Robust Anomaly Detection in Videos Using Multilevel Representations

Hung Vu,[†] Tu Dinh Nguyen,[‡] Trung Le,[‡] Wei Luo,[†] Dinh Phung[‡]
[†]Center for Pattern Recognition and Data Analytics, Deakin University, Geelong, Australia
{hungv, wei.luo}@deakin.edu.au
[‡]Monash University Clayton, VIC 3800, Australia
{Tu.Dinh.Nguyen, trunglm, Dinh.Phung}@monash.edu

Abstract

Detecting anomalies in surveillance videos has long been an important but unsolved problem. In particular, many existing solutions are overly sensitive to (often ephemeral) visual artifacts in the raw video data, resulting in false positives and fragmented detection regions. To overcome such sensitivity and to capture true anomalies with semantic significance, one natural idea is to seek validation from abstract representations of the videos. This paper introduces a framework of robust anomaly detection using multilevel representations of both intensity and motion data. The framework consists of three main components: 1) representation learning using Denoising Autoencoders, 2) level-wise representation generation using Conditional Generative Adversarial Networks, and 3) consolidating anomalous regions detected at each representation level. Our proposed multilevel detector shows a significant improvement in pixel-level Equal Error Rate, namely 11.35%, 12.32% and 4.31% improvement in UCSD Ped 1, UCSD Ped 2 and Avenue datasets respectively. In addition, the model allowed us to detect mislabeled anomalies in the UCDS Ped 1.

1 Introduction

With the increasing popularity of surveillance cameras in public places, there have been growing interests in systems that automatically detect anomalous events in videos. Such systems are crucial in domains like traffic monitoring and security control, where the sheer volume of video data makes manual video analysis infeasible. Video anomaly detectors follow two main approaches based on the availability of data labels. Supervised methods (Singh and Mohan 2017; Zhang et al. 2016; Li, Mahadevan, and Vasconcelos 2014) usually learn the regular objects from the set of normal examples. Often the labels for supervised learning are costly to obtain and are prone to human errors. Without label information, unsupervised methods assume that abnormality has a lower probability of occurrence than regular objects in reality (Sodemann, Ross, and Borghetti 2012) and identify any event that rarely occurs in videos to be an anomaly. Since unsupervised learning can utilize huge unlabeled video archives on a daily basis, this is the approach we take to design our system.

Both supervised and unsupervised methods usually work on low-level features of pixel/edge/motion information to detect abnormality (Ribeiro, Lazzaretti, and Lopes 2017; Hasan et al. 2016; Chong and Tay 2017). For example, a query frame is compared with its reconstructed frame that is produced by deep Convolutional Autoencoders (CAEs) (Ribeiro, Lazzaretti, and Lopes 2017; Hasan et al. 2016) or CAEs/Long Short Term Memories (Chong and Tay 2017; Luo, Liu, and Gao 2017). Regions with high reconstruction errors (over a given threshold) indicate the presence of anomaly objects. Madhdyar Ravanbakhsh et al. (Ravanbakhsh et al. 2017a; 2017b) apply the idea of adversarial learning in Generative Aversarial Networks (GANs) to localize anomaly behaviors. Two conditional GANs are trained on the data pair of frames and their optical flow features to produce generated frames/optical flow maps given the other data. An irregular object is usually associated with a high generation error in (Ravanbakhsh et al. 2017a) or a low value of GAN's discriminator (Ravanbakhsh et al. 2017b). Again, the inputs of these GAN-based systems are raw video frames.

Detecting abnormality using low-level features encounters two issues: i) low-level detection usually causes fragmented and interrupted regions because an anomalous object may contains very normal pixels, for example, a white car also has pixels similar to a white footpath, and ii) low-level information is sensitive to noise and is significantly affected by environment changes and thus low-level detectors usually make a lot of false detections. These problems indicate the unreliability and ineffectiveness of low-level detectors. To address the first issue, we propose to detect anomaly objects at abstract representation. Such abstractness can be discovered via multilayer architecture of deep networks (Zeiler and Fergus 2014), where low layers encode visual features such as edges, corners and colors whilst higher layers describe semantic identities such as objects and their positional relationships. Object-level detection allows us to obtain complete anomaly objects without fragments or interruption. To tackle the second problem of false detections, we are based on the idea that true anomaly objects should be highlighted at many levels of detections and therefore we combine low-level detection with abstract-level detection to accurately isolate abnormality. This combination has three benefits: a) increase the reliability of detection; b) reduce false detections at low levels and also abstract levels; and c) more detected objects found by abstract-level detectors but not by low-level ones. In particular, we firstly use Denoising Autoencoders (DAEs) to extract the high representations of low-level data, i.e., pixel intensity and optical flow features, and then follow the Conditional Adversarial Generative Networks (cGAN) approach in (Ravanbakhsh et al. 2017a; 2017b) to detect anomaly objects at each representation level. Detected objects that show a strong agreement between detection results at all levels are considered as anomalies.

It is worthy to note that there are several studies (Tran and Hogg 2017; Xu et al. 2017; Sabokrou et al. 2017) that can detect anomalies at deep representation. More specifically, these systems divide frames into local patches, feed these patches into a deep network to obtain higher representation and then identify anomaly patches using these representations. Our work stands out from such systems in two aspects: i) our *multilevel* anomaly detector is completely different from the existing *single* level detectors, which work on either low-level data or high-level representations but not both; and ii) our detection is done on the representation of the *whole frame*, which completely preserves objects and their interactions in the scene, whilst image partition corrupts and disconnects objects and therefore, semantic information is lost dramatically in aforementioned patch setting.

To summarize, the main contributions of this work are three-fold:

- A multilevel detection framework is introduced to detect anomaly objects in a video sequence at different levels of semantic representations and consolidate these layer-wise detections for more reliable and accurate results.
- Thorough experiments and analysis show that our multilevel detectors significantly outperform other state-of-the-art anomaly detectors (11.35%, 12.32% and 4.31% improvement in pixel-level Equal Error Rate) in three benchmarks of UCSD Ped 1, UCSD Ped 2 and Avenue datasets. A new record is obtained by our system in UCSD Ped 2.
- We also discover annotation mistakes of missing abnormality (video 32) and mislabeled partially occluded/distant objects in the UCSD Ped 1 dataset. We correct these errors and introduce a new annotation for the widely-used benchmark dataset UCSD Ped 1.

