



Real-time adversarial GAN-based abnormal crowd behavior detection

Qiulei Han¹ · Haofeng Wang² · Lin Yang² · Min Wu² · Jinqiao Kou² · Qinsheng Du¹ · Nianfeng Li¹

Received: 25 March 2020 / Accepted: 6 October 2020 / Published online: 31 October 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Detecting abnormal events in the crowd is a challenging problem. Insufficient samples make those traditional model-based methods cannot cope with sophisticated anomaly monitoring. Therefore, we design a real-time generative adversarial network plus an add-on encoder to deal with the continually changing environment. After the generator reconstructs the compressed pattern to generate the design to the latent vector, a discriminator is used to construct better videos by minimizing the adversarial loss function. We calculated the abnormal score by the distance between the two underlying patterns encoded by the first and the second encoders. The unusual event is detected when the anomaly score is above the threshold. To accelerate the processing efficiency, we introduced the grouped pointwise convolution method to decrease the computing complexity. The frame-level and video-level experiments on the benchmark dataset show the accuracy and reliance of our approach. The acceleration approach can increase the efficiency of the network with only limited accuracy loss.

Keywords Generative adversarial network (GAN) · Anomaly detection · Auto-encoder · Grouped pointwise convolution

1 Introduction

Public places use more and more surveillance cameras, e.g., public transportation systems, hospitals, shopping malls, parks, etc. The enormous safety cameras create a massive amount of videos, and potential application includes object detection, tracking, image retrieval, and so on. At the same time, limited human monitoring cannot keep up with the demand for surveillance. Automated alarming abnormal events is an urgent problem to conquer.

In surveillance videos, suspicious events are challenging to describe and predict because of the intravariance and inter-similarity. We assume that regular competition indicates commonality. Therefore, an unusual event is associated with the distinctness of the activity. For example, we interpret a running person in a everybody-walks scene as abnormal. Unfortunately, due to the elaborate scenes and the uncertainty of anomaly, this abnormal detection is still challenging and hard to be distinguished.

In relation to abnormal activity detection, numerous studies declare they can solve the problem in different application environments. For unsupervised and semi-supervised problems, a generative adversarial network (GAN) has become a representative method in the artificial intelligence field. In the traditional network, the high dimensional vector needs to be transferred to a latent vector to resemble the source data. We use the following data to separate the transferred and primitive data. Many approaches have improved training stages problems. GAN shows better performance compared to traditional methods. We select adversarial training to handle abnormal activity detection [1–6]. There have been many recent developments of variational auto-encoder (VAE) and GAN. GAN shows more promising results in generative natural images than conventional generative models. GAN is semi-supervised learning of rich feature representations for arbitrary data distributions. It is especially appropriate for anomaly detection. A double-encoder network enables the model to generate images to the underlying representation in the training phase. GAN learns the regular events by minimizing the distance between the adjacent video frame and the latent vectors. By inference, a high anomaly score shows an anomaly by reporting evidential improvement over the previous work [7–14].

We employ adversarial auto-encoder within an encoder–decoder–encoder pipeline, capturing the training

✉ Qiulei Han
hql4444@hotmail.com

¹ School of Computer Science, Changchun University, Changchun 130022, People's Republic of China

² Beijing Institute of Computer Technology and Applications, Beijing 110000, People's Republic of China

data distribution within both image and latent vector space. A different training architecture, such as this, practically based on only regular training data examples, produces superior performance over challenging benchmark problems. We propose a double-encoder GAN architecture to detect abnormal crowd events. It can automatically detect abnormal events by learning the inner patterns of changing sceneries. A generative adversarial network based on a double auto-encoder is proposed:

Exact and fuse the motion feature of normal status.
Learn the regular pattern by generator and discriminator.
Detect the anomaly by the score calculated by the add-on encoder.

The contributions of this paper are:

- A semi-supervised abnormal activity learning network is designed for anomaly detection. This network contains a double-encoder sub-network and GAN.
- A grouped pointwise convolution method is used to accelerate the processing efficiency and decrease the computing complexity.

Experiments on the publicly available dataset show that the proposed approach is competitive both in effectiveness and efficiency compared with state-of-the-art approaches in crowd behavior detection tasks.

