Convolutional Transformer based Dual Discriminator
Generative Adversarial Networks for Video Anomaly Detection

Xinyang Feng
Columbia University
New York, New York, USA
xf2143@columbia.edu

Dongjin Song
University of Connecticut
Storrs, Connecticut, USA
dongjin.song@uconn.edu

Yuncong Chen
NEC Laboratories America, Inc.
Princeton, New Jersey, USA
yuncong@nec-labs.com

Zhengzhang Chen
NEC Laboratories America, Inc.
Princeton, New Jersey, USA
zchen@nec-labs.com

Jingchao Ni
NEC Laboratories America, Inc.
Princeton, New Jersey, USA
jni@nec-labs.com

Haifeng Chen
NEC Laboratories America, Inc.
Princeton, New Jersey, USA
haifeng@nec-labs.com

ABSTRACT
Detecting abnormal activities in real-world surveillance videos is an important yet challenging task as the prior knowledge about video anomalies is usually limited or unavailable. Despite that many approaches have been developed to resolve this problem, few of them can capture the normal spatio-temporal patterns effectively and efficiently. Moreover, existing works seldom explicitly consider the local consistency at frame level and global coherence of temporal dynamics in video sequences. To this end, we propose Convolutional Transformer based Dual Discriminator Generative Adversarial Networks (CT-D2GAN) to perform unsupervised video anomaly detection. Specifically, we first present a convolutional transformer to perform future frame prediction. It contains three key components, i.e., a convolutional encoder to capture the spatial information of the input video clips, a temporal self-attention module to encode the temporal dynamics, and a convolutional decoder to integrate spatio-temporal features and predict the future frame. Next, a dual discriminator based adversarial training procedure, which jointly considers an image discriminator that can maintain the local consistency at frame-level and a video discriminator that can enforce the global coherence of temporal dynamics, is employed to enhance the future frame prediction. Finally, the prediction error is used to identify abnormal video frames. Thoroughly empirical studies on three public video anomaly detection datasets, i.e., UCSD Ped2, CUHK Avenue, and Shanghai Tech Campus, demonstrate the effectiveness of the proposed adversarial spatio-temporal modeling framework.

CCS CONCEPTS
• Computing methodologies → Scene anomaly detection;
Adversarial learning; Anomaly detection; Neural networks.

1 INTRODUCTION
With the rapid growth of video surveillance data, there is an increasing demand to automatically detect abnormal video sequences in the context of large-scale normal (regular) video data. Despite a substantial amount of research effort has been devoted to this problem [3, 8, 13, 14, 16, 19, 22, 31, 34], video anomaly detection, which aims to identify the activities that do not conform to regular patterns in a video sequence, is still a challenging task. This is because real-world abnormal video activities can be extremely diverse while the prior knowledge about these anomalies is usually limited or even unavailable.

With the assumption that a model can only generalize to data from the same distribution as the training set, abnormal activities in the test set will manifest as deviance from regular patterns. A common approach to resolve this problem is to learn a model that can capture regular patterns in the normal video clips during the training stage, and check whether there exists any irregular pattern that diverges from regular patterns in the test video clips. Within this framework, it is crucial to not only represent the regular appearances but also capture the normal spatio-temporal dynamics to differentiate abnormal activities from normal activities in a video sequence. This serves as an important motivation for our proposed methods.

Early studies have used handcrafted features to represent video patterns [13, 16, 19, 29]. For instance, Li et al. [13] introduced mixtures of dynamic textures and defined outliers under this model as anomalies. These approaches, however, are usually not optimal for video anomaly detection since the features are extracted based upon a different objective.
Recently, deep neural networks are becoming prevalent in video anomaly detection, showing superior performance over handcrafted feature-based methods. For instance, Hasan et al. [8] developed a convolutional autoencoder (Conv-AE) to model the spatio-temporal patterns in a video sequence simultaneously with a 2D CNN. The temporal dynamics, however, are not explicitly considered. To better cope with the spatio-temporal information in a video sequence, convolutional long short-term memory (LSTM) autoencoder (ConvLSTM-AE) [17, 27] was proposed to model the spatial patterns with fully convolutional networks and encode the temporal dynamics using convolutional LSTM (ConvLSTM). ConvLSTM, however, suffers from computational and interpretation issues. A powerful alternative for sequence modeling is the self-attention mechanism [33]. It has demonstrated superior performance and efficiency in many different tasks, e.g., sequence-to-sequence machine translation [33], time series prediction [24], autoregressive model based image generation [23], and GAN-based image synthesis [39]. However, it has seldom been employed to capture regular spatio-temporal patterns in the surveillance videos.