2 Preliminaries

We firstly introduce the basic concepts of Denoising Autoencoders and Conditional Adversarial Generative Networks that are the fundamental building-blocks of our system.

2.1 Denoising Autoencoders

Denoising Autoencoder (DAE) (Vincent et al. 2008) is a neural network that is trained to reconstruct a data sample $\boldsymbol{v} \in \mathcal{D}$ from its corrupted version $\tilde{\boldsymbol{v}} \sim q_{\text{noise}}(\tilde{\boldsymbol{v}}|\boldsymbol{v})$, where q_{noise} can be any noise distribution, e.g., a Gaussian or uniform distribution. The network is divided into two parts: an encoder and a decoder. The encoder $f_\theta(\tilde{\boldsymbol{v}})$, parameterized by θ, takes the input $\boldsymbol{v}$ and maps it to a code $\boldsymbol{h}$ in the hidden space. The decoding function $g_\phi(\boldsymbol{h})$ projects the code back to the input space. In DAE, f_θ and g_ϕ are usually constructed as deep convolutional networks of weight and bias parameters $\theta = \left\{\boldsymbol{W}_{\text{e}}^{(l)}, \boldsymbol{b}_{\text{e}}^{(l)}\right\}_{l=1}^{N_{\text{e}}}$ and $\phi = \left\{\boldsymbol{W}_{\text{d}}^{(l)}, \boldsymbol{b}_{\text{d}}^{(l)}\right\}_{l=1}^{N_{\text{d}}}$, where N_{e} and N_{d} are the numbers of encoding and decoding hidden layers respectively. At training time, the network learns to reconstruct the original data point $\boldsymbol{v}$. In particular, θ and ϕ are updated via the gradient descent procedure to minimize the following objective:

$$\begin{aligned}\min_{\theta,\phi} \mathcal{J}_{\text{DAE}} &= \min_{\theta,\phi} \frac{1}{|\mathcal{D}|} \sum \left\|\boldsymbol{v}_i - g_\phi\left(f_\theta\left(\tilde{\boldsymbol{v}}_i\right)\right)\right\|_2^2 \\ &+\gamma\left(\sum_{l=1}^{N_{\text{e}}} \left\|\boldsymbol{W}_{\text{e}}^{(l)}\right\|_2^2 + \sum_{l=1}^{N_{\text{e}}} \left\|\boldsymbol{W}_{\text{d}}^{(l)}\right\|_2^2\right) \quad (1)\end{aligned}$$

wherein the second term sets the penalty for the weights' sparsity and γ is a regularization hyper-parameter. Training AEs on perturbed data $\tilde{\boldsymbol{v}}$ not only prevents the model from learning an identity function - a trivial solution of AEs, but also allows to obtain better representations, which are more robust to noise in images (Vincent et al. 2008).

2.2 Conditional Generative Adversarial Networks

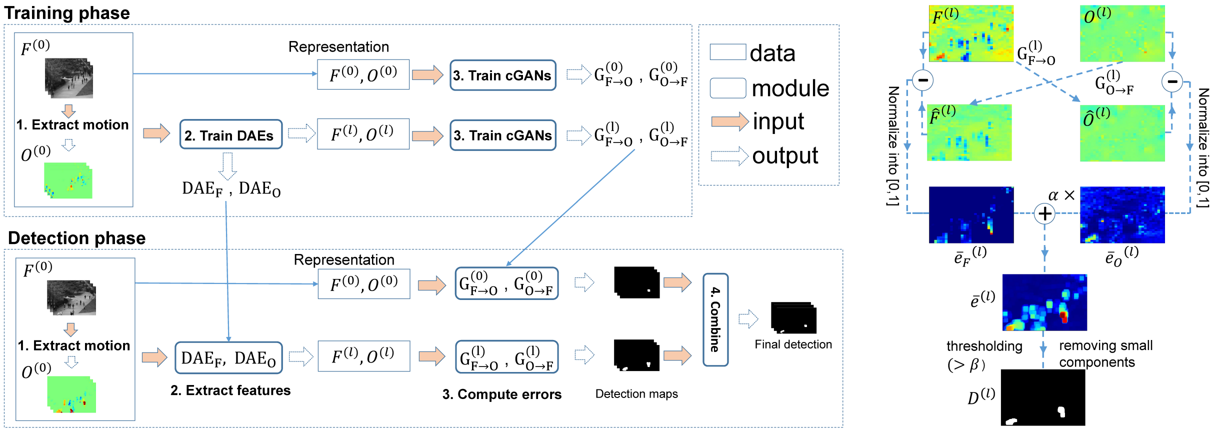
Similar to (Ravanbakhsh et al. 2017a), our anomaly detection system is also based on the generation errors of different image feature types and thus we are interested in Conditional Generative Adversarial Networks (cGAN) (Isola et al. 2017) to learn the transformation between two image representations, e.g., frames to optical flow images and vice versa. More specifically, cGAN learns a generative model G that outputs an image $\mathrm{G}(\boldsymbol{x}, \boldsymbol{z})$ (e.g., an optical flow image) from a source image $\boldsymbol{x}$ (e.g., a pixel intensity image) and a random vector $\boldsymbol{z}$. Using the adversarial learning mechanism, the generator aims at generating realistic images, which look like target images $\boldsymbol{y}$ and cannot be distinguished by a deterministic neural network, named discriminator $\mathrm{D} : \{\boldsymbol{x}, \boldsymbol{o}\} \to [0, 1]$, where $D(\boldsymbol{x}, \boldsymbol{o})$ is indicates how correct $\boldsymbol{o}$ is a transformed image of $\boldsymbol{x}$. By contrast, D is optimized to discriminate the "fake" pair of images generated by G and the real pair from the data. This training phase is summarized through the following objective function:

$$\begin{aligned}\mathcal{J}_{\text{cGAN}} &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\left[\log\left(1 - \mathrm{D}\left(\boldsymbol{x}, \mathrm{G}\left(\boldsymbol{x}, \boldsymbol{z}\right)\right)\right)\right] + \\ &\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\log \mathrm{D}\left(\boldsymbol{x}, \boldsymbol{y}\right)\right] + \lambda \mathcal{J}_{L_1}\left(\boldsymbol{x}, \boldsymbol{y}\right) \quad (2)\end{aligned}$$

wherein the additional L_1 loss $\mathcal{J}_{L_1}(\boldsymbol{x}, \boldsymbol{y}) = \left\|\boldsymbol{y} - \mathrm{G}(\boldsymbol{x}, \boldsymbol{z})\right\|_1$ forces G to generate images as close to the target images as possible and the hyper-parameter λ balances the losses. At training time, G tries to minimize Eq. 2 whilst D learns to maximize this equation. We update the discriminator with one gradient step and then the generator with one step in each training epoch (Goodfellow et al. 2014). We refer readers to (Isola et al. 2017) for more details of the networks and their training protocol.

3 Multilevel Anomaly Detection

In this section, we describe our proposal of MultiLevel Anomaly Detector (MLAD) in detail. Since representation



(a) (b)

Figure 1: Our proposed MLAD system: a) framework overview and b) single level detection.

learning enables to express the scene at various levels of abstractness, detecting unusual objects at these levels can bring the benefit of discovering different aspects of abnormality and then improve the detection performance. Our system can be split into the training and detection phases. In the training phase, we i) compute the optical flow image for every frame; ii) train one separate DAE on each type of frame data and optical flow data; iii) feed each data type into the corresponding trained DAE to extract its high-level features; and iv) train a pair of cGANs on these high-level representations of the frame and motion data.

To detect anomalies in testing videos, the system performs the following steps: i) extracting optical flow features of testing frames; ii) obtaining the high-level representations of frames and motion features using the DAEs; iii) applying trained cGANs to compute generation error maps at one level; iv) thresholding these maps to obtain binary detection maps at each level; and finally v) combining these detection maps into a final detection result. The overview of both training and detection processes is illustrated in Fig. 1a.

3.1 Learning Multilevel Representations

The training step starts with extracting motion features and learning abstract representations of all low-level data. Given a training video collection $\mathcal{D}_F = \{F_i\}_{i=1}^{N_f}$ of N_f frames, we firstly resize all frames into 256 × 256 images and scale their values into $[-1, 1]$. The method in (Brox et al. 2004) is adopted to compute the motion map O_i for every two consecutive frames F_i and F_{i+1}. Each motion map is a 3-channel image of optical flow description along the x-axis, the y-axis and their magnitude. Next, we train two Denoising Autoencoders DAE_F and DAE_O separately on the frame data $\mathcal{D}_F$ and the motion data $\mathcal{D}_O = \{O_i\}$ by minimizing Eq. 1. These networks have the same number of layers for both encoder and decoder. For the encoding path, we use convolutional layers with the stride of 2 and the kernel size of 5 × 5, followed by batch normalization layers and leaky ReLU activation functions. In the decoder, we use the similar architecture but replace convolutional layers with deconvolutional ones. All networks are trained using Adagrad optimizer (Duchi, Hazan, and Singer 2011), $\gamma = 1$, a learning rate of 0.1 and 500 epochs.

When DAE_F has been learned from the training data, we feed each frame F_i into the network and achieve the activations at each encoding layer. Since leaky ReLU activations are unbounded, we normalize them to zero-mean and unit-variance and then clip them to $[-1, 1]$ to obtain $F_i^{(l)}$ as the l^{th} level of abstract representation for the frame data. We apply the same procedure to extract the abstract representation $O_i^{(l)}$ for the motion data. Finally, both $\mathcal{D}_F^{(l)} = \left\{F_i^{(l)}\right\}$ and $\mathcal{D}_O^{(l)} = \left\{O_i^{(l)}\right\}$ are coupled as data to train cGANs for the l^{th} level in the next step. It is noteworthy that although $1 \le l \le N_{\mathrm{e}}$, we assume that $l = 0$ goes back to the low-level data, where $F_i^{(0)} = F_i$ and $O_i^{(0)} = O_i$. Thus, our notations of $F_i^{(l)}$ and $O_i^{(l)}$ imply $0 \le l \le N_{\mathrm{e}}$ in all remaining sections.

3.2 Training Detectors

For each level of representations, we train two generative networks: $\mathrm{G}^{(l)}_{F\to O}$ tries to generate the motion representation $O_i^{(l)}$from a frame representation $F_i^{(l)}$ and the other $\mathrm{G}^{(l)}_{O\to F}$ outputs $F_i^{(l)}$ from $O_i^{(l)}$. To that end, $\mathrm{G}^{(l)}_{F\to O}$ is jointly trained with a discriminator on the input $\mathcal{D}_F^{(l)}$ and the label data $\mathcal{D}_O^{(l)}$ in the adversarial learning as described in Sec. 2.2. For $\mathrm{G}^{(l)}_{O\to F}$, the input and label data are $\mathcal{D}_O^{(l)}$ and $\mathcal{D}_F^{(l)}$ respectively. We follow the network architecture and the setting in (Isola et al. 2017) to train these models using a learning rate of 0.0002, $\lambda = 100$ and the batch size of 1. At the end of training, all discriminators are discarded and we come up with N_{e} pairs of generators $\left(\mathrm{G}^{(l)}_{F\to O}, \mathrm{G}^{(l)}_{O\to F}\right)$ at all abstract levels and two generators $\left(\mathrm{G}^{(0)}_{F\to O}, \mathrm{G}^{(0)}_{O\to F}\right)$ at the low level, all of which are used to detect irregularities in testing videos.

