): a probabilistic encoder E_ϕ, which map input x to the latent representation z to approximate the true posterior distribution $p(z|x)$. This recognition network can be represented as the approximate posterior $q_\phi(z|x)$.

$$\mu, \log(\sigma) = E_\phi(x) \qquad (3)$$

(2) Sampling process: $\epsilon \sim \mathcal{N}(\mu, \sigma)$

$$z = \mu + \sigma \odot \epsilon \qquad (4)$$

(3) Generative network (decoder network): a generative decoder G_θ, which reconstructs the latent representation z to the input value $\widetilde{x}$, does not rely on any particular input x. This generative network can be represented as $p_\theta(x|z)$.

$$\widetilde{x} = G_\theta(z) \qquad (5)$$

where ϕ, θ denote the parameters of recognition and generative network, respectively.

The data distribution $p_\theta(x)$ is intractable by analytic methods, so variational inference methods are introduced to solve the maximum likelihood $\log p_\theta(x)$. The loss of the VAE is represented as Equation (6).

$$\begin{aligned}
&\mathcal{L}_{VAE}(x, \phi, \theta)\\
&\quad = \log p_\theta(x) - KL[q_\phi(z|x) \parallel p_\theta(z|x)]\\
&\quad = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x) + \log p_\theta(z|x) - \log q_\phi(z|x)]\\
&\quad = \mathbb{E}_{z \sim q_\phi(z|x)} - D_{KL}(q_\phi(z|x) \parallel p_\theta(z)) \qquad (6)
\end{aligned}$$

In order to estimate this maximum likelihood, a VAE needs to maximize the evidence lower bound (ELBO) $\mathcal{L}_{VAE}$. KL is a similarity measure between two distributions. To optimize the KLD between $q_\phi(z|x)$ and $p_\theta(z)$, the encoder estimates the parameter vectors of Gaussian distribution $q_\phi(z|x)$, mean μ and standard deviation σ. There is an analytical expression for their KL divergence, because both $q_\phi(z|x)$ and $p_\theta(z)$ are Gaussian. For optimizing the second term of Equation(6), the VAE minimizes the reconstruction errors between the inputs and the outputs. The objective function of the VAE can be rewritten as:

$$\mathcal{L}_{VAE} = \mathcal{L}_{MSE}(\widetilde{x}, x) + \mathcal{L}_{KLD}(\mu, \sigma)$$

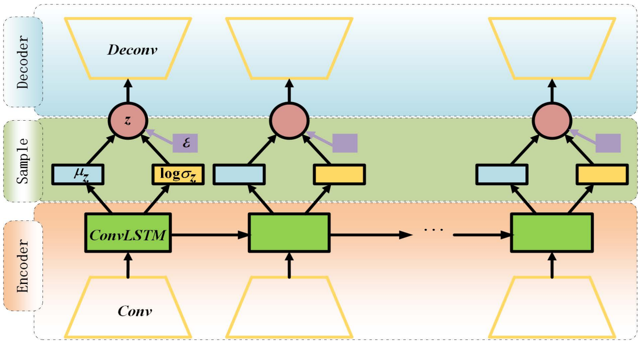


FIGURE 1. Illustration of our unrolled asymmetric convolutional LSTM-VAE model (ConvLSTM-VAE(Asymmetric)).

$$\begin{aligned}\mathcal{L}_{MSE}(\widetilde{x}, x) &= \| \widetilde{x} - x \|^2 \\ \mathcal{L}_{KLD}(\mu, \sigma) &= KL(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, I)) \\ &= \frac{1}{2}(1 + \log(\sigma^2) - \mu^2 - \sigma^2) \qquad (7)\end{aligned}$$

where the first term $\mathcal{L}_{MSE}$ is the reconstruction error (MSE, the mean squared error) between the inputs and their reconstructions. The second term $\mathcal{L}_{KLD}$ is the Kullback-Leibler divergence between the inference model $q_\phi(z|x)$ and $p_\theta(z)$. And regularize the encoder by encouraging the approximate posterior $q_\phi(z|x)$ to match the prior $p_\theta(z)$. Use the "reparameterization trick", ϕ and θ can be obtained by optimizing Equation(7) via stochastic gradient variational bases.

AE uses the reconstruction error as the anomaly score in the test phase, while VAE defines reconstruction probability for anomaly detection. To estimate the probabilistic anomaly score, a VAE samples z according to the prior $p_\theta(z)$ for L times and calculates the average reconstruction as reconstruction probability. That is why the VAE works more robustly than the traditional AE in the anomaly detection domain.

Algorithm 1 Training Algorithm for the ConvLSTM-VAE(Asymmetric) Network

```
Input: Normal training dataset $X$ for every frame $x_t, t = 1, \ldots, T$.
Output: probabilistic encoder $E_\phi$, probabilistic decoder $G_\theta$.
($E_\phi = Conv + ConvLSTM$, $G_\theta = Deconv$)
$\phi, \theta, C_0, h_0 \leftarrow$ Initialize parameters
repeat
  for $t = 1$ to $T$ do
    $F_t = Conv(x_t)$
    $\mu, \sigma, C_t, h_t = ConvLSTM(F_t, C_{t-1}, h_{t-1})$
    $z \leftarrow$ samples from $\mathcal{N}(\mu, \sigma^2)$
    $\widetilde{x}_t = Deconv(z)$
    calculate $\mathcal{L}_t = \mathcal{L}_{MSE}(\widetilde{x}_t, x_t) + \mathcal{L}_{KLD}(\mu, \sigma)$
  end for
  $\phi, \theta \leftarrow$ update parameters using gradients of $\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t$
until convergence of parameters
```

III. PROPOSED METHODS

A. THE CONVLSTM-VAE(ASYMMETRIC) MODEL

In this work, we combine ConvLSTM units with the VAE to model the video sequences for anomaly detection. Due to the traditional network based on the VAE structure, it is easy to train the VAE into the AE model in the training process. We artificially weaken the decoder from the structure, to design an asymmetric model. Figure 1 provides the structure of the ConvLSTM-VAE(Asymmetric) model which is composed of the following three parts: encoder, sample, and decoder. The encoder consists of two modules: *Conv* and *ConvLSTM*, where *Conv* represents a set of convolutional layers for extracting spatial features from each frame, and *ConvLSTM* denotes convolutional long short-term memory units for learning temporal patterns of video sequences from spatial features. In the sampling process, z is sampled from the encoder of the ConvLSTM-VAE(Asymmetric) model. The sampled data z has temporal and spatial properties. The decoder is made up of only one module: *Deconv*, which represents a set of deconvolutional layers, corresponding to the *Conv* module of encoder to generate new realistic input.