More recently, adversarial learning has shown impressive progress on video anomaly detection. For instance, Ravanbakhsh et al. [25] developed a GAN based anomaly detection approach following conditional GAN framework [10]. Liu et al. [14] proposed an anomaly detection approach based on future frame prediction. Tang et al. [31] extended this framework by adding a reconstruction task. The generative models in these two works were based on U-Net [26]. Similar to Conv-AE, the temporal dynamics in the video clip were not explicitly encoded and the temporal coherence was enforced by a loss term on the optical flow. Moreover, the potential discriminative information in the form of consistency at frame-level and global coherence of temporal dynamics in video sequences were not fully considered in previous works.

In this paper, to better capture the regular spatio-temporal patterns and cope with the potential discriminative information at frame-level and in video sequences, we propose Convolutional Transformer based Dual Discriminator Generative Adversarial Networks (CT-D2GAN) to perform unsupervised video anomaly detection. We first present a convolutional transformer to perform future frame prediction. The convolutional transformer is essentially a encoder-decoder framework consisting of three key components, i.e., a convolutional encoder to capture the spatial patterns of the input video clip, a novel temporal self-attention module adapted for video temporal modeling that can explicitly encode the temporal dynamics, and a convolutional decoder to integrate spatio-temporal features and predict the future frame. Because of the temporal self-attention module, convolutional transformer can capture the underlying temporal dynamics efficiently and effectively. Next, in order to maintain the local consistency of the predicted frame and the global coherence conditioned on the previous frames, we adapt dual discriminator GAN to deal with video frames and employ an adversarial training procedure to further enhance the prediction performance. Finally, the prediction error is adopted to identify abnormal video frames. Thoroughly empirical studies on three public video anomaly detection datasets, i.e., UCSD Ped2, CUHK Avenue, and Shanghai Tech Campus, demonstrate the effectiveness of the proposed framework and techniques.

2 RELATED WORK

The proposed Convolutional Transformer based Dual Discriminator Generative Adversarial Networks (CT-D2GAN) is closely related to deep learning based video anomaly detection and self-attention mechanism [33].

Note that we focus our discussions on methods based on unsupervised settings, which are efficient in generalization without the time-consuming and error-prone process of manual labeling. We are aware that there are numerous works on weakly supervised or supervised video anomaly detection, e.g., Sultani et al. (2018) proposed a deep multiple instance ranking framework using video-level labels and achieves better performance than convolutional autoencoder (Conv-AE) based method [8], but it employs both normal and abnormal video clips for training which is different from our setting.

Deep neural networks based video anomaly detection methods demonstrate superior performance over traditional methods based on handcrafted features. Hasan et al. (2016) developed Conv-AE method to simultaneously learn the spatio-temporal patterns in a video with 2D convolutional neural networks by concatenating the video frames in the channel dimension. The temporal information is mixed with the spatial information in the first convolutional layer, thus not explicitly encoded. Xu et al. (2017) proposed appearance and motion DeepNet (AMDN) to learn video feature representations, which however still requires a decoupled one-class SVM classifier applied on learned representation to generate anomaly score. Dong et al. (2019) proposed a memory-augmented autoencoder (MemAE) that uses a memory module to constrain the reconstruction.

More recently, adversarial learning has demonstrated flexibility and impressive performance in multiple video anomaly detection studies. A generative adversarial networks (GANs) based anomaly detection approach [25] was developed following cGAN framework of image-to-image translation [10]. Specifically, it employs image and optical flow as source domain and target domain, and vice versa, and trains cross-channel generation through adversarial learning. The reconstruction error is used to compute anomaly score. The only temporal constraint is imposed by the optical flow calculation. Liu et al. (2018) proposed an anomaly detection approach based on future frame prediction in GAN framework and U-Net [26]. Similar to Conv-AE, the temporal information is not explicitly encoded and the temporal coherence between neighboring frames is enforced by a loss term on the optical flow. Tang et al. (2020) extended the future frame prediction framework by adding a reconstruction task. One way to alleviate the temporal encoding issue in video spatio-temporal modeling is to use convolutional LSTM autoencoder (ConvLSTM-AE) based methods [4, 17, 27, 38], where the spatial and temporal patterns are encoded with fully convolutional networks and convolutional LSTM, respectively. Despite its popularity, ConvLSTM suffers from issues such as large memory consumption. The complex gating operations add to the computational cost and complicate the information flow, making interpretation difficult.