Algorithm 1 Combining multilevel detection maps

```
Require: Detection maps $\left\{D^{(l)}\right\}$, score maps $\left\{E^{(l)}\right\}$, object
    lists $\left\{C^{(l)}\right\}$, anomaly threshold $\beta$ and overlapping threshold
    $\rho$
Ensure: Final detection $D$, $E$ and $C$
 1: $D \leftarrow D^{(0)}$;   $E \leftarrow E^{(0)}$;   $C \leftarrow C^{(0)}$
 2: for $l \leftarrow 1, \ldots, N_\mathrm{e}$ do
 3:     for $c \in C$ and $c_l \in C^{(l)}$ do
 4:         if $L(c \cap c_l) / L(c) \geq \rho$ then
 5:             $D(c) \leftarrow D(c) \cup D^{(l)}(c_l)$
 6:             $E(c_l \cup c) \leftarrow \max\left(E(c_l \cup c), E^{(l)}(c_l \cup c)\right)$
 7:             $C(c) \leftarrow C(c) \cup C^{(l)}(c_l)$
 8: $E \leftarrow \min(E, 2\beta)$
 9: $E \leftarrow \frac{E - \min(E)}{\max(E) - \min(E)}$
```

3.3 Anomaly Detection

Single level detection At testing time, MLAD inputs a sequence of frames F_i and computes the corresponding motion maps O_i, analogously in the training phase. Then, the high-level features $F_i^{(l)}$ and $O_i^{(l)}$ are extracted by passing F_i and O_i into DAE_F and DAE_O respectively. For every representation level, we run two trained cGANs on these high-level features to gain the generated motion image $\hat{O}_i^{(l)} = \mathrm{G}_{F\to O}^{(l)}\left(F_i^{(l)}, \boldsymbol{z}\right)$ and generated frame $\hat{F}_i^{(l)} = \mathrm{G}_{O\to F}^{(l)}\left(O_i^{(l)}, \boldsymbol{z}\right)$. Since we assume that abnormality usually occurs in regions with motion (static objects exist in every frame and then they are regular objects), we set the values of $F_i^{(l)}, O_i^{(l)}, \hat{F}_i^{(l)}$ and $\hat{O}_i^{(l)}$ to 0 at zero-optical flow locations. This assumption also helps to limit the search space and speed up the detection.

Next, we compute generation error maps as the difference between the original features and the generated ones, $e_{F,i}^{(l)} = F_i^{(l)} - \hat{F}_i^{(l)}$ and $e_{O,i}^{(l)} = O_i^{(l)} - \hat{O}_i^{(l)}$. To consolidate these error maps, we firstly normalize them into $[0, 1]$ for each channel as $\bar{e}_{F,i}^{(l)} = \left[e_{F,i,j}^{(l)} / m_{F,j}\right]_{j=1}^{N_F^{(l)}}$ and $\bar{e}_{O,i}^{(l)} = \left[e_{O,i,j}^{(l)} / m_{O,j}\right]_{j=1}^{N_O^{(l)}}$, wherein $N_F^{(l)}$ and $N_O^{(l)}$ are the number of the channels of the error maps whilst $m_{F,j} = \max_{i,x,y} e_{F,i,j}^{(l)}(x, y)$ and $m_{O,j} = \max_{i,x,y} e_{O,i,j}^{(l)}(x, y)$ are the maximum errors across all locations in the video for the j^{th} channel. The channel-wise normalization is crucial to reduce the negative effect of false detections in one channel. The combined error map at one level is obtained by summing these normalized maps $\bar{e}_i^{(l)} = \bar{e}_{F,i}^{(l)} + \alpha \bar{e}_{O,i}^{(l)}$, where α is a coefficient to control the contribution of each feature type. We set $\alpha = 2$, analogously to (Ravanbakhsh et al. 2017a) in all experiments. The error maps of the whole video $E^{(l)} = \left\{\bar{e}_i^{(l)}\right\}$ is smoothen by taking the average of consecutive frames in a sliding frame window of 5. By comparing $E^{(l)}$ with a predefined anomaly threshold β, we obtain a binary detection map as $D_i^{(l)}(x, y) = 1$ for abnormal pixels if $\bar{e}_i^{(l)}(x, y) > \beta$, otherwise $D_i^{(l)}(x, y) = 0$, where (x, y) is a pixel in the i^{th} frame.

We adopt the connected-component-finding procedure in (Vu et al. 2017) to filter out false detections and noise. Specifically, we construct a sparse graph of vertices at $D_i^{(l)}(x, y) = 1$ and edges that connect two vertices (i, x, y) and $(i + t, x + u, y + v)$, satisfying $t, u, v \in (-1, 0, 1)$ and $|t| + |u| + |v| > 0$. By finding all connected components $C^{(l)}$ in this graph and discarding one $c \in C^{(l)}$ whose lifespan $L(c)$ (the number of frames in which c occurs) is less than 30 contiguous frames, we obtain the refined detection map $D^{(l)}$ and the corresponding object list $C^{(l)}$, where an anomalous object is one connected component. Finally, we apply dilation operations to the refined map $D^{(l)}$ to fill noisy holes inside detected regions. Fig. 1b summarizes all aforementioned steps to compute $D^{(l)}$ at one level.

Multilevel detection combination Since detections at all levels provide different views of anomaly objects, the result at one level can support and correct wrong detections at the other levels and therefore combining these results can improve the performance. Alg. 1 describes our proposal to merge anomaly objects over levels. Starting with the object list at the pixel-level $C = C^{(0)}$, we travel across all higher levels and merge the abstract object list $C^{(l)}$ into the current list C. In particular, given two objects $c \in C$ and $c^{(l)} \in C^{(l)}$, if their intersection $c \cap c^{(l)}$ is large enough (greater than an overlapping threshold ρ) in terms of the lifespan ratio, we update c and its corresponding score map E with anomaly pixels in $c^{(l)}$(lines $5 - 7$ in Alg. 1). Finally, E is normalized into $[0,\ 1]$ by clipping any values that are greater than 2β and shifting and scaling E by its minimum and maximum. We use 2β to balance the value ranges of regular and irregular objects.

4 Experiments

In this section, we show that our proposal of multilevel detection can improve the performance of localizing anomalies in a video sequence.

