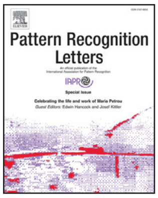
Pattern Recognition Letters 155 (2022) 143–150

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Video anomaly detection using deep residual-spatiotemporal translation network

Thittaporn Ganokratanaa[a], Supavadee Aramvith[b,*], Nicu Sebe[c]

[a] Department of Electrical Engineering, Chulalongkorn University, Bangkok, 10330, Thailand
[b] Multimedia Data Analytics and Processing Research Unit, Department of Electrical Engineering, Chulalongkorn University, Bangkok 10330, Thailand
[c] Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy

ARTICLE INFO

Article history:
Received 23 September 2020
Revised 12 October 2021
Accepted 1 November 2021
Available online 3 November 2021

Edited by: Jiwen Lu

MSC:
41A05
41A10
65D05
65D17

Keywords:
Anomaly detection
Generative adversarial network
Surveillance video
Residual unit
Hard negative mining

ABSTRACT

Video anomaly detection has gained significant attention in the current intelligent surveillance systems. We propose Deep Residual Spatiotemporal Translation Network (DR-STN), a novel unsupervised Deep Residual conditional Generative Adversarial Network (DR-cGAN) model with an Online Hard Negative Mining (OHNM) approach. The proposed DR-cGAN provides a wider network to learn a mapping from spatial to temporal representations and enhance the perceptual quality of synthesized images from a generator. During DR-cGAN training, we take only the frames of normal events to produce their corresponding dense optical flow. At testing time, we compute the reconstruction error in local pixels between the synthesized and the real dense optical flow and then apply OHNM to remove false-positive detection results. Finally, a semantic region merging is introduced to integrate the intensities of all the individual abnormal objects into a full output frame. The proposed DR-STN has been extensively evaluated on publicly available benchmarks, including UCSD, UMN, and CUHK Avenue, demonstrating superior results over other state-of-the-art methods both in frame-level and pixel-level evaluations. The average Area Under the Curve (AUC) value of the frame-level evaluation for the three benchmarks is 96.73%. The improvement ratio of AUC in the frame level between DR-STN and state-of-the-art methods is 7.6%.

1. Introduction

An anomaly is a rare event occurring in crowded scenes and there might be more than one anomaly at a time. Generally, for multiple shot video, video anomaly detection (VAD) utilizes a temporal video segmentation algorithm to detect shot boundaries in consecutive video frames [1]. The challenges of VAD relate to complex and crowded scenes, the anomaly localization, small anomaly datasets, and many false-positive detection results. The anomaly localization is required to indicate the position of the abnormalities in a scene and is more challenging than detecting an abnormal frame. Another challenge is the very small number of anomalies present in the available public datasets leading to the difficulty to learn a good classifier. Besides, these challenges result in false-positives in the final output through which the system incorrectly detects normal events as abnormal ones.

In complex real crowded scenes with different occluded and small objects, deep learning methods [6,7,22,34] are more suitable than the previous works using hand-crafted features (e.g., Gaussian regression with Bags of Visual Words [3], trajectories with *K*-means [5], and Histogram of Oriented Gradients [33]) as they are able to generalize the representations of these objects due to the nonlinear transformation performance of learnable models. In addition, many of the deep learning methods [7,25,26,28] are only able to obtain a high detection rate on the frame level while the detection rate at the pixel level is much lower. The reasons are as follows: i) a full frame is fed into the model without prior knowledge on the objects, resulting in insufficient features of objects of interest for performing deep data-hungry learning; ii) patch extraction is not effective in collecting comprehensive features of the object. Recent works [24,29] aim to enhance the accuracy using supervised learning methods that need data labeling for all samples, making it not suitable for VAD as anomalies are varied and unpredictable. Hence, unsupervised deep learning methods are a more suitable solution as they aim to learn only normal events (the majority of patterns in the scene) without the need of labeling data.

[*] Corresponding author.
E-mail address: supavadee.a@chula.ac.th (S. Aramvith).

https://doi.org/10.1016/j.patrec.2021.11.001

Any unknown patterns will be considered as anomalies by their large distance from the normal patterns. Following this consideration, Generative Adversarial Networks (GANs) have gained more attention in anomaly detection research due to their outstanding performance in constructing images, affording data augmentation, and dealing with implicit data in complex scenarios [8]. GANs consist of two competing networks: a generator *G* and a discriminator *D*. With the convolutional networks in *G*, many works have tried to achieve a high visual quality of image reconstruction and to overcome vanishing gradients. U-Net has been proposed in [27] based on the idea of skip connections [9] to enhance the accuracy of image segmentation for biomedical image. Isola et al. proposed [11] an effective translation of sketch images to realistic images based on conditional GANs (cGANs) with the use of U-Net.

In this work, we propose a novel Deep Residual Spatiotemporal Translation Network (DR-STN) approach for video anomaly detection and localization in crowds. Inspired by He et al. [9], Isola et al. [11], we propose a novel Deep Residual cGAN (DR-cGAN) to enhance the accuracy and quality of the synthesized image. Different from previous works [7,15,25,26] which are based on [11], our DR-cGAN is built by designing the residual units and the residual connections in *G* to learn the translation of objects of interest (the foreground object in the scene) from appearance (spatial) to motion (temporal) representations. The object can be a walking crowd, vehicle, wheelchair, paper, bag, etc. Our goal is to learn only normal events which refer to walking crowds in the benchmark datasets [17–19]. Specifically, we do not find the difference between the normal events, instead, we aim to learn more of various normal events from both spatial and temporal representations to be able to differentiate them from the unknown events during testing. This means that our DR-cGAN can learn all normal objects at once during training. Fig. 1 shows the overview of our proposed framework during testing in which a powerful object detector [2] is initially applied to extract the objects in the frame to be fed into our DR-cGAN. The reconstruction error is computed by finding a pixel-by-pixel difference between the generated and the real temporal frames, representing the possible abnormal events. Online Hard Negative Mining (OHNM) and semantic region merging methods are then implemented to obtain only the true positive anomaly detection results for the final output.