A more effective and efficient alternative for sequence modeling is the self-attention mechanism [33], which is essentially an attention mechanism relating different positions of a single sequence to compute a representation of the sequence, in which the keys, values,
and queries are from the same set of features. Some related applications include autoregressive model based image generation [23], GAN-based image synthesis [39].

In this work, based on related works, we introduce the convolutional transformer by extending the self-attention mechanism to video sequence modeling and develop a novel self-attention module specialized for spatio-temporal modeling in video sequences. Compared to existing approaches for video anomaly detection, the proposed convolutional transformer model has the advantage of being able to explicitly and efficiently encode the temporal information in a sequence of feature maps, where the computation of attentions can be fully parallelized via matrix multiplications. Based on the convolutional transformer, a dual discriminator generative adversarial networks (D2GAN) approach is developed to further enhance the future frame prediction through enforcing local consistency of the predicted frame and the global coherence conditioned on the previous frames. Note that the proposed D2GAN differs from existing works on dual discriminator based GAN which have been applied to different scenarios [5, 21, 35, 37].

3 CT-D2GAN

In this section, we first introduce the problem formulation and input to our framework. Then, we present the motivation and technical details of the proposed CT-D2GAN framework including convolutional transformer, dual discriminator GAN, the overall loss function, and lastly the regularity score calculation. An overview of the framework is illustrated in Figure 1.

In CT-D2GAN, a convolutional transformer is employed to generate future frame prediction based on past frames, an image discriminator and a video discriminator are used to maintain the local consistency and global coherence.

3.1 Problem Statement

Given an input representation of video clip of length $T$, i.e., $I = (I_0, I_{+1}, ..., I_T) \in \mathbb{R}^{h \times w \times c \times T}$, where $h$, $w$, $c$ are the height, width and number of channels, we aim to predict the $(t + 1)$-th frame as $\hat{I}_{t+1} \in \mathbb{R}^{h \times w \times c}$ and identify abnormal activities based upon the prediction error, i.e., $\text{MSE}_{t} = \frac{1}{h \times w \times c} \sum_{t=1}^{T} \left\| \hat{I}_{t+1} - I_{t+1} \right\|_{F}^2$, where $I_{t+1} \in \mathbb{R}^{h \times w \times c}$.

3.2 Input

As appearance and motion are two characteristics of video data, it is common to explicitly incorporate optical flow together with the still images to describe a video sequence [28], e.g. optical flow has been employed to represent video sequences in the cGAN framework [25] and used as a motion constraint [14].

In this work, we stack image with pre-computed optical flow maps [2, 9] in channel dimension as inputs similar to Simonyan et al. [28] for video action recognition and Ravanbakhsh et al. [25] for video anomaly detection. The optical flow maps consist of a horizontal component, a vertical component and a magnitude component. To be noted that, the optical flow map is computed from the previous image and current image, thus does not contain future frame information. Therefore, the input can be given as $I \in \mathbb{R}^{h \times w \times 4 \times T}$, and we used 5 consecutive frames as inputs, i.e., $T = 5$, similar to Liu et al. [14].

3.3 Convolutional Transformer

Convolutional transformer is developed to obtain a future frame prediction based on past frames. It consists of three key components: a convolutional encoder to encode spatial information, a temporal self-attention module to capture the temporal dynamics, and a convolutional decoder to integrate spatio-temporal features and predict future frame.

3.3.1 Convolutional Encoder: The convolutional encoder [15] is employed to extract spatial features from each frame of the video. Each frame of the video is first resized to $256 \times 256$ and then fed into the convolutional encoder. The convolutional encoder consists of 5 convolutional blocks. And the convolutional block follows common structure in CNN. All the convolutional kernels are set as $3 \times 3$ pixels. For brevity, we denote a convolutional layer with stride $s$ and number of filters $n$ as $\text{Conv}_{s,n}$, a batch normalization layer as BN, a scaled exponential linear unit [12] as SELU, and a dropout operation with dropout ratio $r$ as dropout.