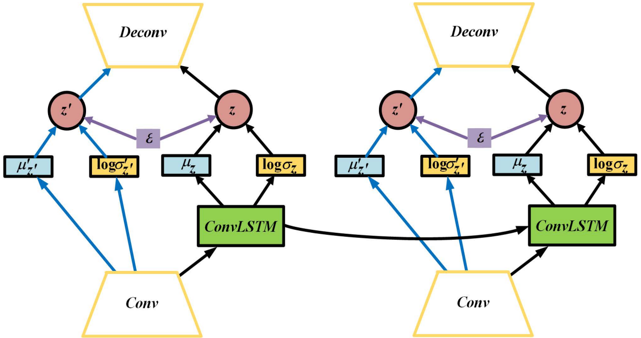


FIGURE 2. Illustration of two consecutive video blocks of our unrolled double-flow convolutional LSTM-VAE model (DF-ConvLSTM-VAE). The blue arrows and black arrows represent two different flows of the DF-ConvLSTM-VAE model.

The objective function of the ConvLSTM-VAE (Asymmetric) model can be expressed in Equation(8).

$$\mathcal{L} = \mathcal{L}_{MSE}(\widetilde{x}, x) + \mathcal{L}_{KLD}(\mu, \sigma) \qquad (8)$$

More details and configuration about our ConvLSTM-VAE(Asymmetric) model is presented in Table 1 of Section IV, and the algorithm for training the ConvLSTM-VAE(Asymmetric) is shown in algorithm 1.

B. THE DF-CONVLSTM-VAE MODEL

We believe that the decoder of the ConvLSTM-VAE (Asymmetric) model consisting only of deconvolutional layers cannot adequately decode the sampled spatio-temporal information. Meanwhile, inspired by traditional symmetric structures of many VAE-based networks, we propose an improved model, namely DF-ConvLSTM-VAE model to improve performance of networks. The DF-ConvLSTM-VAE model is a non-traditional symmetric structure variational autoencoder for processing time series data.

Figure 2 displays the structure of the DF-ConvLSTM-VAE model consisting of the following two flows: the left flow and the right flow. In Figure 2, the blue arrows represent the left flow, and the black arrows denote the right flow. Note that the structure of the right flow is the same as the ConvLSTM-VAE(Asymmetric) model. The left flow is different from the right flow. The left flow is a model which is composed of

Algorithm 2 Training Algorithm for the DF-ConvLSTM-VAE Network

```
Input: Normal training dataset $X$ for every frame $x_t$, $t = 1, \ldots, T$.
Output: probabilistic encoder$E_\phi, E'_{\phi'}$. probabilistic decoder $G_\theta, G'_{\theta'}$.
($E_\phi = Conv + ConvLSTM$, $G_\theta = Deconv$, $E'_{\phi'} = Conv$,
$G'_{\theta'} = Deconv$)
$\phi$,$\theta$, $\phi'$, $\theta'$,$C_0$,$h_0$,$C'_0$,$h'_0$ ← Initialize parameters
repeat
  for $t = 1$ to $T$ do
    $\mu', \sigma', F_t = Conv(x_t)$
    $\mu, \sigma, C_t, h_t = ConvLSTM(F_t, C_{t-1}, h_{t-1})$
    $z$← samples from $\mathcal{N}(\mu, \sigma^2)$
    $z'$ ← samples from $\mathcal{N}(\mu', \sigma'^2)$
    $\widetilde{x}_t = Deconv(z', z)$
    calculate $\mathcal{L}_t = \mathcal{L}_{MSE}(\widetilde{x}_t, x_t) + \mathcal{L}_{KLD}(\mu, \sigma) + \mathcal{L}'_{KLD}(\mu', \sigma')$
  end for
  $\phi, \theta, \phi'$,$\theta'$← update parameters using gradients of $\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t$
until convergence of parameters
```

the following three parts: encoder *Conv*, sample, and decoder *Deconv*. In particular, the left flow skips the *ConvLSTM* model directly.

Many networks often improve the network performance by increasing the depth and width of the spatial view. At the same depth of the network, we increase the network width from spatial and temporal views to improve the utilization of features, and thus improve the performance of the model. We offer a new option to learn the temporal pattern of video sequences. The DF-ConvLSTM-VAE model is composed of the following three parts: encoder, sample, and decoder. Different from the three parts of ConvLSTM-VAE(Asymmetric) model, the encoder of DF-ConvLSTM-VAE model comprises two modules: *Conv*, and *ConvLSTM* of the right flow, the sampling process consists of two sample processes: the data z of the right flow sampled from $\mathcal{N}(\mu, \sigma^2)$ and the data z' of the left flow sampled from $\mathcal{N}(\mu', \sigma'^2)$, and the decoder is a module: *Deconv*.

The objective function of DF-ConvLSTM-VAE model can be represented in Equation(9).

$$\mathcal{L} = \mathcal{L}_{MSE}(\widetilde{x}, x) + \mathcal{L}_{KLD}(\mu, \sigma) + \mathcal{L}'_{KLD}(\mu', \sigma') \qquad (9)$$

where the second term $\mathcal{L}_{KLD}$ and the third term $\mathcal{L}'_{KLD}$ represent the Kullback-Leibler divergence of the right and left flow, respectively. The algorithm for training the DF-ConvLSTM-VAE is shown in algorithm 2. More details and configurations about our DF-ConvLSTM-VAE model are provided in Table 1 and Table 2 of Section IV.

C. ANOMALY DETECTION

In this paper, we propose video anomaly detection models to calculate the anomaly score from the reconstruction error probability(REP). Given a frame x_t of the test video clip as the input, the encoder estimates the parameters of latent gaussian variables μ and σ as the output. Then the reparameterization trick is used to sample z for L times according to the latent distribution $N(\mu, \sigma^2)$, i.e. $z^{(l)} = \mu + \sigma \odot \epsilon^{(l)}$, where $\epsilon \sim N(0, I)$ and $l = 1, \ldots, L$. The generative network receives $z^{(l)}$ as input data and outputs the reconstructed frame $\widetilde{x}_t^{(l)}$. We compute the reconstruction error probability of a pixel's intensity value I at location(u, v) in frame x_t of a given video sequence by the Equation(10).

$$REP_{(u,v,t)} = \frac{1}{L}\sum_{l=1}^{L}\|\widetilde{I}^{(l)}_{(u,v,t)} - I_{(u,v,t)}\|_2 \qquad (10)$$

where $\widetilde{I}^{(l)}_{(u,v,t)}$ denotes a pixel's intensity value I at location(u, v) in reconstructed frame $\widetilde{x}_t^{(l)}$.