The remainder of the paper is organized as follows. In Sect. 2, we review the relevant works, including anomaly detection, motion feature, deep learning structure, and efficient neural network. We describe the detail of our approach in Sect. 3. We conducted experiments to evaluate and analyze our method in Sect. 4. Finally, conclusions are provided in Sect. 5.

2 Related work

In this section, we briefly review related work in anomaly detection, motion feature, and deep learning structure. We present the advantages and drawbacks of some approaches [15–18].

2.1 Anomaly detection

Anomaly detection in a video is a challenging problem in surveillance applications. The approaches can be classified into supervised [17, 19, 20], semi-supervised, and unsupervised [3, 21] approaches. The supervised method needs a lot of labeled standard data and abnormal data to train a model. Nevertheless, in the real world, a large number of labeled data are often not available.

Recently, many semi-supervised or unsupervised methods are presented, using limited annotated data or even do not use any annotation information, which brings abnormal event detection to be more practical for end-to-end use. Those unsupervised methods only need standard samples which do not include anomalies. The approach designs a model for learning standard patterns. Any deviations from this normal can be identified as an anomaly by measuring a loss or score function [22–24].

2.2 Motion feature

The traditional manual designed features, such as histogram of oriented gradients (HOG) and the histogram of oriented flow (HOF), show excellent performance in capturing the motion features. Recently, C3D [25] is mostly efficient networks for specific spatiotemporal features from videos. However, the manner it combines the traditional function and the deep learning-based functionality is attractive but underexplored.

Hasan et al. [26] combine HOG and HOF to catch the motion features. Weixin li and Vijay Mahadevan detect anomaly in crowd scenes using a set of a mixture of dynamic textures [27]. Waqas and Chen use C3D to exact features and put them into multiple instances learning to predict high anomaly in videos [28]. Rohit et al. build a two-stream network to classify actions. One stream is for appearance features from raw images, and another is for motion features from optical flows [29]. However, these methods need to detect the motion features first, and this procedure can influence the effectiveness of anomaly detection.

2.3 Deep learning structure

In the related work, the reconstruction model, predictive model, and generative model are mostly efficient structures in anomaly detection [30–33].

Auto-encoder is the reconstruction model that reduces the dimensionality of data by minimizing the reconstruction error. Mohammad and others decompose the deep auto-encoder and CNN into several cascaded classifiers in each sub-stage [34]. But the auto-encoder algorithm requires an objective function for evaluating the accuracy of encoded/decoded input data. In most applications, it is not possible to represent the real objective.

The long short-term memory model (LSTM) is used to predict unusual events. They build a convolutional LSTM network for the precipitation new-casting problem by deriving the adjacent video frame from the reconstruction errors of a set of predictions. It is an end-to-end trainable model. They LSTM yields lower regularity scores with abnormal video sequences from the original series over time. However,

LSTM is relatively complex, with more parameters and heavier computation.

Both VAE and GAN are generative models that approximate the data distribution of high-dimensional data input. The shortcoming of VAE is that the generated samples can be blurry because of the injected noise and imperfect elementwise measures. GAN uses adversarial training to learn patterns with the generative and discriminative models. By a novel anomaly scoring scheme, they design a deep convolutional GAN to trace a manifold of normal anatomical variability [35]. The samples generated from GAN often are far from natural because GAN can not cover the training data. Recently, some papers combine the auto-encoders with GAN. Dimokranitou [36] uses an arbitrary prior distribution to match the aggregated posterior of the hidden code vector of the auto-encoder. We get a low adversarial error of the learned auto-encoder for regular events and get a high error for irregular activities. Larsen et al. combined VAE with a generative adversarial network to learn feature representations [37].

2.4 Efficient neural network

Deep convolutional neural network (CNN) models are computationally expensive and memory intensive. Therefore, many techniques concentrate on model compression and acceleration. But the magnitude of the model is not effectively simplified. For example, those parameters which are not sensitive to effectiveness can be reduced. However, we need to manually eliminate parameters so that the fine-tuning process can be infelicitous in practical applications [38, 39].