The structure of the convolutional encoder is: $[\text{Conv}_{1,64} \text{-SELU}-\text{BN}]-[\text{Conv}_{2,64} \text{-SELU}-\text{BN}-\text{Conv}_{1,64} \text{-SELU}]-[\text{Conv}_{2,128} \text{-SELU}-\text{BN}-\text{Conv}_{1,128} \text{-SELU}]-[\text{Conv}_{2,256} \text{-SELU}-\text{BN}-\text{dropout}_0.25]-[\text{Conv}_{1,256} \text{-SELU}]-[\text{Conv}_{2,256} \text{-SELU}-\text{BN}-\text{dropout}_0.25]-[\text{Conv}_{1,256} \text{-SELU}]$, where each [ ] represents a convolutional block.

At the $l$-th convolutional block $\text{conv}^l$, $F_{t-1} \in \mathbb{R}^{h \times w \times c_1}$, $t \in [0, ..., T - 1]$ denotes the input feature maps to the self-attention module with $h_t$, $w_t$, $c_t$ as the height, width, and number of channels, respectively. The temporal dynamics among the spatial feature maps of different time steps will be encoded with temporal self-attention module.

3.3.2 Temporal Self-attention Module: To explicitly encode the temporal information in the video sequence, we extend self-attention mechanism in the transformer model [33] and develop a novel temporal self-attention module to capture the temporal dynamics of the multi-scale spatial feature maps generated from the convolutional encoder. This section applies to all layers, thus we omit the layer for clarity. An illustration of the multi-head temporal self-attention module is shown in the upper panel of Figure 1, Spatial Feature Vector. We first use global average pooling (GAP) to extract a feature vector $f_t$ from the feature map $F_t$ extracted in the convolutional encoder. The feature vector in current time step $f_t$ will be used as part of the query and each historical feature vector $f_{t-i}$, $i \in [1, T - 1]$ will be used as part of the key to index spatial feature maps.

Positional Encoding. Different from sequence models such as LSTM, self-attention does not model sequential information inherently, therefore it is necessary to incorporate temporal positional information into the model. We generate a positional encoding vector $\text{PE} \in \mathbb{R}^{d_p}$ following [33]:

$$
\text{PE}_{p, 2i} = \sin(p/10000^{2i/d_p})
$$

$$
\text{PE}_{p, 2i+1} = \cos(p/10000^{2i/d_p})
$$

(1)

where $d_p$ denotes the dimension of PE, $p$ denotes the temporal position and $i \in [0, ..., (d_p/2 - 1)]$ denotes the index of the dimension. Empirically, we fix $d_p = 8$ in our study.

Temporal Self-Attention. We concatenate the positional encoding vector with the spatial feature vector for each time step and
use the concatenated vectors as the queries and keys, and the feature maps as the values in the setting of self-attention mechanism. For each query frame at time $t$, the current concatenated feature vector $q_t = [f_t; \text{PE}] \in \mathbb{R}^{c_t \times d_p}$ is used as query, and compared to the feature vector of each frame from the input video clip i.e. memory $m_{t-i} = [f_{t-i}; \text{PE}] \in \mathbb{R}^{c_{t-i} \times d_p}, i \in [1, T-1]$ using cosine similarity:

$$D(q_t, m_{t-i}) = \frac{q_t \cdot m_{t-i}}{\|q_t\| \|m_{t-i}\|}$$

(2)

Based on the similarity between $q_t$ and $m_{t-i}$, we can generate the normalized attention weights $a_{t,i} \in \mathbb{R}$ across the temporal dimension using a softmax function:

$$a_{t,i} = \frac{\exp(\beta D(q_t, m_{t-i}))}{\sum_{j=1}^{T-1} \exp(\beta D(q_t, m_{t-j}))}$$

(3)

where a positive temperature variable $\beta$ is introduced to sharpen the level of focus in the softmax function and is automatically learned in the model through a single hidden densely-connected layer with the query as the input.