4.1 Experimental settings

We compare our system with the state-of-the-art methods on three datasets of UCSD Ped 1 (Li, Mahadevan, and Vasconcelos 2014), USCD Ped 2 (Li, Mahadevan, and Vasconcelos 2014) and Avenue (Lu, Shi, and Jia 2013). Each dataset consists of two sets of training videos and testing videos We resize all videos into the same size of 256×256 pixels. Since our method is completely unsupervised, we discard all label information during training. We set the thresholds $\beta = 0.8$ and $\rho = 0.75$ in all experiments. These thresholds give the best performance in all datasets.

To evaluate an anomaly detection system, we are based on the criteria of *frame-level, pixel-level* in (Li, Mahadevan, and Vasconcelos 2014) and *dual-pixel level* (Sabokrou et al. 2015). These metrics compute the pair of the true-positive

	UCSD Ped 1		UCSD Ped 1*		UCSD Ped 2		Avenue	
	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓
MLAD$_0$	66.07	**22.38**	64.41	22.32	92.96	5.47	47.07	43.90
MLAD$_{0+\text{Alex}}$	63.48	24.35	61.89	24.24	94.33	**4.43**	40.60	46.33
MLAD$_{0+3}$	**66.60**	22.65	**66.95**	**21.08**	**94.45**	4.58	**52.82**	**38.82**

Table 1: Pixel-level evaluation (%) on three datasets and new ground-truth of UCSD Ped 1 (denoted by *). The best/second best values are bold/underlined. ↑/↓ means higher/lower is better.

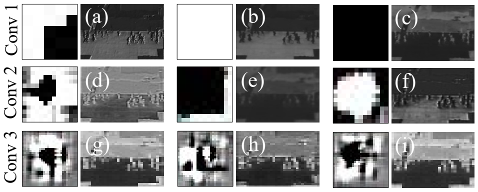


Figure 2: The pairs of filters and their activations learned by our DAE$_F$. The white values in activation images indicate where the left filter is activated most.

rate (TPR) and the false-positive rate (FPR) that are TPR $= \frac{\text{\# of true-positive frames}}{\text{\# of positive frames}}$ and FPR $= \frac{\text{\# of false-positive frames}}{\text{\# of negative frames}}$.

At the frame-level evaluation, we do not care about finding anomaly objects in the scene and therefore one frame is true-positive if it contains abnormality in its ground-truth and any pixel is detected as an anomaly. Meanwhile, the locations of detected objects are important in the pixel-level evaluation; in particular, a frame is true-positive if at-least 40% of abnormal ground-truth pixels are detected by the system. However, the drawback of the pixel-level is that when 40% of the abnormal ground-truth are overlapped, all false detected regions are ignored. As a result, the system can give as many (false) detections as possible to have more chance to cover the ground-truth. To address this problem, the dual-pixel level metric adds one constraint to the pixel-level conditions. In particular, at least ξ (i.e., 0.05 or 5%) of detected regions are true abnormal pixels. Therefore, if a large irrelevant region is detected, it is not considered as a true positive by this metric.

For each criterion, a Receiver Operating Characteristic (ROC) curve is drawn using pairs of TPR and FPR. Finally, Area Under Curve (AUC) and Equal Error Rate (EER), that is the ratio of misclassification at FPR $= 1 -$ TPR, are usually reported to compare methods.

4.2 Abstract feature representations

In our first experiment, we evaluate the effectiveness of DAEs to represent the scene at different levels of abstractness. Two DAE networks of 3 hidden layers are trained separately on the pixel-intensity data and the optical flow data. The numbers of the filters of convolutional layers are $32, 16$ and 8 respectively and the stride is set to 2. We compare three MLAD versions of using: a) low-level data only (MLAD$_0$) including raw frames and optical flow images; b) low-level data combined with the top abstract representations extracted by DAEs (MLAD$_{0+3}$) and c) combined with Conv5 of AlexNet (Krizhevsky, Sutskever, and Hinton 2012) (MLAD$_{0+\text{Alex}}$). Table 1 shows that MLAD$_{0+3}$ can improve the detection performance but MLAD$_{0+\text{Alex}}$ cannot. This is because AlexNet is trained on ImageNet (Russakovsky et al. 2015) for image classification problem, where the objects are highly-distinctive at the global scale, whereas abnormal objects in surveillance videos can appear in local and small regions of the scene. As a result, scene-driven training for abstract feature extraction is better than using a general network like AlexNet.

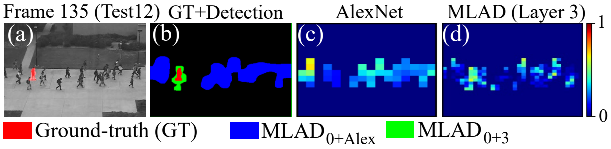


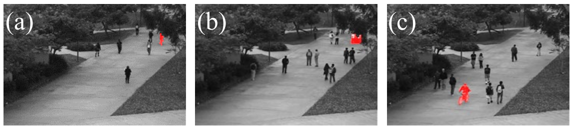
Figure 3: An example of detecting anomaly objects at AlexNet-Conv5 and our MLAD's layer 3 on UCSD Ped 2: a) original frame and ground-truth; b) ground-truth and detection results on AlexNet-Conv5 and our DAE$_O$'s layer 3; c-d) optical flow error maps produced by detectors on AlexNet-Conv5 and our DAE$_O$'s layer 3. Best view in color.

To further understand the trained DAEs, we have a look at their filters and activations (Fig. 2). The filters at the low layer (the top row) usually describe colors and edges, for example, the filter in Fig. 2a describes the pattern of two black and white regions and it responses significantly at the boundaries of the footpath and pedestrians. Conv 2 and 3 encode highly abstract levels of objects and their relationships in the scene. Since these filters are trained on UCSD Ped 2, where pedestrians move on a white footpath, the dark blobs in the filters Fig. 2d-f represent pedestrians and therefore Conv 2 layer focuses on objects. Similarly, multiple dark blobs in Conv 3's filters (Fig. 2g-i) describe many objects and their relative locations. It is noteworthy that our networks tend to learn region/pattern-like filters rather than corner/edge filters as shown in (Erhan et al. 2009) since the surveillance scene contains more low-frequency features such as homogeneous regions than high-frequency details in close-up object images of ImageNet. It also explains why Conv5 features of AlexNet are not as effective as our learned features in this anomaly detection problem.