From each frame, we compute the REP of a frame x_t by summing up all the pixel-wise errors probabilities: $REP_{(t)} = \sum_{(u,v)} REP_{(u,v,t)}$. We compute the regularity scores $s(t)$ of a video sequence through the Equation(11):

$$s(t) = 1 - \frac{REP(t) - min_t REP(t)}{max_t REP(t)} \qquad (11)$$

In addition, in order to know the number of abnormal events in a given video, we explore local minima that are very noisy and not all meaningful in the time-series of regularity score to detect abnormal events. Distinct local minima indicate that video frames are most likely to contain anomalies. We use the Persistence1D [39] algorithm to identify meaningful local minima. In this step, if the distance of two local minima is less than 50 frames, they are identified as a part of the same abnormal event.

IV. EXPERIMENTS

A. DATASETS

To test our two methods, we conduct experiments on several challenging datasets, namely USCD Ped1 and Ped2, Avenue datasets.

1) USCD DATASET

UCSD ped dataset [12] consists of two sub-datasets,namely UCSD ped1 and UCSD ped2. In UCSD ped1 dataset, there are 34 training video clips for training and 36 video clips for testing. The resolution of each frame is 238 × 158 pixels. UCSD ped2 dataset consists of 16 training and 12 testing video clips, each with 360 × 240 resolution. Anomaly events mainly contain two categories in UCSD ped dataset, the movement of non-pedestrian entities and anomalous pedestrian motions. Anomalous events of UCSD ped dataset include bikers, skaters, carts, wheelchairs and people walking off the walkway.

2) AVENUE DATASET

There are 16 training and 21 testing video clips in AVENUE dataset [40]. The resolution of each frame is 640 × 360 pixels. Each video clip is around 2 minutes long. The training video clips contain mostly normal activities, but do include a few anomalous events. There are several typical anomalous events, including running, throwing objects and walking in the wrong direction in testing video clips. In addition, it is worth noting that the camera in this dataset has jitter problems, while the other datasets are from stationary cameras.

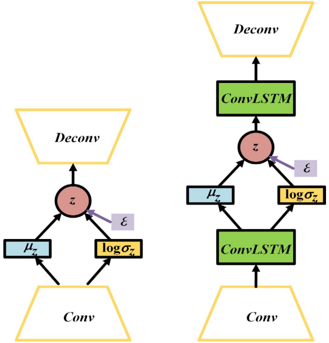


FIGURE 3. Left:Illustration of single frame structure of VAE; Right:Illustration of single frame structure of ConvLSTM-VAE(Symmetric).

B. IMPLEMENTATION DETAILS

In order to verify the performance of our asymmetric structure models, two symmetric structure models are designed for comparison. These two models are VAE and ConvLSTM-VAE(symmetric), respectively. As shown in Figure 3, the left model is VAE model consisting only of symmetric *Conv* and *Deconv* modules. The right model is ConvLSTM-VAE(Symmetric) model which symmetrically adds one ConvLSTM layer compared with VAE model. In detail, the corresponding modules parameters of the two symmetric models are the same as our model.

1) EVALUATION METRIC

In the field of video anomaly detection, two commonly used anomaly detection evaluation criteria are Equal Error Rate (EER) and Area Under Receiver Operating Characteristic Curve(AUC). These two criteria are derived from Receiver Operating Characteristic Curve(ROC), which is well suited for comparison of algorithm performance. ROC curve evaluates the detection effect of abnormal events. The ROC curve takes False positive rate(FPR) as abscissa and True positive rate(TPR) as ordinate. Here, TP(True Positive) indicates true positives, FN(False Negative) indicates false negatives, FP(False Positive) indicates False negatives, TN(True Negative) indicates true negatives. We compute FPR and TPR through the Equation(12):

$$\begin{aligned} TPR &= \frac{TP}{TP+FN} \\ FPR &= \frac{FP}{FP+TN} \end{aligned} \tag{12}$$

We select different threshold and calculate the TPR and FPR respectively to make ROC curve. EER is the point where the TPR and FPR are equal on the ROC curve, namely, the intersection of the ROC curve and the diagonal (line [0,1]-[1,0]) in the ROC space. If the EER in the ROC curve of an algorithm is smaller and the AUC is larger, it indicates that the performance of this method is better.

2) CONFIGURATIONS OF OUR MODELS

The input images are resized to 224 × 224 pixels and converted to gray-scale. The input length of two networks is ten ($T = 10$). Figure 4 gives comparison of average $\mathcal{L}_{MSE}$ of sequence of the ConvLSTM-VAE(Asymmetric) model with respect to different learning rate (Figure 4(a)), mini-batch (Figure 4(b)) and optimizer (Figure 4(c)) on USCD ped1 dataset. The three blue curves show that our asymmetric model performs best with its corresponding hyperparameters. From the Figure 4, we use an Adam optimizer with a learning rate of 10^{-4} to train our two networks from a Xavier uniform random weights initialization. Our two networks are L_2 regularized with a weight decay of 5×10^{-4}. On USCD ped1 and Avenue, the batch size is set to 4, and on USCD ped2, it is set to 8. Figure 5 shows that comparison of AUC and EER of different dimension of the learned hidden representation z on USCD ped1 dataset. As can be seen from Figure 5, the performance of the ConvLSTM-VAE(Asymmetric) model is the best when the dimension of the hidden representation z is set to 256.

Figure 6 and Table 1 provide the structure and corresponding parameters of the ConvLSTM-VAE(Asymmetric) model, respectively. The ConvLSTM-VAE(Asymmetric) model concatenates the outputs of three recurrent ConvLSTM layers and sends it to next two fully connected layers to calculate the mean and the variance (ConvLSTM4, ConvLSTM5, ConvLSTM6 → FC7, FC8).

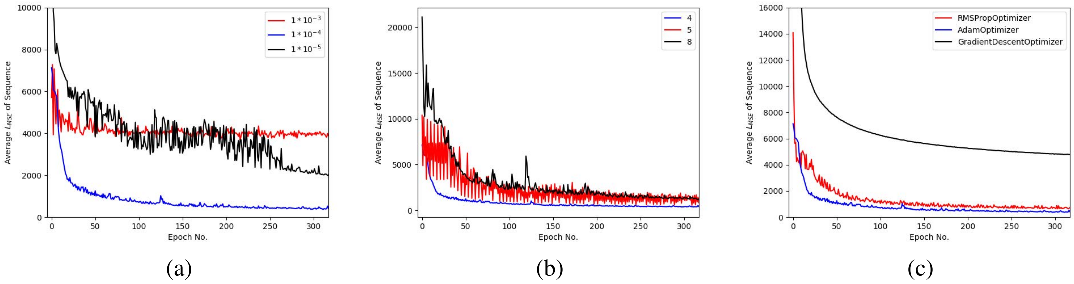
Table 2 presents the structure of the DF-ConvLSTM-VAE model. It should be noted that the output data of C3 are sent to the left flow (C3 → **L**-FC7) and the right flow (C3 → **R**-ConvLSTM4). The DF-ConvLSTM-VAE model concatenates the outputs of three recurrent ConvLSTM layers of the right flow, and sends them to next layer (**L**-ConvLSTM4, **L**-ConvLSTM5, **L**-ConvLSTM6 → **R**-FC7, **R**-FC8). The corresponding parameters of the DF-ConvLSTM-VAE model are the same as those of the ConvLSTM-VAE(Asymmetric) model.