The final attended feature maps $H_t$ are a weighted sum of all feature maps $F$ using the attention weights in Eq. (3):

$$H_t = \sum_{i=1}^{T-1} a_{t,i} \cdot F_{t-i}.$$  

(4)

Multi-head Temporal Self-Attention. Multi-head self-attention [33] enables the model to jointly attend to information from different representation subspaces at different positions. We adapt it to spatio-temporal modeling by first mapping the spatial feature maps to $n_h = 8$ groups, each using 32 $1 \times 1$ convolutional kernels. For each group of feature maps with dimension $c_h = 32$, we then perform the single head self-attention as described in the previous subsection and generate attended feature maps for head $k$
as $H_t^{(k)}$:

$$H_t^{(k)} = \sum_{i \in [1, ..., T-1]} a_{t, i}^{(k)} \cdot F_{t, i}^{(k)}, \quad (5)$$

where $F_{t, i}^{(k)} \in \mathbb{R}^{h \times w \times c_k}$ is the transformed feature map at frame $t - i$ for head $k$, $a_{t, i}^{(k)}$ is the corresponding attention weight. The final multi-head attended feature map $H_t^{MH} \in \mathbb{R}^{h \times w \times (c_\theta \cdot c_k)}$ is the concatenation of the attended feature maps from all the heads along the channel dimension:

$$H_t^{MH} = \text{Concat}(H_t^{(1)}, ..., H_t^{(n_h)}). \quad (6)$$

In this way, the final attended feature maps not only integrate spatial information from the convolutional encoder, but also capture temporal information from multi-head temporal self-attention mechanism.

**Spatial Selective Gate.** The aforementioned module extends the self-attention mechanism to the temporal modeling of 2D image feature maps, however, it comes with the loss of fine-grained spatial resolution due to the GAP operation. To compensate this, we introduce spatial selective gate (SSG), which is a spatial attention mechanism to integrate the current and historical information. The attended feature maps from the temporal self-attention module and the feature maps of the current query are concatenated, on which we learn a spatial selective gate using a sub-network $N_{SSG}$ with structure: Conv$_{1, 256}$-BN-SELU-Conv$_{1, 256}$-BN-SELU-Conv$_{1, 256}$-BN-SELU-Conv$_{1, 256}$-BN-SELU-Conv$_{1, 256}$-Sigmoid. The final output is a pixel-wise weighted average of the attended maps $H_t^{MH}$ and the current query’s multi-head transformed feature maps $F_t^{MH} \in \mathbb{R}^{h \times w \times (c_\theta \cdot c_k)}$, according to SSG:

$$S_t = \text{SSG} \circ F_t^{MH} + (1 - \text{SSG}) \circ H_t^{MH} \quad (7)$$

where $\circ$ denotes element-wise multiplication.

We add SSG at each level of temporal self-attention module. As the spatial dimensions are larger at shallower layers and we want to include contextual information while preserving the spatial resolution, we use dilated convolution [36] with different dilation factors at the 4 convolutional blocks in the sub-network $N_{SSG}$: specifically from conv$^2$ to conv$^5$, the dilation factors are (1,2,4,1), (1,2,2,1), (1,1,2,1), (1,1,1,1). Note that SSG is computationally more efficient than directly forwarding the concatenated feature maps to the convolutional decoder.

### 3.3.3 Convolutional Decoder

The outputs of the temporal self-attention module $S_t$ are fed into the convolutional decoder. The convolutional decoder predicts the video frame using 4 transposed convolutional layers with stride 2 on the feature maps in a reverse order of the convolutional encoder. The fully-scaled feature maps then go through one convolutional layer with 32 filters and one convolutional layer with $c$ filters of size $1 \times 1$ that maps to the same size of channels $c$ in the input. In order to predict finer details, we utilize the skip connection [26] to connect the spatio-temporally integrated maps at each level of the convolutional encoder to the corresponding level of the convolutional decoder, which allows the model to further fine-tune the predicted frames.

### 3.4 Dual Discriminator GAN

We propose a dual discriminator GAN using both an image discriminator and a video discriminator to further enhance the future frame prediction of convolutional transformer via adversarial training. The image discriminator $D_I$ critiques on whether the current frame is generated or real just on the basis of one single frame to assess the local consistency. The video discriminator $D_V$ performs critique on the prediction conditioned on the past frames to assess the global coherence. Specifically, we stack the past frames with current generated or real frame in the temporal dimension and the video discriminator is essentially a video classifier. This idea of combining local and global (contextual) discriminator is similar to adversarial image inpainting [37] but is used in a totally different context.

The network structures of the two discriminators are kept the same except that we use 2D operations in image discriminator and the corresponding 3D operations in the video discriminator. We use PatchGAN architecture as described in [10] and use spectral normalization [20] in each convolutional layer. In the 3D version, the stride and kernel size in the temporal dimension were set at 1 and 2 respectively.