Our contribution can be concluded as four-fold: (i) our unsupervised DR-STN learns only normal events without using any handcrafted features and effectively translates comprehensive information of the objects of interest from appearance to motion representations in crowded scenes; (ii) we propose DR-cGAN, a novel end-to-end unsupervised deep residual connection network, to improve perceptual information of reconstructed images from the generator. DR-cGAN provides a wider network that extensively passes information from the previous to the next layer of encoder and decoder. To the best of our knowledge, this is the first attempt to build deep residual connections (projection and identity shortcuts) on the U-Net architecture of cGAN for VAD; (iii) we introduce the object detector as the pre-processing process to extract only the objects of interest to feed into the DR-cGAN model to help in learning the pattern of normal objects. This provides better object localization for the pixel level; (iv) we introduce OHNM and a semantic region merging as the post-processing processes to eliminate the false-positives without retraining the model and integrate the intensity of objects for the final anomaly output, providing more reliable and remarkable results than the state-of-the-art works.

2. Related works

Among existing works, the deep learning approaches are the most successful ones. The main approaches include supervised and unsupervised learning.

The supervised learning methods typically provide higher accuracy on classification problems. Ramachandra et al. [24] proposed anomaly localization in videos using Siamese CNN to compute a distance between the ground truth label on normal and abnormal video patches, causing over-fitting issues as the input of the network is limited to small patches of the abnormal event. Singh et al. [29] proposed Aggregation of Ensembles (AOE) of different fine-tuned CNNs with additional multiple SVM and Softmax classifiers to detect anomalies in crowds. This network is not end-to-end trainable and has a high cost of data annotation for obtaining a sufficient amount of data.

On the other hand, unsupervised learning is considered as being a more flexible approach for VAD. Xu et al. [32] proposed appearance and motion anomaly detection network using Stacked Denoising AutoEncoders (SDAEs) as the feature extractor with the One-Class SVM classifier. Prawiro et al. [22] proposed a two-stream autoencoder where the decoder is used to learn the static background and the dynamic foreground objects. Ravanbakhsh et al. [25] proposed two cross-channel networks between appearance and motion and vice versa based on cGANs. This fusion strategy for the two networks makes it more complex to reconstruct images. Similarly, the adversarial discriminator based on cGANs is proposed in [26], where the discriminator is used as the classifier during testing, making it faster than [25] but yielding lower accuracy. Tang et al. [30] proposed the combination of future frame prediction and reconstruction error method using two U-net blocks in the generator for detecting anomalies. The network is trained in an adversarial manner along with the use of gradient, intensity, and temporal image difference constraints, obtaining in better results than the baseline method proposed by Liu et al. [15]. Ganokratanaa et al. [7] proposed a deep spatiotemporal translation network (DSTN) based on GAN with pre- and post-processing pro-

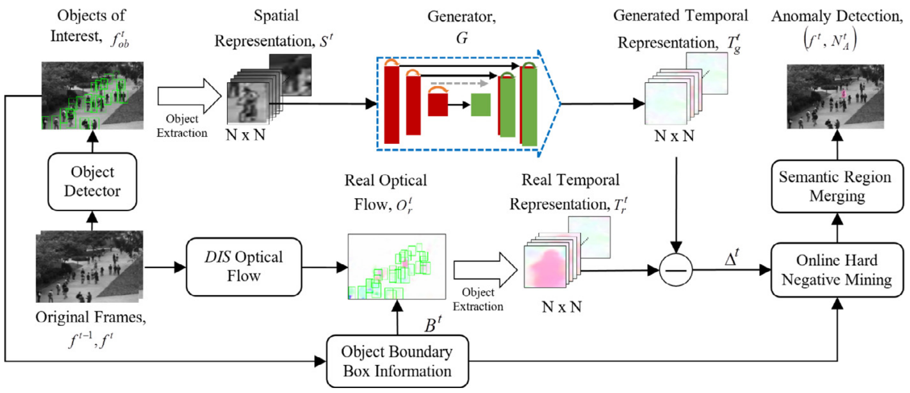

Objects of Interest, f_{ob}^t
Spatial Representation, S^t
Generator, G
Generated Temporal Representation, T_g^t
Anomaly Detection, (f^t, N_A^t)
Object Extraction
N x N
N x N
Object Detector
Real Optical Flow, O_r^t
Real Temporal Representation, T_r^t
Semantic Region Merging
Δ^t
Original Frames, f^{t-1}, f^t
DIS Optical Flow
Object Extraction
N x N
Online Hard Negative Mining
B^t
Object Boundary Box Information

Fig. 1. Overview of proposed framework.

cedures, resulting in a good frame-level anomaly detection. However, their background removal is quite sensitive to shadow and illumination changes and the patch extraction is not always able to obtain the full object appearance.

The proposed DR-cGAN is different from other previous works since we do not rely on hand-crafted features or require any labeled data as in the supervised-based approaches. Specifically, we are different from Ganokratanaa et al. [7] as we build the deep residual cGAN architecture with the object detector without any pre-defined background subtraction model. Additionally, the OHNM method [12] has been implemented to explicitly address anomaly localization and false-positive detection problems, providing more robust and reliable results.