Fig. 3 is the anomaly frame that fails to be detected by AlexNet-Conv5. The frame contains an abnormal skater, which is moving among a lot of pedestrians. Due to the complexity of the crowded footpath, AlexNet-based detector does not understand the scene comprehensively and therefore, it cannot distinguish between the true abnormality and most of the pedestrians (many wrong detections in Fig. 3b and c). Conversely, MLAD$_{0+3}$ does the task very well and isolates the skater correctly (Fig. 3b and d).

4.3 New UCSD Ped 1 ground-truth

The experimental results in Table 1 show that MLAD$_{0+3}$ has better performance than MLAD$_0$ in both UCSD Ped 2 and Avenue but there is marginal improvement in UCSD Ped 1. This motivates us to investigate the results in this dataset deeply. Surprisingly, we discover that MLAD$_{0+3}$ can



Test 028 - Frame 109 Test 036 - Frame 019 Test 032 - Frame 100

Figure 4: Some incorrect labels in UCSD Ped 1: a-b) mislabeled anomaly objects because of partial occlusion and c) a complete abnormal biker missing from the ground-truth. Best view in color.

Layers		3	4	5	0+3	0+4	0+5	0+all	0+≥3
Ped 1*	(A)	57.01	_	_	***68.36***	_	_	63.14	_
	(B)	42.60	_	_	62.20	_	_	57.76	_
	(C)	36.87	53.98	_	63.23	*64.61*	_	56.89	64.13
	(D)	38.14	46.88	33.49	64.24	63.46	*64.67*	62.60	63.47
	MLAD_0 : 64.41								
Ped 2	(A)	59.83	_	_	*94.45*	_	_	92.51	_
	(B)	47.04	_	_	*93.06*	_	_	92.78	_
	(C)	63.25	66.20	_	*96.12*	*93.69*	_	*96.87*	*96.98*
	(D)	58.68	78.65	64.73	*97.22*	*96.67*	*97.36*	*97.61*	***98.28***
	MLAD_0 : 92.96								
Avenue	(A)	*52.33*	_	_	***52.82***	_	_	*48.66*	_
	(B)	*51.82*	_	_	*49.98*	_	_	*49.69*	_
	(C)	*50.04*	*47.19*	_	*50.31*	*48.43*	_	*50.93*	*51.59*
	(D)	46.35	*52.45*	*52.43*	*50.1*	*50.21*	*48.74*	*51.36*	*51.82*
	MLAD_0 : 47.07								

(A) $32/16/8$ (B) $32/64/128$ (C) $32/64/128/256$
(D) $64/128/256/512/1024$

Table 2: AUCs (%) at the pixel-level criterion are reported on different networks and abstract layers. Ped1* is the relabeled UCSD Ped 1. The bold/underlined numbers are the best/second best in each dataset. Italic values indicate improvement compared with MLAD_0.

find a lot of correct anomaly frames, which were unfortunately labeled as normal frames in the ground-truth. This problem not only degrades its results but also counts the false detections of MLAD_0 as true-positives and therefore, the power of multilevel detector is not evaluated properly in UCSD Ped 1. We review the whole dataset and correct its two main mistakes of mislabeling partially occluded objects (e.g., Fig. 4a and b) and missing some anomaly objects (e.g., Fig. 4c). Overall, 745 frames of 27 videos (out of 36 videos in UCSD Ped 1) are re-annotated. The performance in this new ground-truth (Ped 1*) is reported in Table 1, where MLAD_{0+3} shows the improvements of 2.54% in AUC at the pixel-level against MLAD_0. This result confirms the benefit of multilevel detection and the effectiveness of our proposed framework. It is worthy to note that since the labeling problem does not take place in Ped 2, where all occluded objects are correctly labeled, we do not re-annotate this dataset. Our new ground-truth is available online [1].

4.4 Abstract detector and combined detector

We test our framework on four different network structures of (A) $32/16/8$, (B) $32/64/128$, (C) $32/64/128/256$ and (D) $64/128/256/512/1024$, where (A) $32/16/8$ indicates a 3-layer network (A) of 32, 16 and 8 filters in its convolutional layers. For each network configuration, we evaluate the contribution of each abstract layer and their combination with low-level data to the overall detection performance.

Table 2 shows that a single abstract-level detector does not always outperform a low-level detector MLAD_0. This is because it also has its own false detections. For UCSD Ped 1 and 2, the detector at an abstract level can be fooled by the diversity of human poses or different configurations of the group of pedestrians (e.g., Fig. 5a). By contrast, for the Avenue dataset with large-size objects, MLAD_0 makes more mistakes (e.g., false and fragmented detections as shown in Fig. 5c) and thus the abstract-level detector is better. However, since an abnormal object, if exists, should be present in detection results at many levels and therefore, combining both low and abstract-level detectors can help to highlight this object more correctly and eliminate detection mistakes at each level (Fig. 5a). As a result, the combined detector boosts performance dramatically as shown in Table 2.

We also investigate different strategies to combine detectors: MLAD_0 cooperates with (I) *one* abstract-level detector, (II) *all* abstract-level detectors and (III) detectors at *the highest layers* (≥ 3). For (III), we choose all layers from the 3$^{\text{rd}}$ layer because it is the lowest layer, where MLAD_{0+3} shows improvement in most of the datasets. We observe that (II) and (III) do not increase the accuracy much but add more computational cost to the entire system. For this reason, we conclude that the strategy (I) is the best choice to balance the accuracy and the speed.