The method in Liu et al. [14] is similar to using the image discriminator only. Different from the video discriminator in Tulyakov et al. [32], which applies on the whole synthetic video clip, our proposed video discriminator conditions on the real frames.

### 3.5 Loss

For the adversarial training, we use the Wasserstein GAN with gradient penalty (WGAN-GP) setting [1, 7]. The generator $G$ is the mapping $: I \rightarrow \tilde{I}_{t+1}$. For discriminators, $D_V : (I, \tilde{I}_{t+1}) \rightarrow p(I, \tilde{I}_{t+1})$ is real and $D_I : \tilde{I}_{t+1} \rightarrow p[\tilde{I}_{t+1}$ is real] are video and image discriminators respectively. The GAN loss is:

$$L_{adv}(G, D_I, D_V) = \mathbb{E}_{I, \tilde{I}_{t+1}}[D_V(I, \tilde{I}_{t+1}) - \mathbb{E}_{I, \tilde{I}_{t+1}}[D_V(I, \tilde{I}_{t+1})]]$$

$$+ \lambda \mathbb{E}_{\tilde{I}_{t+1}}[[\|
abla D_V(I, \tilde{I}_{t+1})\|_2 - 1]^2]$$

$$+ \mathbb{E}_{I, \tilde{I}_{t+1}}[D_I(\tilde{I}_{t+1}) - \mathbb{E}_{I, \tilde{I}_{t+1}}[D_I(\tilde{I}_{t+1})]]$$

$$+ \lambda \mathbb{E}_{\tilde{I}_{t+1}}[[\|
abla D_I(\tilde{I}_{t+1})\|_2 - 1]^2] \quad (8)$$

where $\tilde{I}_{t+1} = \epsilon I_{t+1} + (1 - \epsilon) \tilde{I}_{t+1}, \epsilon \sim U[0, 1]$. The penalty coefficient $\lambda$ is fixed as 10 in all our experiments.

In addition, we consider the pixel-wise $L_1$ loss of the prediction. Therefore the total loss $L$ is:

$$L = L_{adv} + \|I_{t+1} - \tilde{I}_{t+1}\|_1 \quad (9)$$

We trained our models on each dataset separately by minimizing the loss above using ADAM [11] algorithm with learning rate $0.0002$ and a batch size of 5.

### 3.6 Regularity Score

A regularity score based on the prediction error $\epsilon_t$ is calculated for each video frame:

$$r_{\epsilon_t} = 1 - \frac{\epsilon_t - \min_{t'} \epsilon_{t'}}{\max_{t'} \epsilon_{t'} - \min_{t} \epsilon_{t}} \quad (10)$$

In Hasan et al. [8], $\epsilon_t$ is the frame-wise reconstruction error $MSE_{\epsilon_t}$. In Liu et al. [14], $\epsilon_t$ is equivalently negative frame-wise prediction error.
Table 1: Video anomaly detection datasets details

| Dataset       | Total # frames/clips | Training # frames/clips | Testing # frames/clips | Anomaly Types                  |
|---------------|----------------------|-------------------------|------------------------|--------------------------------|
| UCSD Ped2     | 4,560/28             | 2,550/16                | 2,010/12               | biker, skater, vehicle         |
| CUHK Avenue   | 30,652/37            | 15,328/16               | 15,324/21              | running, loitering, object throwing |
| ShanghaiTech  | 315,307/437          | 274,516/330             | 40,791/107             | biker, skater, vehicle, sudden motion |

PSNR (Peak Signal to Noise Ratio): $\text{PSNR} = 10 \log_{10} \frac{\max(I_t)}{\text{MSE}_t}$. In this study, we use similar setting to the two methods above with: $e_t = \log_{10} \text{MSE}_t$.

4 EXPERIMENTS

In this section, we first introduce the three public datasets used in our experiments, which follow the same setup as other similar unsupervised video anomaly detection studies. Then, we report the video anomaly detection performance and comparison with other methods. Finally, we perform ablation studies to demonstrate the contribution of each component and interpret the results based on the proposed CT-D2GAN.

4.1 Datasets

We evaluate our framework on three widely used public video anomaly detection datasets, i.e., UCSD Ped2 dataset [13]¹, CUHK Avenue dataset [16]², and ShanghaiTech Campus (SH-Tech) dataset [18]³. We describe the dataset-specific characteristics and the effects on video anomaly detection performance, some details can be found in Table 1:

4.1.1 UCSD Ped2. UCSD Ped2 includes pedestrians, vehicles largely moving in parallel to the camera plane.

4.1.2 CUHK Avenue. CUHK Avenue includes pedestrians and objects both moving parallel to or toward/away from the camera. Slight camera motion is present in the dataset. Some of the anomalies are staged actions.