3. Methodology

The proposed DR-STN consists of four main components as described next.

3.1. Pre-processing DR-STN

The object detection is introduced at the first stage of DR-STN to detect, locate, and extract the objects of interest for the input of our DR-cGAN model, allowing us to gain more meaningful semantic information. We use You Only Look Once (YOLO) [2], which is trained on the Microsoft COCO dataset [14] for object detection, to handle the challenges from the realistic scenes (e.g., noise, illumination changes, and object scaling and occlusions) due to its high robustness on images in different environments and its optimal speed-accuracy tradeoff. The pre-trained YOLO is applied on each frame f to predict a set of bounding boxes for the objects. These bounding boxes aim to extract spatial information of the objects from each frame f and temporal information of the objects from each dense optical flow O_r to pass into the DR-cGAN for model learning.

3.2. DR-cGAN in DR-STN

Our DR-cGAN is proposed for learning the translation from spatial to temporal information (dense optical flow). In training, we input only the objects of interest in the frames of normal events to G. G translates the spatial object f_{ob} to the synthesized dense optical flow object O_{obg} in such a way that it is challenging for D to differentiate it from the real dense optical flow object O_{obr}. Our G and D architectures are adopted from Ioffe and Szegedy [10], Radford et al. [23]. The residual units in G are designed based on [9]. The details of our architecture are explained in the following subsections.

3.2.1. Generator with residual connections

The generator G is the core model used both in training and in testing in DR-cGAN. In the common GAN [8], G learns a random noise z as an input to construct an output image $\hat{y}$. Differently, cGAN [20] learns a conversion from an image x with a random noise z to output an image $\hat{y}$, $\hat{y} = G(x, z)$. However, the use of random noise z is not essential in G as G can still learn the mapping without the noise [11]. Following [11], we apply the noise in the form of dropout in the decoder, resulting in $\hat{y} = G(x)$.

A concerning issue of translating the spatial to temporal information is mapping the difference in surface appearance from a high-resolution input to a high-resolution output grid. Thus, we design the generator architecture to effectively align the input structure to the output structure as shown in Fig. 2. This generator network consists of two models: encoder and decoder. The encoder functions as the data compressor, while the decoder reversely functions as the data decompressor. In the encoder, the spatial image is input to a series of down-sampling layers until reaching a bottleneck layer. Then, the decoder performs the reconstruction process to generate a semantic output image. Our structure of the encoding and decoding blocks are defined in [7].

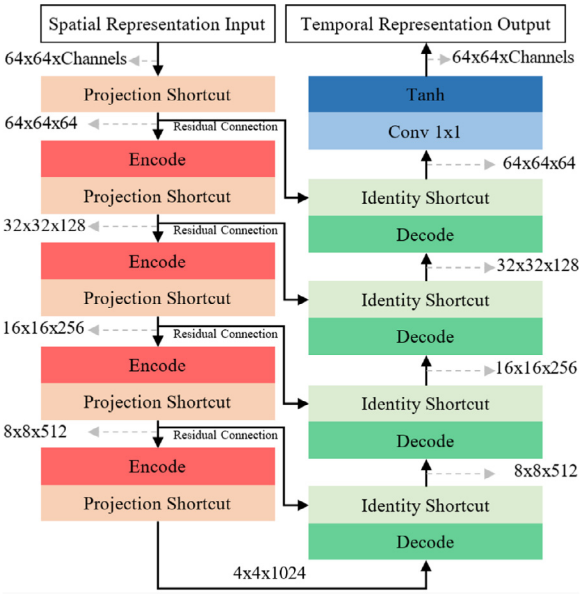


Fig. 2. The proposed generator architecture of DR-cGAN.

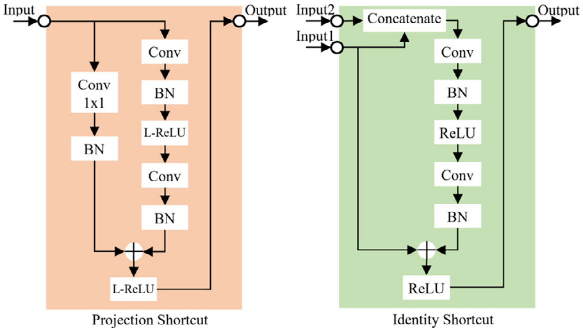


Fig. 3. Structure of the residual unit.

To achieve finer semantic results, the low-level information is required to be shared between the input and the output in order to propagate the information through the network without degradation while maintaining the high-level information. Following this consideration, we introduce the novel generator architecture as: i) we add the residual unit in each layer of the encoder and decoder to achieve a wider feature learning network; and ii) we apply the residual connections from encoder layers to decoder layers to share the low-level information. Suppose n is the total number of layers. The residual unit is added after each encoder layer i and decoder layer $n - i$, while the residual connections are added from each encoder layer i to the decoder layer $n - i$. This implies better generalization and easier optimization for image translation as discussed in Section 4.6. Specifically, our residual units consist of projection and identity shortcuts as shown in Fig. 3. The projection shortcut is used to match the dimensions. Since the dimensions of our input and output in the encoder are not the same, we define the projection shortcut to increase the dimensions of the input features to be able to add with the output features. For the decoder, its residual unit has two inputs: the output from the decoder layer and the residual connections from the encoder layer. The identity shortcut is then defined to add the concatenated inputs with the output using the same dimensions.