Using matrix decomposition, those low-rank factorization-based techniques, such as MobileNets, can estimate the informative parameters of the deep CNNs. Because of the computational complexity of decomposition operation, global parameter compression cannot be executed layer by layer. We design structural convolutional filters for compact convolutional filters to decrease parameters. However, to refine the explosive results on some datasets, we need to set up exact transfer assumptions.

Knowledge distillation methods like using a distilled model train a more compact neural network to recreate the larger network's output to distillate knowledge, like FitNets. The shortcoming of the method is that it is too stern and cannot be implemented in practical use [40, 41].

3 The proposed system

In this section, we first review the GAN and the improvements on it. Next, we introduce our network structure, the encoder, the decoder, the add-on encoder, and the

discriminator. The combination of encoder and decoder acts as the generator of the VAE/GAN network. Finally, we describe the model optimization, initialization, and test stage.

3.1 GAN

GAN has two adversarial parts; each of them is a competing neural network. The generative model catches the data distribution and the discriminative model calculates the probability. Training a GAN network does not need to make any Monte Carlo approximations.

However, searching a Nash equilibrium of a game is the core work of training a GAN; it often accompanies instability problems. A lot of practice is made to optimize the method of GAN. Radford and Chintala proposed the Deep Convolutional GAN (DCGAN) [2]. A fully convolutional generative network replaces pooling layers with stridden convolutions and fractional-stridden convolutions. It used batch norm in both the generator and the discriminator. We remove fully connected hidden layers for deeper architectures. We use ReLU activation in a generator for all layers except for the output. In the discriminator, Tanh and Leaky ReLU activation is used for all layers.

3.2 Overall scheme

As shown in Fig. 1, we build a three-channel stream auto-encoder and GAN architecture for crowd anomaly detection. The network contains three sub-networks: A fully connected auto-encoder and a decoder part. It is an adversarial auto-encoder. The input data are mapped to latent space and are remapped back to input data space. An add-on auto-decoder part is an additional encoder. It maps the data from the image space to the latent space. The discriminator part has a classical

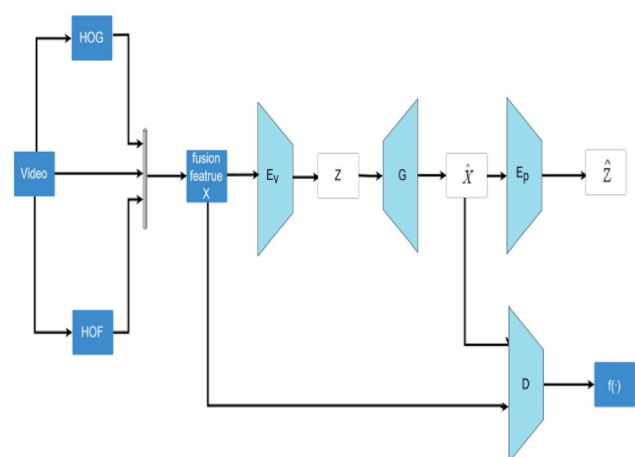


Fig. 1 The architecture of double encoded GAN

classification architecture. It is used for determining the validity of the data.

To detect abnormal events in a crowd scene, we provide competing auto-encoders within a double-encoder model, learning the training data organization with both motion features and potential vector space. This design can produce combatant performance over challenging bench problems. A 3D fully convolutional network E_v is used as the encoder to get latent vector z , and a 3D deconvolutional network G is applied as the decoder. There are three inputs. (1) The gray image video: $I[1:t]$, denoted by V_{xyz}^I , (2) the HOG sequence: $G[1:t]$, denoted by V_{xyz}^G , (3) the optical flow sequence: $F[1:t]$, denoted by V_{xyz}^F . We concatenate V_{xyz}^I , V_{xyz}^G , and V_{xyz}^F into V_f . This fusion can enhance the motion features' ability in the algorithm.

The video combined with the traditional features (HOG and HOF) is used as the input, E_v is the encoder to get latent vector z , and G is applied as the decoder to get output. E_p is encoder in the second sub-network, and D is the discriminator in the third sub-network, and it classifies the input x and out as real or fake.