4.1.3 ShanghaiTech. Different from the other datasets, the ShanghaiTech dataset is a multi-scene dataset (13 scenes), and includes pedestrians, vehicles, and sudden motions, and the ratios of each scene in the training set and test set can be different.

4.2 Evaluation

The model was trained and evaluated on a system with an NVIDIA GeForce 1080 Ti GPU and implemented with PyTorch. To measure the effectiveness of our proposed CT-D2GAN framework for video anomaly detection, we report the area under the receiver operating characteristic (ROC) curve i.e., AUC. Specifically, AUC is calculated by comparing the frame-level regularity scores with frame-level ground truth labels.

Table 2: Frame-level video anomaly detection performance (AUC).

| Method            | UCSD Ped2 | CUHK   | SH-Tech |
|-------------------|-----------|--------|---------|
| MPPCA+SF [19]     | 61.3      | -      | -       |
| MDT [13, 19]      | 82.9      | -      | -       |
| Conv-AE [8]       | 85.0 †    | 80.6 † | 60.9 †  |
| 3D Conv [40]      | 91.2      | 80.9   | -       |
| Stacked RNN [18]  | 92.2      | 81.7   | 68.0    |
| ConvLSTM-AE [17]  | 88.1      | 77.0   | -       |
| memAE [6]         | 94.1      | 83.3   | 71.2    |
| memNormality [22] | 97.0      | 88.5   | 70.5    |
| ClusterAE [3]     | 96.5      | 86.0   | 73.3    |
| AbnormalGAN [25]  | 93.5      | -      | -       |
| Frame prediction [14] | 95.4   | 85.1   | 72.8    |
| Pred+Recon [31]   | 96.3      | 85.1   | 73.0    |
| CT-D2GAN          | 97.2      | 85.9   | 77.7    |

† Evaluated in [14]; -: Not evaluated in the study.

Ordered in publication year. The best performance in each dataset is highlighted in **boldface**.

4.3 Video Anomaly Detection

To demonstrate the effectiveness of our proposed CT-D2GAN framework for video anomaly detection, we compare it against 12 different baseline methods. Among those, MPPCA (mixture of probabilistic principal component analyzers) + SF (social force) [19], MDT (mixture of dynamic textures) [13, 19] are handcrafted feature based methods; Conv-AE [8], 3D Conv [40], Stacked RNN [18], and ConvLSTM-AE [17] are encoder-decoder based approaches; MemAE [6], MemNormality [22] and ClusterAE [3] are recent encoder-decoder based methods enhanced with memory module or clustering; AbnormalGAN [25], Frame prediction [14], and Pred+Recon [31] are methods based on adversarial training.

Table 2 shows the frame-level video anomaly detection performance (AUC) of various approaches. We observed that encoder-decoder based approaches in general outperform handcrafted feature based methods. This is because the handcrafted features are usually extracted based upon a different objective and thus can be sub-optimal. Within encoder-decoder based approaches, ConvLSTM-AE outperforms Conv-AE since it can better capture temporal information. We also notice that adversarial training based methods perform better than most baseline methods. Finally, our
Figure 2: Examples of video anomaly detection. The blue lines in the line graphs delineate frame-level regularity scores. The green and red shaded segments in the line graphs indicate the ground truth normal and abnormal video segments respectively. The frames in the green boxes are regular frames from the regular video segments, the frames in the red boxes are abnormal frames from abnormal video segments. The abnormal objects are annotated.

4.4 Ablation Studies

To understand how each component contributes to the anomaly detection task, we conducted ablation studies with different settings: (1) convolutional transformer only without the adversarial training (Conv Transformer), (2) Conv Transformer with image discriminator only, (3) Conv Transformer with video discriminator only, (4) U-Net based generator (as has been utilized in image-to-image translation [10] and video anomaly detection [14]) with dual discriminator, and compare with our full CT-D2GAN model. The performance comparison can be found in Table 3. We observed that adversarial training can enhance the performance for anomaly detection, either with the image discriminator or the video discriminator. Video discriminator alone achieves nearly similar performance as using dual discriminator, but we observed the loss decreased faster when combined with image discriminator. Using image discriminator alone was not as effective, and the loss was less stable. Finally, we observed that CT-D2GAN achieved superior performance than U-Net with dual discriminator, suggesting that convolutional transformer can better capture the spatio-temporal dynamics and thus can make a more accurate detection.