3.2.2. Discriminator

We use the discriminator D only during the training process. D classifies two classes of spatiotemporal objects: a real class $\{x = f_{ob}, y = O_{obr}\}$ and a fake class $\{x = f_{ob}, O_{obg} = G(x)\}$. We train D to maximize the correct classification problem on both real and fake classes. A binary cross-entropy loss with logits loss is computed as the objective function of D. In contrast, G is trained to minimize the objective function of D with a reconstruction error between O_{obg} and O_{obr}. In other words, the adversarial D and G learn a two-player minimax game with value function $V(D, G)$:

$$\min_{G}\max_{D} V(D, G) = \mathcal{L}_{cGAN}(G, D) + \lambda\mathcal{L}_{L1}(G) \tag{1}$$

where $\mathcal{L}_{cGAN}(G, D)$ presents as a cGAN loss, and $\mathcal{L}_{L1}(G)$ is a reconstruction loss in G. Both losses are determined as below,

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}\Big[\big\|y - G(x)\big\|_1\Big], \tag{2}$$

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) = {} & \mathbb{E}_{x,y}\Big[\log\big[\sigma(D(x, y))\big]\Big] \\ & + \mathbb{E}_{x}\Big[\log\big[1 - \sigma(D(x, G(x)))\big]\Big]\end{aligned} \tag{3}$$

where σ is a sigmoid function, $\sigma(D) = 1/(1 + e^{-D})$.

Our DR-cGAN provides good feature learning of the learned normal events which is less complex than learning anomalies. Since we do not train with the abnormal event, the model understands only the normal patterns at the training time and then can observe the irregular objects following the reconstruction error at the testing time. The anomaly detection process is explained in detail in the following section.

3.3. Anomaly detection

At testing time, only G is applied to translate f_{ob} of test video frame to O_{obg} in order to compare with its corresponding O_{obr} for obtaining the irregular object. Specifically, the spatial objects $S^t = \{f_{ob_1}, f_{ob_2}, \ldots, f_{ob_K}\}_t$ and their corresponding bounding boxes $B^t = \{b_1, b_2, \ldots, b_K\}_t$ are extracted from each frame at time t, where K is the total number of the detected objects in a frame. To detect the irregular object, the reconstruction error $\Delta^t = \{\Delta_{ob_1}, \Delta_{ob_2}, \ldots, \Delta_{ob_K}\}_t$ is computed by differentiating between the real temporal objects $T_r^t = \{O_{obr_1}, O_{obr_2}, \ldots, O_{obr_K}\}_t$ and the synthesized temporal objects generated from G, $T_g^t = \{O_{obg_1}, O_{obg_2}, \ldots, O_{obg_K}\}_t$. The reconstruction error on kth object is:

$$\Delta_{ob_k} = O_{obr_k} - O_{obg_k} > 0 \tag{4}$$

Δ_{ob_k} provides an irregular score representing the possible anomalous event in the scene when the value of Δ_{ob_k} is greater than 0. However, the output of Δ_{ob_k} may result in a false positive, meaning that the normal object (negative sample) is incorrectly detected as the abnormal object (positive sample). This false-positive object represents a hard negative example. To ensure that we obtain the actual abnormal object, we determine the high confidence score to decide whether Δ_{ob_k} belongs to the normal or abnormal object. Then OHNM is proposed to get rid of the negative example in the anomaly detection. The probability of anomaly score P_{a_k} on kth object is computed as:

$$P_{a_k} = \sum_{(i,j)\in\Delta_{ob_k}} \Delta_{ob_k}(i, j) \Bigg/ \sum_{(i,j)\in O_{obr_k}} O_{obr_k}(i, j) \tag{5}$$

Since the model is trained only with the normal patterns, it performs a good reconstruction on the normal objects, causing a low value of Δ_{ob_k} and P_{a_k}. In contrast, the model is not able to correctly reconstruct the abnormal object, causing high value of Δ_{ob_k} and P_{a_k}. Following these characteristics, the high confidence scores of the normal and abnormal objects are set based on two-interval thresholds: confident normal threshold C_n and confident abnormal threshold C_a. After this setting, we obtain a true detection of normal and abnormal objects. However, there are some objects which are not enrolled in these two criteria ($C_n < P_{a_k} < C_a$). Then, we take these objects into consideration of the OHNM examples to finalize the true detection of anomaly outputs.

To observe hard negative examples, the template matching is performed as a short tracklet to match each detected object in f^t to the search patch p in its adjacent frames within a window ±1 frame. The size of p is assigned to extensively cover the displacement of the object by enlarging the bounding box b_k of the kth reference object f_{ob_k} to the size of 20 × 20 pixels. This size of p is defined due to the small movement of the object between frames. Specifically, the main idea of our OHNM is to move f_{ob_k} (template) at f^t over p in its adjacent frames (f^{t-1} and f^{t+1}) in order to measure the highest similarity patch and record the template as a normal object. The highest similarity of the pattern between f_{ob_k} and p is determined via block matching by shifting f_{ob_k} with the distance (u, v) in the horizontal and vertical directions within the corresponding sub-patch of p. To find the similarity score from the best-matching position between f_{ob_k} and p, we use the standard normalized cross-correlation (NCC) algorithm which is formulated as:

$$NCC(u, v) = \frac{\sum\limits_{(i,j)\in f_{ob_k}} p(u + i, v + j) \cdot f_{ob_k}(i, j)}{\sqrt{\sum\limits_{(i,j)\in f_{ob_k}} p^2(u + i, v + j) \cdot \sum\limits_{(i,j)\in f_{ob_k}} f_{ob_k}^2(i, j)}} \tag{6}$$

After acquiring the NCC similarity score, we are able to determine whether the object is abnormal or not based on the confident similarity score C_s. If there is a large appearance change between frames, we assign the object as being abnormal otherwise, we consider it to be a normal object or an isolated object yielded by flicker noise. Finally, the semantic region merging is implemented by combining all the detected abnormal objects into a full semantic frame A computed as follows,

$$A(i, j) = \begin{cases} \Delta_{ob_k}(i, j), & \text{non-overlapping object} \\ 1/K \sum\limits_{k\in K} \Delta_{ob_k}(i, j), & \text{otherwise} \end{cases} \tag{7}$$

where K is the total number of the final abnormal objects and (i, j) are the pixel positions of A.