C. EXPERIMENTAL RESULTS

1) DF-CONVLSTM-VAE VS. OTHER MODELS

Figures 7–9 describe the comparison of average $\mathcal{L}_{MSE}$ and KL divergence of sequence with different models on three datasets. In detail, the Figure 8(b) is a partial magnification of the Figure 8(c). The Figure 9(b) also is a partial magnification of the right Figure 9(c). From Figures 7–9, it is easy to see that the average $\mathcal{L}_{MSE}$ of sequence curve and the average KL divergence of sequence curve of the ConvLSTM-VAE(Symmetric) model are obviously different from the other three models.

From three Figures 7(a), 8(a) and 9(a), in the early training process, we find that the Convlstm-VAE(Symmetric) model tends to fall into local optima or saddle point, and lingers for a long time before jumping out and continuing to optimize. The other three models do not present this phenomenon. Obviously, the convergence rate of ConvLSTM-VAE(Symmetric) model is slower than that of the other three models. In addition, non-convergence sometimes occurs when the ConvLSTM-VAE(Symmetric) model is trained on the AVENUE dataset.



(a) (b) (c)

FIGURE 4. Comparison of average $\mathcal{L}_{MSE}$ of sequence of the ConvLSTM-VAE(Asymmetric) model with respect to different learning rate(a), mini-batch(b) and optimizer(c) on USCD ped1 dataset.

TABLE 1. Specifications of the ConvLSTM-VAE(Asymmetric) model. I=input layer, C=convolutional layer, ConvLSTM= convolutional long short-term memory, FC=fully connected layer, S=sampling layer, D=deconvolutional layer, O=output layer.

Layer	Input	Kernel	Stride/Pad	Output	Last/Next Layer
O14	224×224×1				
D13	56×56×128	11×11	4/0	224×224×1	D12/O14
D12	28×28×64	5×5	2/0	56×56×128	D11/D13
D11	28×28×32	3×3	1/0	28×28×64	FC10/D12
FC10	1×256			1×25088//(28×28×32)	S9/D11
S9	1×256			1×256	FC7,FC8/FC10
FC8	1×75264//(28×28×32×3)			1×256	ConvLSTM4,5,6/S9
FC7	1×75264//(28×28×32×3)			1×256	ConvLSTM4,5,6/S9
ConvLSTM6	28×28×32	5×5	1/0	28×28×32	ConvLSTM5/FC7,FC8
ConvLSTM5	28×28×32	5×5	1/0	28×28×32	ConvLSTM4/ConvLSTM6,FC7,FC8
ConvLSTM4	28×28×32	5×5	1/0	28×28×32	C3/ConvLSTM5,FC7,FC8
C3	28×28×64	3×3	1/0	28×28×32	C2/ConvLSTM4
C2	56×56×128	5×5	2/0	28×28×64	C1/C3
C1	224×224×1	11×11	4/0	56×56×128	I0/C2
I0	224×224×1			1×224×224	

TABLE 2. The table provides corresponding structure of the DF-ConvLSTM-VAE model. R-=right flow, L-=left flow, I=input layer, C=convolutional layer, ConvLSTM= convolutional long short-term memory, FC=fully connected layer, S=sampling layer, D=deconvolutional layer, O=output layer.

Left Flow		Right Flow	
Layer	Last/Next Layer	Last/Next Layer	Layer
O14			O14
D13	D12/O14		D13
D12	D11/D13		D12
D11	**L**-FC10, **R**-FC10/D12		D11
L-FC10	**L**-S9/D11	**R**-S9/D11	**R**-FC10
L-S9	**L**-FC7, **L**-FC8/**L**-FC10	**R**-FC7,**R**-FC8/**R**-FC10	**R**-S9
L-FC8	C3/**L**-S9	**R**-ConvLSTM4,5,6/**R**-S9	**R**-FC8
L-FC7	C3/**L**-S9	**R**-ConvLSTM4,5,6/**R**-S9	**R**-FC7
L-		**R**-ConvLSTM5 /**R**-FC7,**R**-FC8	**R**-ConvLSTM6
L-		**R**-ConvLSTM4/**R**-ConvLSTM6	**R**-ConvLSTM5
L-		C3/**R**-ConvLSTM5	**R**-ConvLSTM4
C3	C2/**L**-FC7,**L**-FC8, **R**-ConvLSTM4		C3
C2	C1/C3		C2
C1	I0/C2		C1
I0			I0

z-dim	AUC/EER(%)
128	87.0/18.6
256	88.2/17.6
512	86.4/19.1
1024	81.3 /25.2

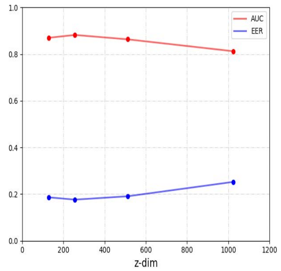


FIGURE 5. Comparison of area under ROC curve(AUC) and Equal Error Rate(EER) of different dimension of the learned hidden representation *z* on USCD ped1 dataset.

As can be seen from three Figures 7(a), 8(a) and 9(a), the average $\mathcal{L}_{MSE}$ of sequence curve of the DF-ConvLSTM-VAE model is at the bottom compared to the other curves. From Figures 7(b), 8(b) and 9(b), it is obvious that the average KL divergence of sequence curve of the DF-ConvLSTM-VAE model lies between VAE and the ConvLSTM-VAE(Asymmetric). Therefore, the structural design of the DF-ConvLSTM-VAE model composed of the VAE and the ConvLSTM-VAE(Asymmetric) model is effective. Obviously, compared with the ConvLSTM-VAE(Asymmetric) model, which can avoid falling into the saddle point for a long time, the training time of the DF-ConvLSTM-VAE model is relatively less.