4.5 Interpretation

We illustrate an example of predicted future frame $t+1$ and compare it with the previous frame $t$ and the ground truth future frame $t+1$ in Figure 3. The prediction performance is poor for the anomaly (red box). And also we noted that the model is able to capture the temporal dynamics by predicting the future behavior in normal part of the image (green box).

Self-attention weights under perturbation. It is not straightforward to directly interpret the temporal self-attention weight vector, as temporal self-attention is applied to an abstract representation of the video. Therefore, to further investigate the effectiveness of temporal self-attention, we perturb two frames of the video and run the inference on this perturbed video segment. For one frame (Figure 4, red), we added a random Gaussian noise with zero mean.
We plot the temporal attention weights for the frame right after the input noise. A module can adaptively select relevant feature maps and is robust to the contribution to the attended map. This suggests that self-attention was employed to perform future frame prediction. A dual discriminator based adversarial training approach was used to maintain the local consistency of the predicted frame and the global coherence conditioned on the previous frames. Thorough experiments on three widely used video anomaly detection datasets demonstrate that our proposed CT-D2GAN is able to detect anomaly frames with superior performance.

5 CONCLUSIONS
In this paper, we developed Convolutional Transformer based Dual Discriminator Generative Adversarial Networks (CT-D2GAN) to perform unsupervised video anomaly detection. The convolutional transformer which consists of three components, i.e., a convolutional encoder to capture the spatial patterns of the input video clip, a temporal self-attention module to encode the temporal dynamics, and a convolutional decoder to integrate spatio-temporal features, was employed to perform future frame prediction. A dual discriminator based adversarial training approach was used to maintain the local consistency of the predicted frame and the global coherence conditioned on the previous frames. Thorough experiments on three widely used video anomaly detection datasets demonstrate that our proposed CT-D2GAN is able to detect anomaly frames with superior performance.

REFERENCES
[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In International Conference on Machine Learning (ICML). PMLR, 214–223.
[2] Thomas Brox, Andreas Bruhn, Nils Papenberg, and Joachim Weickert. 2004. High accuracy optical flow estimation based on a theory for warping. In European Conference on Computer Vision (ECCV). Springer, 25–36.
[3] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. 2020. Clustering Driven Deep Autoencoder for Video Anomaly Detection. In European Conference on Computer Vision (ECCV). Springer, 329–345.
[4] Yong Shean Chong and Yong Haur Tay. 2017. Abnormal event detection in videos using spatiotemporal autoencoder. In International Symposium on Neural Networks (ISNN). Springer, 189–196.
[5] Fei Dong, Yu Zhang, and XiuShan Nie. 2020. Dual Discriminator Generative Adversarial Network for Video Anomaly Detection. IEEE Access 8 (2020), 88176–88176.
[6] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetla Venkatesh, and Anton van den Hengel. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In IEEE International Conference on Computer Vision (ICCV). IEEE, 1705–1714.
[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Couvillie. 2017. Improved training of Wasserstein GANs. In Advances in Neural Information Processing Systems (NIPS). 5767–5777.
[8] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 733–742.
[9] Eddy Ilg, Nikolaus Mayer, Tommoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2462–2470.
[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 5967–5976.
[11] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) (2015).
[12] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. In Advances in Neural Information Processing Systems (NIPS). 5767–5777.
[13] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 733–742.
[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 3431–3440.
[15] Wenda Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 FPS in MATLAB. In IEEE International Conference on Computer Vision (ICCV). IEEE, 2720–2727.
[16] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. Remembering history with conditional adversarial networks. In IEEE International Conference on Multimedia and Expo (ICME). IEEE, 439–444.

and 0.1 standard deviation to the image to simulate the deterioration in video quality; for another frame (Figure 4, purple), we scaled the optical flow maps by 0.9 to simulate the frame rate distortion. We plot the temporal attention weights for the frame right after the two perturbed frames in Figure 4. The weights assigned to the perturbed frames are clearly lower than the others, implying less contribution to the attended map. This suggests that self-attention module can adaptively select relevant feature maps and is robust to input noise.