A is normalized to get the probability score N_A in a range of [0, 1] of the full semantic frame. The highest pixel intensity value of A, M_A, is considered as the abnormal pixel in the frame. The ROC curve is performed on N_A by slightly shifting the threshold of anomaly scores in a range of [0,1] to determine the best decision threshold. N_A can be defined as follows,

$$N_A(i, j) = 1/M_A \cdot A(i, j) \tag{8}$$

4. Experimental results

In this section, we evaluate the performance of the proposed DR-STN on three anomaly benchmarks and compare it with state-of-the-art methods on both frame level and pixel level. The impact of our proposed DR-cGAN model and OHNM method along with the running time performance are analyzed in detail.

4.1. Datasets

The UCSD dataset [18] includes two sub-folders: Ped1 and Ped2. There are 34 training and 16 test videos in Ped1 and 16 training and 12 test videos for Ped2. The image sizes of Ped1 and Ped2 are 238 × 158 pixels and 360 × 240 pixels, respectively. The abnormal events in this dataset include cycling, skateboarding, vehicles, and wheelchairs.

Table 1
Anomaly Detection Benchmark Datasets.

Datasets	Number of Videos	Number of Frames	Training Samples	Testing Samples
UCSD Ped1	50	8900	5500	3400
UCSD Ped2	28	4202	2550	1652
UMN	11	7740	1200	6540
CUHK Avenue	37	30,652	15,328	15324

The UMN dataset [19] has 11 videos recorded in three different scenes (i.e., one crowded indoor and two crowded outdoor scenes) with an image size of 320 × 240 pixels. We take the first 400 frames of each scene as training samples following [4] and leave the rest for testing. The abnormal events in the UMN dataset refer to running, while the normal events refer the normal walking.

The CUHK Avenue dataset [17] has 16 training and 21 test videos with an image size of 360 × 640 pixels. There are various anomalies in the scenes, e.g., jumping, loitering, running, and throwing objects, while the normal events are the walking crowds.

We include the details of anomaly detection benchmark datasets for our experiments in Table 1, including the number of videos, frames, training, and testing samples.

4.2. Implementation details

Our proposed DR-STN is based on Python and Matlab [32] with PyTorch [21]. The training and testing processes are implemented on NVIDIA GeForce GTX 1080 Ti. Adam optimization is used to optimize our reconstruction loss ($\lambda\mathcal{L}_{L1}$) that targets to 2E-1. The optimization parameters are defined as [11].

In our DR-cGAN, the sizes of the input and output of G for both training and testing processes are set to 64 × 64 pixels. With the encoder network in G, the input image is encoded by using a CNN with a kernel size of 3 × 3 pixels and a stride $s = 2$ to reach a bridge representing the spatial data. For the decoder network in G, each layer is built as the reverse of each encoder layer. To avoid the over-fitting problems on the training dataset, the random noise z is provided in the form of dropout in the decoder with the default probability value $p = .5$. In addition, the residual units for both encoder and decoder are designed by using 3 × 3 convolution and 1 × 1 convolution with $s = 1$, respectively. For D, it takes two input images with the resolution of 64 × 64 pixels to produce the 6 × 6 output feature.

4.3. Evaluation criteria

We evaluate the quantitative performance of the proposed DR-STN considering two criteria: frame level (F) and pixel level (P). In F, the frame is considered as an anomaly if there is at least one abnormal event in a test frame. On the other hand, P specifies the location of the abnormal event. The frame is a true detection when the detected abnormal region overlaps with the ground truth region more than 40% [13].

4.4. Performance evaluation

In this section, we compare Area Under the Curve (AUC) and Equal Error Rate (EER) performance of DR-STN with other state-of-the-art methods as shown in Table. 2. We use the same network configuration and training parameter settings for all three datasets. The experiment on the UCSD dataset is implemented with 10 and 12 videos of the UCSD Ped1 and UCSD Ped2, respectively, along with their pixel-level ground truth. GANs [25] and DSTN [7] are set as the baseline methods due to their success in leveraging frame-level and pixel-level detection accuracy and achieving state-of-the-art performance under the unsupervised manner. Table 2 shows that our DR-STN surpasses not only the baseline methods but also most of the competing works in both F and P criteria in which we achieve higher AUC and lower EER than other works, except only for the AUC of the UCSD Ped1 dataset at P in [24]. This is probably due to their supervised learning on labeled abnormal data. However, our experimental results can significantly overcome other criteria in [24] and all criteria in [29] which also relies on a supervised-based method, showing the competitive performance of DR-STN in anomaly detection and localization tasks. In addition, the examples of our detection and localization results on three datasets are shown in Fig. 4 where we can detect and localize both single and multiple abnormal events in the crowded scenes even when they are occluded (e.g., a bicycle and a skateboard in Fig. 4(b)).