For the network size, better detection usually comes with larger networks. However, deeper layers do not always mean better performance. This is because the depth of a layer is related to object sizes in video frames. More specifically, one unit at a high layer expresses the combination of visual elements (edges, parts, objects) of the previous layer and then this implies that a deeper layer works on larger image regions, whose areas are specified by the size of previous layers' filters. As a result, abstract-level detectors work less effectively in far-view scenes with small objects such as UCSD Ped 1 and 2 than near-view scenes such as Avenue as shown in Table 2 (some samples of these datasets can be seen in Fig. 5). From this table, we choose the configuration MLAD_{0+3} using the network (A) for our next experiments because of its acceptable improvement in most cases.

4.5 Video anomaly detection

We compare our proposed framework with existing systems that are based on conventional machine learning methods and other state-of-the-art deep detectors. From the results in Table 3, we can observe that our method MLAD_{0+3} (A) outperforms all methods in UCSD Ped 2 at least 2.02% in AUC and 4.22% in EER. It also achieves excellent results with the highest dual-pixel value of 51.76% and has at least 9.76% and 4.31% improvement in AUC and EER at the pixel-level evaluation in the Avenue dataset. For UCSD Ped 1, although the performance of our system is lower at the frame-level evaluation, MLAD_{0+3} (A) is still better in dual-pixel level criterion and has much lower EER and comparable AUC in pixel-level evaluation. It is noteworthy that since the pixel-level and dual-pixel level criteria consider object locations, they can evaluate the detection task more precisely than the

[1] https://github.com/SeaOtter/vad_gan

	Ped 1					Ped 2					Avenue				
	Frame		Pixel		Dual	Frame		Pixel		Dual	Frame		Pixel		Dual
	AUC↑	EER↓	AUC↑	EER↓	AUC↑	AUC↑	EER↓	AUC↑	EER↓	AUC↑	AUC↑	EER↓	AUC↑	EER↓	AUC↑
Machine learning methods															
OC-SVM[1]	59.06	42.97	21.78	37.47	11.72	61.01	44.43	26.27	26.47	19.23	**71.66**	33.87	33.16	47.55	33.15
GMM[1]	60.33	38.88	36.64	35.07	13.60	75.20	30.95	51.93	18.46	40.33	67.27	35.84	43.06	43.13	41.64
MDT[2]	81.8	25.0	44.0	55.0	_	85.0	25.0	_	55.0	_	_	_	_	_	_
Deep models															
ConvAE[3]	81.00	27.90	_	_	_	90.00	21.70	_	_	_	70.20	**25.10**	_	_	_
WTA+SVM[1x1][4]	81.3	27.9	56	46.8	_	96.6	8.9	89.3	16.9	_	_	_	_	_	_
AMDN[5]	92.1	16.0	67.2	40.1	_	_	_	90.8	17.0	_	_	_	_	_	_
DeepGMM[6]	92.5	15.1	69.9	64.9	_	_	_	_	_	_	_	_	_	_	_
Plug-and-Play CNN[7]	95.7	8.0	64.5	40.8	_	88.4	18.0	_	_	_	_	_	_	_	_
GAN/gen[8]	**97.40**	8.0	70.30	35.00	_	93.50	14.00	_	_	_	_	_	_	_	_
GAN/dis[9]	96.80	**7.0**	**70.80**	34.00	_	95.50	11.00	_	_	_	_	_	_	_	_
MLAD$_{0+3}$(A)	82.34	23.50	66.60	**22.65**	**60.79**	97.52	4.68	94.45	4.58	93.99	71.54	36.38	**52.82**	**38.82**	**51.76**
MLAD (best for each dataset)	82.34	23.50	66.60	**22.65**	**60.79**	**99.21**	**2.49**	**97.22**	**1.74**	**96.75**	71.54	36.38	**52.82**	**38.82**	**51.76**
	MLAD$_{0+3}$(A)					MLAD$_{0+3}$(D)					MLAD$_{0+3}$(A)				

[1](Vu et al. 2017), [2](Mahadevan et al. 2010), [3](Hasan et al. 2016), [4](Tran and Hogg 2017), [5](Xu et al. 2015), [6](Feng, Yuan, and Lu 2017), [7](Ravanbakhsh et al. 2018), [8](Ravanbakhsh et al. 2017a), [9](Ravanbakhsh et al. 2017b)

Table 3: Anomaly detection results at the frame-level, pixel-level and dual pixel-level ($\alpha = 5\%$) criteria. Higher AUC and lower EER indicate better performance. Meanwhile, high dual-pixel values point out more accurate localization. We do not report EER for dual-pixel level because this number does not always exist. The best scores are in bold whilst the next best is underlined. A cell with "_" indicates "value not reported".

Methods	AUC	EER	Methods	AUC	EER
WTA+SVM	25.3	18.9	Plug-and-Play CNN	31.2	32.8
AMDN	24.9	24.1	GAN/gen	27.1	27.0
DeepGMM	22.6	49.8	GAN/dis	26.0	27.0
Minimum	22.6	18.9	MLAD$_{0+3}$(A)	15.74	-0.85

Table 4: The average AUC and EER gaps of existing deep models and our MLAD$_{0+3}$(A) in the UCSD Ped 1 dataset, where AUC *gap* = AUC (frame-level) - AUC (pixel-level) and EER *gap* = EER (pixel-level)-EER (frame-level).

frame-level evaluation. For this reason, although existing models have higher frame-level values than our system, they are only slightly higher (about 4.2% in AUC) at the other criteria in UCSD Ped 1. This reveals that these models are finding many wrong anomalous pixels in video frames.

This observation is also confirmed by considering the *gaps* between the frame-level and pixel-level criteria in UCSD Ped 1 (Table 4). The minimum AUC and EER *gaps* of deep methods are 22.6% and 18.9%, whereas these *gaps* are smaller for MLAD$_{0+3}$(A), 15.74% and -0.85% respectively. More interestingly, the negative value of our EER gap indicates that our pixel-level EER is even better than the frame-level EER. Furthermore, there is slight difference between AUC of MLAD at the pixel-level criterion and its stricter version of the dual-pixel level criterion, showing that MLAD is focusing on object localization intensively and thus it can highlight most anomalous objects correctly with low false detections. Fig. 5 visualizes some cases detected by MLAD$_{0+3}$(A): a) filtering out false detections at the abstract level; b) an anomaly object missed by the low-level detector and c) fragmented and false detections at the low level. Finally, we report the best performance and the optimal network configuration for each dataset in Table 3 (the last row). Overall, excellent performance in video anomaly detection task proves that our proposed idea of multilevel detection is useful to localize abnormality in surveillance videos.