The output of the 3D network is calculated by,

$$v_{xyz}^I = \sigma \left(\sum_{i=1}^w \sum_{j=1}^h \sum_{k=1}^d \omega_{ijt}^I k_{(x+i)(y+j)(z+k)}^I \right), \quad (1)$$

$$v_{xyz}^G = \sigma \left(\sum_{i=1}^w \sum_{j=1}^h \sum_{k=1}^d \omega_{ijt}^G k_{(x+i)(y+j)(z+k)}^G \right), \quad (2)$$

$$v_{xyz}^F = \sigma \left(\sum_{i=1}^w \sum_{j=1}^h \sum_{k=1}^d \omega_{ijt}^F k_{(x+i)(y+j)(z+k)}^F \right), \quad (3)$$

where w, h, d is the kernel size of 3D convolution, σ is the activation function, and b is the bias. ω_{ijk} is the weight at position (i, j, k) , k^{IGF} is the kernel function.

The encoder has three convolutional layers and two pooling layers. The decoder has three deconvolutional layers and two unpooling layers. The input sinogram image has a uniform size of 512×512 . The three 3D convolutional layers have 512, 256, and 128 filters, respectively. The first layer is with a stride of 4, and it produces 512 feature maps with a resolution of 55×55 pixels. The pooling layers have a kernel size of 2×2 pixels. The encoder provides 128 feature maps of size 13×13 pixels. The deconvolving and unpooling the data are in reverse order of magnitude. They are reconstructed into the decoder networks.

3.3 The frame encoder and the adversarial discriminator

We design an add-on encoder E_p to compress the reconstructed video. This training encoder with an adversarial

setting can not only better reconstruct the data but also control the latent space. The network architecture is shown in Fig. 1. We aim to map the E_v pattern of common video features. The add-on encoder E_p compresses the video by minimizing the distance between \hat{X} and \hat{Z} .

At the test phase, this add-on encoder can detect the anomaly by minimizing the latent vectors' distance with its parameterization.

$$\hat{Z} = E_p(\hat{X}). \quad (4)$$

Besides this, at the testing stage, we can detect the anomaly at the test stage by calculating the abnormal score. At the test phase, this add-on encoder can detect the anomaly by minimizing the latent vectors' distance with its parameterization. By the adversarial learning of GAN, the standard motion feature can be differentiated better. We follow the architecture proposed in DCGAN [2].

3.4 Model optimization and initialization

To achieve better performance, we train our model with three object function shown as follows.

Reconstruction function brings less blurry results than we use to restrain G by calculating the difference between the real video and the created. The reconstruction loss is defined as,

$$\ell_{\text{rec}} = E_{x \sim p_x} \|x - G(x)\|_1. \quad (5)$$

Encoder function: We build a predicting loss to decrease the distance between the compressed feature Z and the encoded features of generated video(). The is formally defined as

$$\ell_{\text{pre}} = E_{x \sim p_x} \|E_v(X) - E(G(Z))\|_1. \quad (6)$$

Adversarial function. Instead of updating G by the output of discriminator, we update generator G by the compressed pattern of discriminator D . We define function $f(x)$ as the output of the discriminator D . An input x draw from the input data distribution P_x is used to calculate the distance between the feature representation of the real data and the generated data.

The antagonistic loss for the discriminator is written as:

$$\ell_{\text{adv}} = E_{x \sim p_x} \left\| f(x) - E_{x \sim p_x} f(G(x)) \right\|_2. \quad (7)$$

Therefore, the total loss function to be optimized in our approach is

$$\ell_{\text{total}} = \ell_{\text{rec}} + \lambda_1 \ell_{\text{adv}} + \lambda_2 \ell_{\text{pre}}, \quad (8)$$

where λ_1 and λ_2 are weight factors to balance the effect of three different terms, and their values are determined by the

experimental results which are introduced in the ablation study.

We train the network by the normal motion video in datasets. We observed that the Xavier format is more stable than the Gaussian boot. So Xavier algorithm is used in our algorithm.

3.5 Model testing

At the test stage, the predicting encoder detects a given anomaly by the scoring calculated by Eq. 8.

The anomaly score is defined as

$$S_{abn} = \left\| E_{pre}(\hat{x}) - E_v(G(\hat{x})) \right\|_1. \tag{9}$$

Then, by normalizing S_{abn} , anomaly detection in crowds S_{abn} can be defined as

$$S'_{abn} = \frac{S_{abn} - \min S_{abn}}{\max S_{abn} - \min S_{abn}}. \tag{10}$$

The Eq. 10 will yield an anomaly score for the anomaly data.