4.5. Computational cost analysis

It is important to compare our computational time performance with state-of-the-art works. Table 3 shows the comparison of computational time with the state-of-the-art methods in seconds per frame on the UCSD Ped1, UCSD Ped2, UMN, and CUHK Avenue datasets where the data are available from the original papers. During testing, all competing methods are evaluated by CPUs with different specifications as shown in Table 3, except Tang et al. [30] using GPU. The experimental results of the AUC and EER performance in Table 2 and the computational time in Table 3 clearly show the speed-accuracy tradeoff between DSTN [7] and the proposed DR-STN. As DR-STN has a more complex generative network, it has a higher computational time than DSTN. However, DR-STN is still faster than AMDN (Double Fusion) [32] both for UCSD Ped1 and Ped2 datasets. This is due to the fact that AMDN uses a very small image patch-based extraction (i.e., 15 × 15 pixels) while our DR-STN does not rely on image patches as we use the powerful object detector [2] to extract only objects of interest from the scene where all types of anomalies can be described by objects alone, without the interaction with other objects. Besides, it is natural for Detection at 150 fps [17] to have the lowest computational time because their sparse learning network has less neuron connections than other methods. Thus, it can be concluded that the proposed DR-STN has a good overall performance for surveillance videos, considering its high accuracy that significantly outperforms all other methods and its good running time for various benchmark datasets even if we run the testing with the CPU. Our proposed DR-STN runs at 8 frames per second (fps) using an NVIDIA 1080Ti GPU. To address the relatively high computational time of our method for real-time applications we can consider running multiple GPUs in parallel.

4.6. Analysis of DR-STN

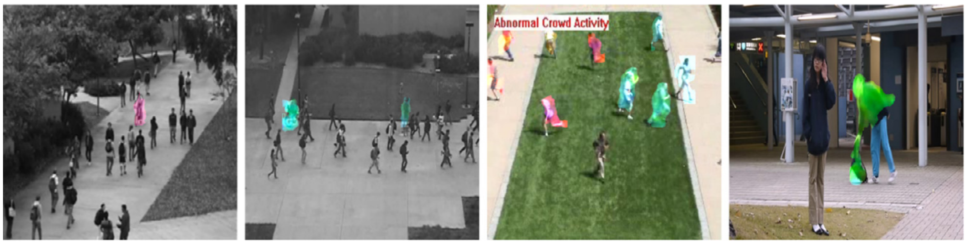
To emphasize the importance of our DR-STN, we analyze two main components of the proposed framework: i) the performance of DR-cGAN compared with the baseline methods including U-Net [11] and autoencoder which is simply built by removing the skip connections in U-Net and ii) the impact of OHNM on DR-STN with

Table 2
AUC and EER comparison with state-of-the-art methods on UCSD, CUHK Avenue, and UMN datasets.

Method	UCSD Ped1 (F) AUC/EER	UCSD Ped1 (P) AUC/EER	UCSD Ped2 (F) AUC/EER	UCSD Ped2 (P) AUC/EER	CUHK Avenue (F) AUC/EER	UMN (F) AUC/EER
Social force (SF) [19]	67.5%/31.0%	19.7%/79.0%	55.6%/42.0%	-/80.0%	-/-	96.0%/-
Sparse reconstruction [4]	46.1%/19.0%	45.3%/54.0%	-/-	-/-	-/-	97.8%/-
Detection at 150fps [17]	91.8%/15.0%	63.8%/43.0%	-/-	-/-	80.9%/-	-/-
AMDN (double fusion) [32]	92.1%/16.0%	67.2%/40.1%	90.8%/17.0%	-/-	-/-	-/-
GANs [25]	97.4%/8.0%	70.3%/35.0%	93.5%/14.0%	-/-	-/-	99.0%/-
Liu et al. [15]	83.1%/23.5%	33.4%/-	95.4%/12.0%	40.6%/-	85.1%/-	-/-
Adversarial discriminator [26]	96.8%/7.0%	70.8%/34.0%	95.5%/11.0%	-/-	-/-	99.0%/-
AnomalyNet [34]	83.5%/25.2%	45.2%/-	94.9%/10.3%	52.8%/-	86.1%/22.0%	99.6%/-
Tang et al. (optical flow) [30]	84.7%/-	-/-	96.3%/-	-/-	85.1%/-	-/-
DSTN [7]	98.5%/5.2%	77.4%/27.3%	95.5%/9.4%	83.1%/21.8%	87.9%/20.2%	99.6%/-
GMM-FCN [6]	94.9%/11.3%	71.4%/36.3%	92.2%/12.6%	78.2%/19.2%	83.4%/22.7%	-/-
Siamese [24]	86.0%/23.3%	80.4%/-	94.0%/14.1%	**93.0%/-**	-/-	-/-
AOE [29]	94.6%/-	-/-	95.9%/-	-/-	-/-	-/-
Two-stream decoder [22]	84.2%/-	-/-	96.1%/-	-/-	-/-	-/-
DR-STN (Proposed method)	**98.8%/2.9%**	**82.5%/21.5%**	**97.6%/6.9%**	86.4%/**16.3%**	**90.8%/11.0%**	**99.7%/-**

Table 3
Computational time comparison during testing (seconds per frame).