In Table 3, these four models are experimented on three test datasets. Overall, the experimental results show that the result of ConvLSTM-VAE(Symmetric) model is better than the other three models. The performance of the DF-ConvLSTM-VAE model is better than ConvLSTM-VAE(Asymmetric) and VAE models. Although the performance of the DF-ConvLSTM-VAE model is not the

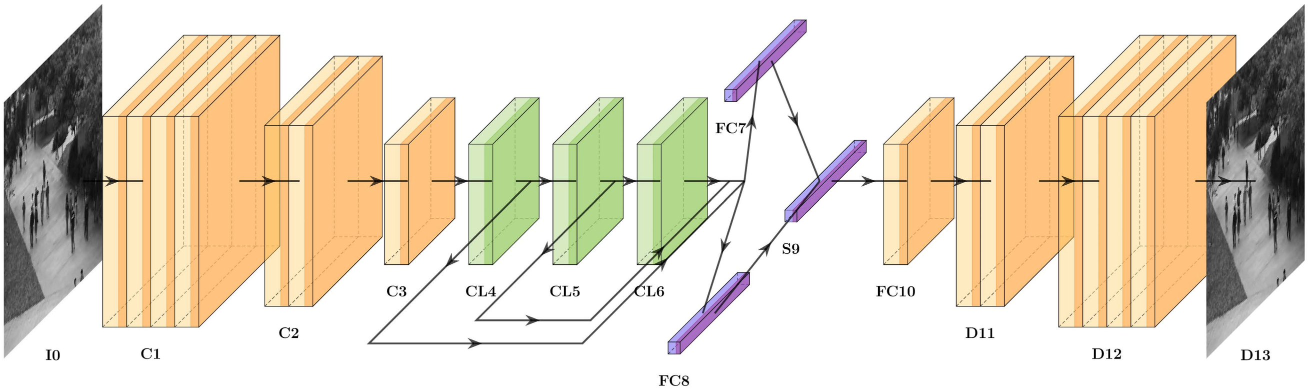


FIGURE 6. The architecture of the ConvLSTM-VAE(Asymmetric) model. I=input layer, C=convolutional layer, CL=convolutional long short-term memory, FC=fully connected layer, S=sampling layer, D=deconvolutional layer.

TABLE 3. Comparison of area under ROC curve(AUC) and Equal Error Rate(EER) of different models.

Method	AUC/EER(%)		
	Ped1	Ped2	Avenue
VAE	87.9/17.0	86.8/12.9	86.8/19.3
ConvLSTM-VAE(Symmetric)	**89.4/16.5**	**88.9**/14.1	87.0/**18.5**
ConvLSTM-VAE(Asymmetric)	88.2/17.6	87.6/15.1	86.8/18.7
DF-ConvLSTM-VAE	88.4/16.7	88.8/**12.2**	**87.2**/18.9

TABLE 4. The time consumed by different models.

Model	Running Time(in seconds)
VAE	0.0011
ConvLSTM-VAE(Symmetric)	0.0017
ConvLSTM-VAE(Asymmetric)	0.0012
DF-ConvLSTM-VAE	0.0015

best, the performance of the DF-ConvLSTM-VAE model for anomaly detection is worth considering and selecting in terms of training and time consumption.

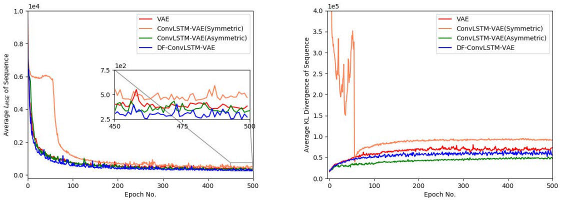
We implement our four models using the Tensorflow Framework. All the test experiments are conducted on a GPU GeForce RTX 2080 Ti. We test the time(in seconds) consumed per frame by these four models on USCD ped1 dataset, and the results are shown in Table 4. The running time taken by the ConvLSTM-VAE(Symmetric) model is the longest, due to its two Symmetric ConvLSTM layers. Our DF-ConvLSTM-VAE model has two sampling processes in the data stream of each frame and thus, takes longer time than that of the ConvLSTM-VAE (Asymmetric) model, but it is less time-consuming than that of the ConvLSTM-VAE (Symmetric) model with one sampling process.

2) QUANTITATIVE ANALYSIS: ROC AND ANOMALOUS EVENT COUNT

Table 5 compares the anomaly detection accuracy of our DF-Convlstm-VAE model against other state-of-the-art methods on three datasets. In Table 5, Adam, SF, MPPCA, MPPCA+SF, and HOFME are traditional methods. It is easy to see that our DF-Convlstm-VAE method is significantly better than these traditional methods in terms of AUC and EER on USCD dataset.

TABLE 5. Comparison of area under ROC curve(AUC) and Equal Error Rate(EER) of different methods."−"denotes the value is not published in their corresponding articles.

Method	AUC/EER(%)		
	Ped1	Ped2	Avenue
Adam [43]	77.1/38.0	−/42.0	−
SF [44]	67.5/31.0	55.6/42.0	−
MPPCA [12]	66.8/40.0	69.3/36.0	−
MPPCA+SF [12]	74.2/32.0	61.3/36.0	−
HOFME [10]	72.7/33.1	87.5/20.0	−
ConvAE [13]	81.0/27.9	90.0/21.7	70.2/25.1
ST-AE [29]	89.9/**12.5**	87.4/12.0	80.3/20.7
Two-stage [45]	77.8/29.2	**96.4/8.9**	85.3/23.9
ISTL [46]	75.2/29.8	91.1/**8.9**	76.8/29.2
Ada-Net [47]	90.4/15.8	90.3/15.5	**89.2/17.6**
ST-CaAE [48]	**90.5**/18.8	92.9/12.7	83.5/23.5
DF-ConvLSTM-VAE	88.4/16.7	88.8/12.2	87.2/18.9



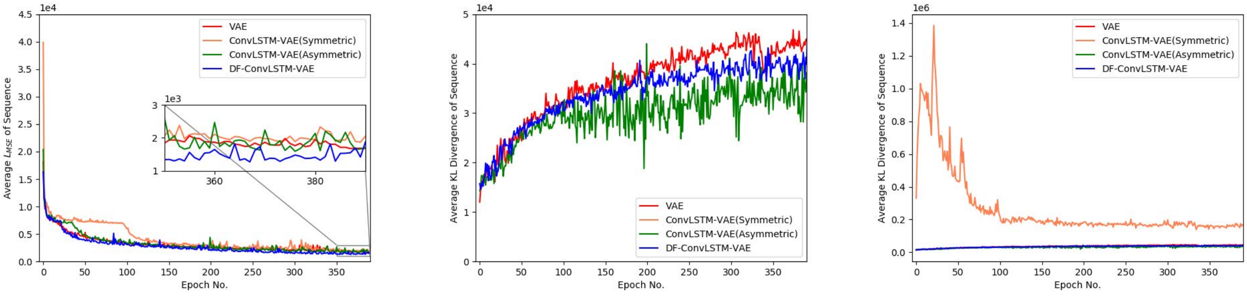
(a) (b)

FIGURE 7. Comparison of average $\mathcal{L}_{MSE}$ (a) and KL divergence(c) of sequence with different models on UCSD Ped1 datasets. (b) is a partial magnification of (c).