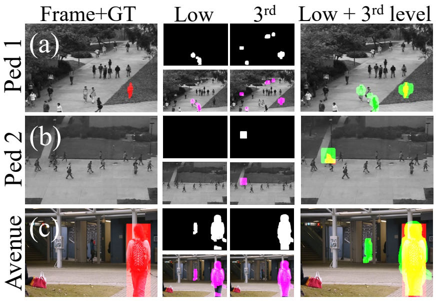


Figure 5: Some examples of detecting anomaly objects in all datasets: (from left to right) original frames and ground-truth (GT), detection results at low-level representations MLAD$_0$, at abstract-level representations MLAD$_3$ (A) and combined detections MLAD$_{0+3}$ (A). Legends for colors: red is ground-truth (GT), magenta is a single level detection, green is the combined detection and yellow is the intersection of red and green regions. Best view in color.

5 Conclusion

We propose a multilevel video anomaly detector. By finding unusual objects at high-level representations besides at low level data and combining these detection results, our detector can localize anomaly regions with high accuracy and low false detections. Experiments on public datasets show that our proposed method outperforms single level detectors and other existing state-of-the-art systems on both the UCSD Ped 2 and Avenue datasets, and is competitive on UCSD Ped 1 dataset. The ground-truth of UCSD Ped 1 is also corrected and published to the research community.

6 Acknowledgement

This work was partially supported by the Defence Science and Technology Group (DST) and the Australian Research Council (ARC) DP150100031.

References

Brox, T.; Bruhn, A.; Papenberg, N.; and Weickert, J. 2004. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, volume 3024 of *Lecture Notes in Computer Science*, 25–36. Springer.

Chong, Y. S., and Tay, Y. H. 2017. Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder. In *Advances in Neural Networks - ISNN 2017: 14th International Symposium, Part II*, 189–196.

Duchi, J. C.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR* 12:2121–2159.

Erhan, D.; Bengio, Y.; Courville, A.; and Vincent, P. 2009. Visualizing higher-layer features of a deep network. Technical report, University of Montreal.

Feng, Y.; Yuan, Y.; and Lu, X. 2017. Learning deep event models for crowd anomaly detection. *Neurocomputing* 219:548 – 556.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*, 2672–2680.

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning Temporal Regularity in Video Sequences. In *CVPR*.

Isola, P.; Zhu, J.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*, 5967–5976.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 1106–1114.

Li, W.-X.; Mahadevan, V.; and Vasconcelos, N. 2014. Anomaly Detection and Localization in Crowded Scenes. In *TPAMI*, volume 36, 18–32.

Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal Event Detection at 150 FPS in MATLAB. In *ICCV*.

Luo, W.; Liu, W.; and Gao, S. 2017. Remembering history with convolutional lstm for anomaly detection. In *ICME*, 439–444.

Mahadevan, V.; LI, W.-X.; Bhalodia, V.; and Vasconcelos, N. 2010. Anomaly Detection in Crowded Scenes. In *CVPR*, 1975–1981.

Ravanbakhsh, M.; Nabi, M.; Sangineto, E.; Marcenaro, L.; Regazzoni, C. S.; and Sebe, N. 2017a. Abnormal Event Detection in Videos using Generative Adversarial Nets. In *International Conference on Image Processing (ICIP)*, 1577–1581.

Ravanbakhsh, M.; Sangineto, E.; Nabi, M.; and Sebe, N. 2017b. Training Adversarial Discriminators for Cross-channel Abnormal Event Detection in Crowds. *CoRR*.

Ravanbakhsh, M.; Nabi, M.; Mousavi, H.; Sangineto, E.; and Sebe, N. 2018. Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection. In *IEEE Winter Conference on Applications of Computer Vision WACV*, 1689–1698.

Ribeiro, M.; Lazzaretti, A. E.; and Lopes, H. S. 2017. A Study of Deep Convolutional Auto-Encoders for Anomaly Detection in Videos. *Pattern Recognition Letters*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. Imagenet large scale visual recognition challenge. *IJCV* 115(3):211–252.

Sabokrou, M.; Fathy, M.; Hosseini, M.; and Klette, R. 2015. Real-Time Anomaly Detection and Localization in Crowded Scenes. *CVPRW*.

Sabokrou, M.; Fayyaz, M.; Fathy, M.; and Klette, R. 2017. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing* 26(4):1992–2004.

Singh, D., and Mohan, C. K. 2017. Graph Formulation of Video Activities for Abnormal Activity Recognition. *Pattern Recognition* 65:265–272.

Sodemann, A. A.; Ross, M. P.; and Borghetti, B. J. 2012. A Review of Anomaly Detection in Automated Surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42:1257 – 1272.

Tran, H., and Hogg, D. 2017. Anomaly Detection using a Convolutional Winner-Take-All Autoencoder. In *BMVC*.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*, 1096–1103.

Vu, H.; Nguyen, T. D.; Travers, A.; Venkatesh, S.; and Phung, D. 2017. Energy-Based Localized Anomaly Detection in Video Surveillance. In *The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

Xu, D.; Ricci, E.; Yan, Y.; Song, J.; and Sebe, N. 2015. Learning deep representations of appearance and motion for anomalous event detection. *BMVC*.

Xu, D.; Yan, Y.; Ricci, E.; and Sebe, N. 2017. Detecting Anomalous Events in Videos by Learning Deep Representations of Appearance and Motion. *Computer Vision and Image Understanding (CVIU)* 156:117–127.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*, 818–833.

Zhang, Y.; Lu, H.; Zhang, L.; and Ruan, X. 2016. Combining Motion and Appearance Cues for Anomaly Detection. *Pattern Recognition* 51:443–452.