Methods	CPU	GPU	Memory	Running Time			
				Ped1	Ped2	UMN	Avenue
Sparse reconstruction [4]	2.6GHz	-	2.0GB	3.8	-	0.8	-
Detection at 150 fps [17]	3.4GHz	-	8.0GB	**0.007**	-	-	**0.007**
AMDN (Double Fusion) [32]	2.1GHz	Nvidia Quadro K4000	32GB	5.2	7.5	-	-
Tang et al. [30]	-	Nvidia Tesla P40	24GB	-	**0.03**	-	-
DSTN [7]	2.8GHz	-	24GB	0.315	0.319	**0.318**	0.334
DR-STN (Proposed Method)	3.4GHz	-	24GB	4.26	4.44	4.07	3.62



(a) bicycle (b) bicycle and skateboard (c) running in an outdoor scene (d) grabbing an object

Fig. 4. Examples of anomaly detection and localization results.

regard to AUC. First, we divide the training folder of the UCSD Ped1 dataset into two subsets: 70% for training samples and 30% for testing samples. We train DR-cGAN model and other baseline methods for 20 epochs to see their effectiveness in minimizing the $\lambda\mathcal{L}_{L1}$ loss as illustrated in Fig. 5 where our DR-cGAN (red square) reaches the lowest error over the training epochs, showing faster and superior performance in model learning than other baseline methods about 50%. To clarify the ability in generating the synthesized image on normal events during testing, we evaluate the proposed network using two common methods. First, FCN-scores for semantic segmentation on pixel accuracy [16] are computed to obtain the probability of correct pixels on a set of defined object classes (foreground and background region classes). The pixel accuracy is defined as $\sum_i n_{ii}/\sum_i n_{ti}$, where n_{ii} is the number of correct classified pixels of class i, and n_{ti} is the total number of pixels of class i. Second, Structural SIMilarity Index (SSIM) metric [31] is used to evaluate the similarity between the synthesized and the real images. For both evaluations, a higher value indicates a better result of the synthesized image. Table 4 shows that our DR-cGAN significantly surpasses all baseline methods regarding both evaluations, providing a good synthesized image quality that is highly similar to the real image.

Table 4
Performance comparison of the Autoencoder, U-Net and DR-cGAN in terms of FCN-scores on pixel accuracy and Structural SIMilarity Index (SSIM) on the UCSD Ped1 dataset.

Method	Pixel accuracy	SSIM
Autoencoder	0.81	0.78
U-Net	0.82	0.8
DR-cGAN	**0.87**	**0.85**

Apart from the above experiments, our OHNM relies on both temporal and spatial conditions. For the temporal condition, we can determine whether the object is normal or abnormal based on P_{a_k} under the criteria of two-interval thresholds, $C_n = 0.1$ and $C_a = 0.8$. The object is classified as normality if its P_{a_k} is less than or equal to 0.1 ($P_{a_k} \leq 0.1$) and as abnormality if its P_{a_k} is greater than or equal to 0.8 ($P_{a_k} \geq 0.8$). This is probably because the model has only the knowledge of the learned normal events at the train-

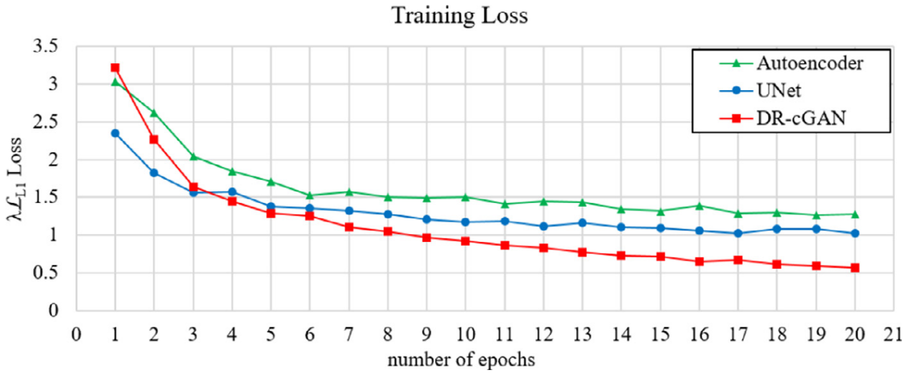


Fig. 5. Training loss comparison between Autoencoder, U-Net and DR-cGAN on the UCSD Ped1 dataset.

Table 5
AUC performance of OHNM on DR-STN.

Method	Ped1 (F)	Ped1 (P)	Ped2 (F)
DR-STN without OHNM	97.85%	72.65%	96.16%
DR-STN with OHNM	**98.83%**	**82.50%**	**97.62%**

ing time. Hence, during testing when we input all objects from each frame into the model, Δ_{ob_k} provides less difference in local pixels between the learned and the test samples in case the input is the normal object, resulting in a small value of P_{a_k} which falls into the criteria of C_n. On the other hand, there is a great difference of Δ_{ob_k} if the input is the abnormal object, resulting in a high value of P_{a_k} which is considered as abnormality following the criteria of C_a. For P_{a_k} value that does not belong to these two criteria $(0.1 < P_{a_k} < 0.8)$, we apply the template matching to observe *NCC* score of the objects between frames to indicate the appearance displacement whether the objects are the same. *NCC* results in a high similarity score if there is a small change in the appearance of the objects between frames, considering as the false-positive anomaly result. Based on the experiment, we set the confident similarity score on the normal object $C_s = 0.8$. We analyze the impact of OHNM on our DR-STN for reducing the false-positive detection results in terms of AUC on the UCSD dataset. With the use of OHNM, the model can remarkably improve the AUC values in both F and P as shown in Table 5. The AUC of P on the UCSD Ped1 dataset is increased up to about 10% compared to the plain DR-STN, providing a more precise location of the abnormal events in the scene. Following these experimental results, it is clear that applying OHNM with the proposed DR-STN benefits both anomaly detection and localization tasks.