In Table 5, ConvAE, ST-AE, two-stage, and ISTL are unsupervised deep learning methods, where ConvAE, ST-AE, ISTL and our DF-Convlstm-VAE algorithm belong to a class of one stage models. Comparing these four models, our algorithm ranked second on USCD ped1 dataset and first on AVENUE dataset in terms of AUC and EER. The two-stage, Ada-net, and ST-CaAE models are more complex networks, where Ada-net and ST-CaAE networks have a high complexity because they are designed with GAN model. In particular, ST-CaAE uses extra optical flow information, and employs 2D/3D convolution methods to extract short-time temporal-spatial features, and integrates classical dual-flow model for video anomaly detection. Our algorithm uses Convsltm units to extract long-time temporal-spatial features of

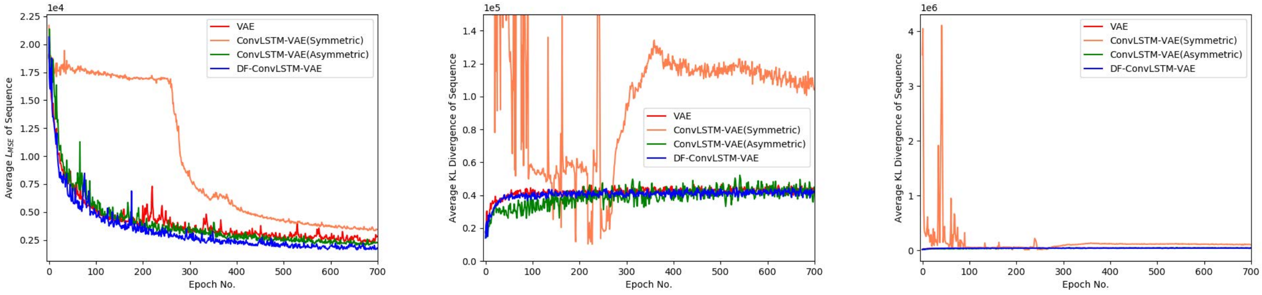
TABLE 6. Anomalous event and false alarm count detected by different methods.

Dataset	Anomalous Event	True Positive/False Alarm				
		Conv-AE [13]	**CC** [28]	**ST-AE** [29]	Ada-Net [47]	**DF-ConvLSTM-VAE**
Ped1	40	38/6	40/7	–	40/6	**40**/7
Ped2	12	12/1	12/1	–	12/1	**12/1**
Avenue	47	45/4	44/6	40/2	45/10	**45/2**



(a) (b) (c)

FIGURE 8. Comparison of average $\mathcal{L}_{MSE}$ (a) and KL divergence(c) of sequence with different models on UCSD ped2 dataset. (b) is a partial magnification of (c).



(a) (b) (c)

FIGURE 9. Comparison of average $\mathcal{L}_{MSE}$ (a) and KL divergence(c) of sequence with different models on Avenue dataset. (b) is a partial magnification of (c).

video sequences without using additional optical flow information which increases the computation. Compared with the ST-CaAE model, our DF-Convlstm-VAE algorithm performs well on EER. Compared to these state-of-the-art deep learning methods, our algorithm ranks third in terms of EER on USCD datasets and second in terms of AUC and EER on AVENUE dataset. In summary, our DF-Convlstm-VAE model is competitive in EER compared with other advanced deep learning models.

The comparions of anomalous events and false alarm counts are provided in Table 6. We employ our DF-Convlstm-VAE model to calculate true positive and false alarm by Persistence1D [39] algorithm. Observing Table 6, it is obvious that our algorithm performs very well in three datesets aspects of True Positive. As for False Alarm, our algorithm performs well on Ped2 and Avenue, except in Ped1 dataset. In summary, the performance of our DF-Convlstm-VAE model is comparable to the state-of-the-art anomalous event detection methods.

As seen in Table 5 and Table 6, compared with other state-of-the art methods, our DF-ConvLSTM-VAE model has competitive advantages.

3) QUALITATIVE ANALYSIS

a: VISUALIZING THE RECONSTRUCTED IMAGES

Figures 10–12 show three examples of generated videos by our DF-ConvLSTM-VAE network, and there are anomalous objects on these ground truth video sequences.

In Figure 10, the first and the third rows are the ground truth video sequences of frames 70 – 80 from UCSD Ped1 testing clip #20, while the second and the fourth rows show the corresponding reconstructed images. We can observe that the pedestrians in the generated images are different from the ground truth images, because the data generated by the network is different from the original dataset but with the same distribution. The network can pay attention to the spatio-temporal characteristics of learning videos and generates continuous foreground information. By observing this figure, the ground truth images show a person in a wheelchair, and at same position, a walking person is generated by our DF-ConvLSTM-VAE model in reconstructed images.

In Figure 11, The first and the third rows are the ground truth video sequences of frames 70 – 80 from UCSD Ped2 testing clip #4, while the second and the fourth rows show the corresponding reconstructed images. We can see that

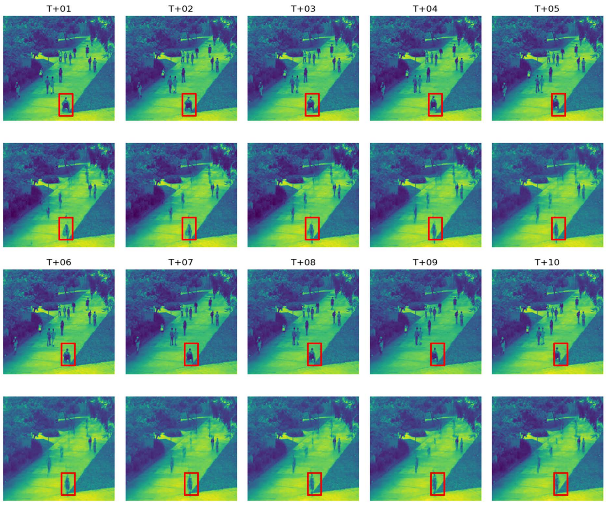


FIGURE 10. The first and third rows are the ground truth video sequences of frames 70-80 from USCDped1 testing clip #20, while the second and fourth rows are corresponding reconstructed images.

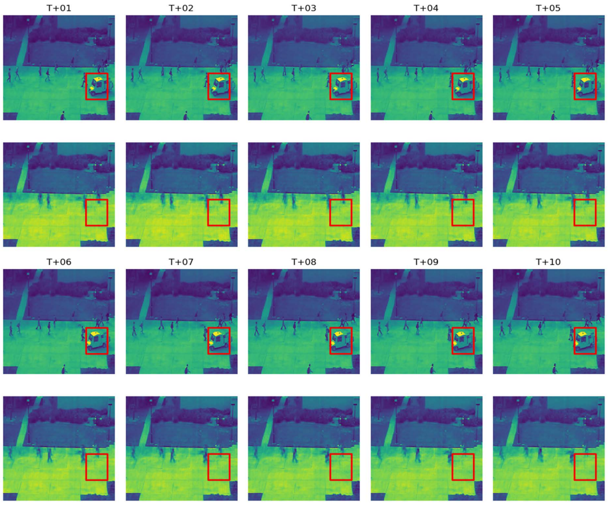


FIGURE 11. The first and the third rows are the ground truth video sequences of frames 50-60 from USCDped2 testing clip #4, while the second and the fourth rows are corresponding reconstructed images.

the ground truth image shows a moving truck while our DF-ConvlSTM-VAE model does not produce any results in the reconstructed images at the same place.

In Figure 12, the first and the third rows are the ground truth video sequences of frames $20 - 30$ from Avenue testing clip #20, while the second and the fourth rows show the corresponding reconstructed images. We can see that a person walking in the wrong direction(walking toward the camera) in the ground truth video. This behavior does not occur in the training set, and therefore, in the generated images, nothing is generated in the corresponding position by our DF-ConvlSTM-VAE model. In addition, observe the ground truth video, we can see that the pillar is obscured by the abnormal object, but it is generated well in our generated images by our DF-ConvlSTM-VAE model. This is because the essence of VAE-based model is a probabilistic graphical model.

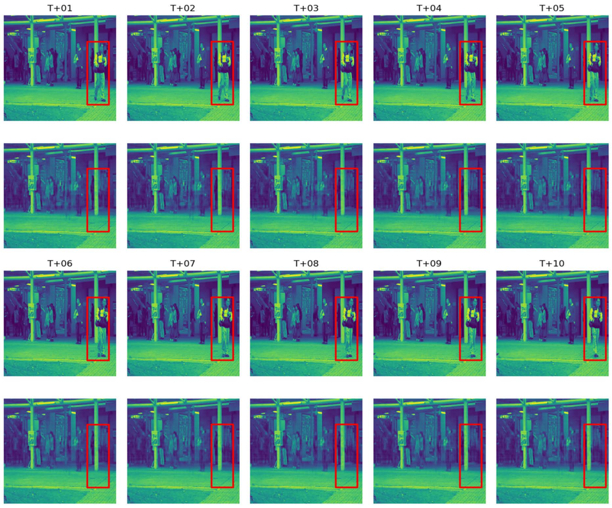


FIGURE 12. The first and third rows are the ground truth video sequences of frames 20-30 from Avenue testing clip #19, while the second and fourth rows are corresponding reconstructed images.

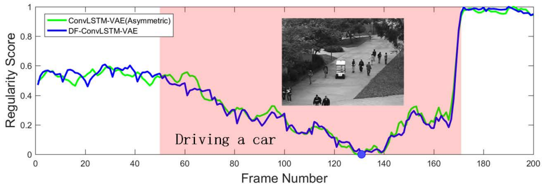


FIGURE 13. Regularity score of video #24 from UCSD ped1 dataset.

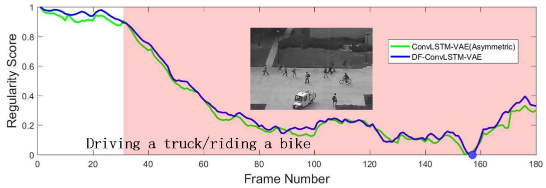


FIGURE 14. Regularity score of video #4 from UCSD ped1 dataset.

Since the distribution of the generated samples with the DF-ConvLSTM-VAE model is the same as and similar to the training datasets, there is no anomalous object in our reconstructed images.

b: VISUALIZING TEMPORAL REGULARITY

In Figures 13–15, we compare our two asymmetric models in terms of the regularity scores on different datasets clips. The anomalous ground truth regions are highlighted in red, and distinct local minima is represented by a blue dot. The lower the regularity score value under the anomalous conditions, the higher the curve value in normal circumstances, indicating that the performance of model is better.

Figure 14 shows that the capability of the DF-ConvLSTM-VAE model is stronger than that of the ConvLSTM-VAE(Asysmetric) model. There are two anomalous objects(a moving truck and a person with bike) in video #4. When two anomalies occur at the same time, the curve only shows that the video frame is anomalous, but cannot indicate that there exist two anomalous objects on this video

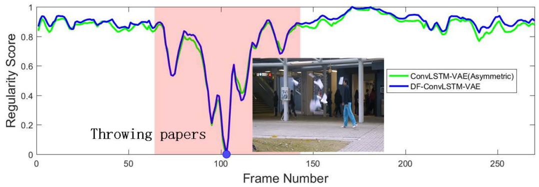


FIGURE 15. Regularity score of video #20 from Avenue dataset.

frame. In Figure 15, the curves show that both of our two asymmetric models can detect the anomalous behavior of a person throwing papers into sky.

From these figures, it is easy to see that when there are irregular motions, the regular score curve drops significantly and forms a nail shape, and the performance of our DF-ConvlSTM-VAE model is slightly better than the ConvlSTM-VAE(Asymmetric) model.

V. DISCUSSION

Although this method takes the whole video frame as the input, it is very advantageous for extracting global features, but when extracting features, we find that the size of the foreground target is relatively small, which brings challenges to extracting the detail features of targets. Therefore, in the subsequent study, we suggest to fully consider removing background information unrelated to the foreground and extract relevant features in the form of patch.

VI. CONCLUSION AND FUTURE WORK

In this paper, both the ConvLSTM-VAE(Asymmtric) model and the DF-ConvLSTM-VAE model consist of ConvLSTM and VAE, and are proposed to learn training data distribution for video anomaly detection. The ConvLSTM-VAE(Asymmetric) model is designed by weakening the decoder. Compared with the ConvLSTM-VAE(Symmetric) model, the ConvLSTM-VAE(Asymmetric) model has some advantages in terms of training time and difficulty. Experiments show that the DF-ConvLSTM-VAE model is superior to the ConvLSTM-VAE(Asymmtric) model. Compared with other typical methods, the experiments verify the validity and competitiveness of our DF-ConvlSTM-VAE on multiple public benchmark data sets. Since the simple gaussian model cannot meet the complexity of real data, in the future, we will try to construct a new probability graph model to accomplish this task by forcing the representation z to obey a more complex model.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their constructive and valuable comments and suggestions.

REFERENCES

[1] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita, and F. Herrera, "Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105590.

[2] T. Hayashi and H. Fujita, "Cluster-based zero-shot learning for multivariate data," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 2, pp. 1897–1911, Feb. 2021.