5. Conclusion

This paper introduced a novel unsupervised deep residual spatiotemporal translation network for video anomaly detection and localization. The proposed DR-STN is embedded with DR-cGAN and OHNM, which benefits in reducing false-positive anomaly detection and increasing anomaly localization accuracy. The DR-cGAN is designed for learning the translation of objects of interest from appearance (spatial) and motion (temporal) representations integrated with the residual units, residual connections, and cGAN. Additionally, our DR-cGAN takes only raw pixels as input from the object detector, which effectively extracts each individual object of interest, without relying on any prior knowledge of hand-crafted features. We conducted extensive experiments on three available benchmarks and showed the strengths of our proposed DR-STN regarding its accuracy, robustness, and effectiveness. DR-STN significantly outperforms the state-of-the-art due to its performance in learning frame-by-frame normal events of the training dataset in various environments, occlusions, and illumination changes, making it flexible to detect and mark any unknown events that are different from the learned normal patterns as abnormal. One of the limitations of our DR-cGAN is the relatively high computational time but this can be addressed by considering a speed-accuracy tradeoff. Besides, since DR-STN is an unsupervised-based learning method, it requires many training samples of normal events and a distinct difference of the learned normal patterns from the abnormal patterns for detecting abnormalities during testing. The concern is that the model may face difficulty in distinguishing the abnormal events if their patterns are too similar to the normal patterns. However, with enough training time and resources on the normal patterns, the model is able to handle the reconstruction task effectively by showing the lowest error over the training epochs and the highest pixel accuracy over other competing methods. Our future work will focus on continuous learning of unknown events, assisting to verify if these are actually abnormal or simply rare normal events.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by Chulalongkorn University Dutsadi Phiphat Scholarship and by the PRIN project PREVUE Prot. 2017N2RK7K.

References

[1] S.H. Abdulhussain, S.A.R. Al-Haddad, M.I. Saripan, B.M. Mahmmod, A. Hussien, Fast temporal video segmentation based on Krawtchouk-Tchebichef moments, IEEE Access 8 (2020) 72347–72359.
[2] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, YOLOv4: optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934 (2020).
[3] K.-W. Cheng, Y.-T. Chen, W.-H. Fang, Video anomaly detection and localization using hierarchical feature representation and gaussian process regression, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2909–2917.
[4] Y. Cong, J. Yuan, J. Liu, Sparse reconstruction cost for abnormal event detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3449–3456.
[5] S. Cosar, G. Donatiello, V. Bogorny, C. Garate, L.O. Alvares, F. Brmond, Toward abnormal trajectory and event detection in video surveillance, IEEE Trans. Circuits Syst. Video Technol. 27 (3) (2016) 683–695.
[6] Y. Fan, G. Wen, D. Li, S. Qiu, M.D. Levine, F. Xiao, Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder, Comput. Vision Image Understanding 195 (2020) 102920.
[7] T. Ganokratanaa, S. Aramvith, N. Sebe, Unsupervised anomaly detection and localization based on deep spatiotemporal translation network, IEEE Access 8 (2020) 50312–50329.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[10] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.

[11] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[12] S. Jin, A. RoyChowdhury, H. Jiang, A. Singh, A. Prasad, D. Chakraborty, E. Learned-Miller, Unsupervised hard example mining from videos for improved object detection, in: European Conference on Computer Vision, 2018, pp. 307–324.

[13] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2013) 18–32.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, C.L. Zitnick, Microsoft COCO: common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[15] W. Liu, W. Luo, D. Lian, S. Gao, Future frame prediction for anomaly detection—a new baseline, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6536–6545.

[16] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[17] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 FPS in MATLAB, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2720–2727.

[18] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1975–1981.

[19] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 935–942.

[20] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, PyTorch: an imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8026–8037.

[22] H. Prawiro, J.-W. Peng, T.-Y. Pan, M.-C. Hu, Abnormal event detection in surveillance videos using two-stream decoder, in: IEEE International Conference on Multimedia and Expo Workshops, 2020, pp. 1–6.

[23] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: International Conference on Learning Representations, 2016.

[24] B. Ramachandra, M. Jones, R. Vatsavai, Learning a distance function with a siamese network to localize anomalies in videos, in: IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 2598–2607.

[25] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe, Abnormal event detection in videos using generative adversarial nets, in: International Conference on Image Processing, 2017, pp. 1577–1581.

[26] M. Ravanbakhsh, E. Sangineto, M. Nabi, N. Sebe, Training adversarial discriminators for cross-channel abnormal event detection in crowds, in: IEEE Winter Conference on Applications of Computer Vision, 2019, pp. 1896–1904.

[27] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2015, pp. 234–241.

[28] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, R. Klette, Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes, Comput. Vision Image Understanding 172 (2018) 88–97.

[29] K. Singh, S. Rajora, D.K. Vishwakarma, G. Tripathi, S. Kumar, G.S. Walia, Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets, Neurocomputing 371 (2020) 188–198.

[30] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, J. Yang, Integrating prediction and reconstruction for anomaly detection, Pattern Recognit. Lett. 129 (2020) 123–130.

[31] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[32] D. Xu, Y. Yan, E. Ricci, N. Sebe, Detecting anomalous events in videos by learning deep representations of appearance and motion, Comput. Vision Image Understanding 156 (2017) 117–127.

[33] Y. Yuan, Y. Feng, X. Lu, Statistical hypothesis detector for abnormal event detection in crowded scenes, IEEE Trans. Cybern. 47 (11) (2016) 3597–3608.

[34] J.T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, R.S.M. Goh, AnomalyNet: an anomaly detection network for video surveillance, IEEE Trans. Inf. Forensics Secur. 14 (10) (2019) 2537–2550.