[3] S. Li, Z. Chen, X. Li, J. Lu, and J. Zhou, "Unsupervised variational video hashing with 1D-CNN-LSTM networks," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1542–1554, Jun. 2020.

[4] B. Fan, H. Liu, H. Zeng, J. Zhang, X. Liu, and J. Han, "Deep unsupervised binary descriptor learning through locality consistency and self distinctiveness," *IEEE Trans. Multimedia*, vol. 23, pp. 2770–2781, 2021.

[5] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, "Model adaptation: Unsupervised domain adaptation without source data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9638–9647.

[6] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Evolving losses for unsupervised video representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 130–139.

[7] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image Vis. Comput.*, vol. 106, no. 6, Feb. 2021, Art. no. 104078.

[8] B. Ramachandra, M. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2293–2312, May 2022.

[9] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Pers. Ubiquitous Comput.*, vol. 3, pp. 1–17, 2021.

[10] T. Wang and H. Snoussi, "Histograms of optical flow orientation for abnormal events detection," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, vol. 5, Jan. 2013, pp. 13–18.

[11] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1446–1453.

[12] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975–1981.

[13] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[14] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal AutoEncoder for video anomaly detection," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1933–1941.

[15] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 843–852.

[18] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level CNN: Saliency-aware 3-D CNN with LSTM for video action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 510–514, Apr. 2017.

[19] H. Ge, Z. Yan, W. Yu, and L. Sun, "An attention mechanism based convolutional LSTM network for video action recognition," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 20533–20556, Jul. 2019.

[20] N. Zhuang, J. Ye, and K. A. Hua, "DLSTM approach to video modeling with hashing for large-scale video retrieval," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3222–3227.

[21] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval," *Future Gener. Comput. Syst.*, vol. 93, pp. 583–595, Apr. 2019.

[22] A. Pfeuffer and K. Dietmayer, "Separable convolutional LSTMs for faster video segmentation," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Auckland, New Zealand, Oct. 2019, pp. 1072–1078.

[23] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, and R. Klette, "STFCN: Spatio-temporal fully convolutional neural network for semantic segmentation of street scenes," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Cham, Switzerland: Springer, 2016, pp. 493–509.

[24] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[25] Y. Yang, J. Zhou, J. Ai, Y. Bin, and A. Hanjalic, "Video captioning by adversarial LSTM," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5600–5611, Nov. 2018.

[26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 4, 2014, pp. 3104–3112.

[27] B. Lindemann, B. Maschler, N. Sahlab, and M. Weyrich, "A survey on anomaly detection for technical systems using LSTM networks," *Comput. Ind.*, vol. 131, no. 3, Oct. 2021, Art. no. 103498.

[28] J. Ryan Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," 2016, *arXiv:1612.00390*.

[29] S. C. Yong and H. T. Yong, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.*, 2017, pp. 189–196.

[30] L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan, "Abnormal event detection in videos using hybrid spatio-temporal autoencoder," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2276–2280.

[31] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2014.

[32] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture IE*, vol. 2, pp. 1–18, Dec. 2015.

[33] *MNIST Dataset*. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[34] M. Sölc, J. Bayer, M. Ludersdorfer, and P. van der Smagt, "Variational inference for on-line anomaly detection in high-dimensional time series," in *Proc. Int. Conf. Learn. Represent. (ICLR) Workshops*, 2016.

[35] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder," *IEEE Robot. Automat. Lett.*, vol. 3, no. 3, pp. 1544–1551, Jul. 2018.

[36] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[37] *Moving-MNIST Dataset*. [Online]. Available: http://www.cs.toronto.edu/~nitish/unsupervised video/

[38] J. Geweke, "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica: J. Econ. Soc.*, vol. 57, no. 6, pp. 1317–1339, 1989.

[39] Y. Kozlov and T. Weinkauf. *Persistence1D: Extracting and Filtering Minima and Maxima of 1D Functions*. [Online]. Available: http://people.mpi-inf.mpg.de/~weinkauf/notes/persistence1d.html

[40] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2720–2727.

[41] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "BETA-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. 5th Int. Conf. Learn. Represent.*, 2017, pp. 1–11.

[42] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS), Disentanglement Workshop*, 2017.

[43] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[44] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.

[45] S. Wang, Y. Zeng, Q. Liu, C. Zhu, E. Zhu, and J. Yin, "Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection," in *Proc. 26th ACM Int. Conf. Multimedia*. Seoul, South Korea, Oct. 2018, pp. 636–644.

[46] R. Nawaratne, D. Alahakoon, D. D. Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 393–402 Jan. 2020.

[47] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia, "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2138–2148, Aug. 2020.

[48] N. Li, F. Chang, and C. Liu, "Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes," *IEEE Trans. Multimedia*, vol. 23, pp. 203–215, 2021.

LIN WANG (Member, IEEE) received the B.S. degree in mathematics and applied mathematics from Weinan Normal University, Shaanxi, China, in 2010, and the M.S. degree in mathematics from the North China University of Technology, Beijing, China, in 2015. She is currently pursuing the Ph.D. degree with the School of Instrumentation and Opto-Electronic Engineering, Beihang University, Beijing. Her current research interests include image processing and deep learning.

HAISHU TAN (Member, IEEE) received the B.S. and Ph.D. degrees in optical engineering from Tianjin University, China, in 1994 and 1998, respectively. He is currently a Professor with the School of Physics and Optoelectronic Engineering, Foshan University, China. His research interests include computer vision, photoelectric measurement, and optical metrology.

FUQIANG ZHOU received the B.S., M.S., and Ph.D. degrees in instrument, measurement, and test technology from Tianjin University, China, in 1994, 1997, and 2000, respectively. In 2000, he joined the School of Automation Science and Electrical Engineering, Beihang University, China, as a Postdoctoral Research Fellow. He is currently a Professor with the School of Instrumentation and Opto-Electronics Engineering, Beihang University. His research interests include precision vision measurement, 3-D vision sensors, image recognition, and optical metrology.

WANGXIA ZUO received the B.S. degree in electric automatization from the Wuhan University of Hydraulic and Electric Engineering, Wuhan, China, in 2001, the M.S. degree in control theory and control engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2006, and the Ph.D. degree in measurement technology and instruments from Beihang University, Beijing, China, in 2022. She is currently a Lecturer with the School of Electrical Engineering, University of South China. Her current research interests include image processing and deep learning.

PENGFEI SUN received the B.S. and M.S. degrees in measuring, testing technologies and instruments from Henan Polytechnic University, Henan, China, in 2012 and 2015, respectively. She is currently pursuing the Ph.D. degree with the School of Instrumentation and Opto-Electronic Engineering, Beihang University, Beijing, China. Her current research interests include machine vision and precision measurement.