3.6 Efficiency

MobileNet uses depthwise convolution to efficiently reduce complexity, extract information independently from each channel of the feature maps of the other layer, and then fuse it across the channels by using pointwise convolution (1×1 convolution).

Suppose DF is the spatial width and height of a square input attribute map, and M is the number of input channels. DK is the spatial dimension of the depth-wise convolution kernel. N is the number of output channels in pointwise convolution. Then, the depthwise convolution (see Fig. 2b) has a computational cost of $DK \times DK \times M \times DF \times DF$, and pointwise convolution (see Fig. 2c) has a value of $N \times M \times DF \times DF$.

We are motivated to accelerate the pointwise convolution because the channel number is usually greater than the square of kernel size. This means that in neural networks, pointwise convolutions only use a limited number of channels to meet quality restrictions, which might hinder the accuracy of the segmentation. To improve the efficiency of pointwise convolution, we have designed grouped pointwise convolution (GPC) in the backbone based on group convolution. Our approach is illustrated in Fig. 2d. The rectangles or squares represent convolution filters.

Suppose we use G groups in GPC. By ensuring that each convolution manipulates only on the corresponding input channel group, group convolution greatly decreases the computation cost. The outputs from a definite channel are

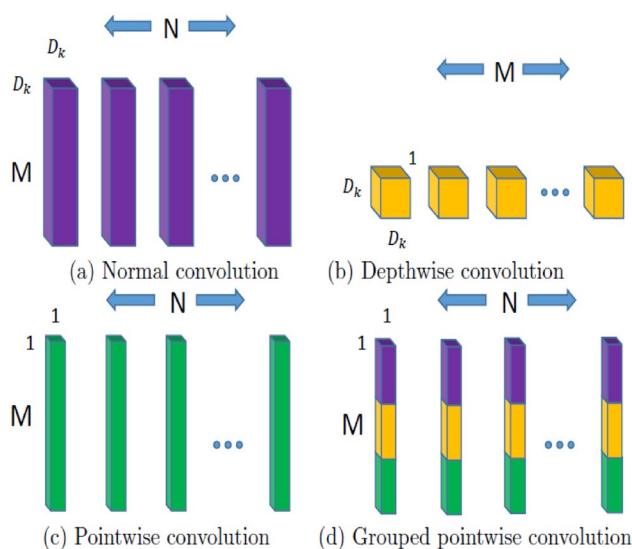


Fig. 2 Illustrations of acceleration network. The rectangles or squares represent convolution filters. Different colors represent different groups in the layer

derived from a small fraction of input channels. We divide the channels into each group into several subgroups to alleviate the side effect. We also select from the next layers each group with subgroups different from groups the other layers. A channel shuffle operation can efficiently and elegantly implement this method. Our approach can effectively decrease computation in pointwise convolution from $N \times M \times DF \times DF$ to $N \times (M/G) \times DF \times DF$. The information between different groups in the current layer exchanges in the group convolution in the next layer. Moreover, channel shuffle is also differentiable, which means it can be embedded into network designs for lengthwise training.

4 Experiments and comparisons

The input video has size 180×180 . To handle the problem that lacking of annotated instances in the dataset, the data augmentation approaches are applied, for instance, randomly crop the video of size 45×45 , rotating a stochastic angle, flipping horizontally and vertically, and adding Gaussian blur to the video frames with a definite probability.

4.1 Implementation details

The input video has a size 180×180 . We apply the data augmentation approaches to handle the problem that was lacking of annotated instances in the dataset. For instance, randomly crop the video of size 45×45 , rotating a stochastic angle, flipping horizontally and vertically, and adding Gaussian blur to the video frames with a definite probability.



Fig. 3 Scene 1 for the detection of the abnormal activities in the UMN data. Scene 1 is an outdoor scene

The parameter is chosen by referencing the experimental results, and the details are introduced in the ablation study. Our proposed GAN is trained on stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decrement of 0.005. We also use the “poly” policy to adjust the learning rate and avoid shocks in the performance curve, and the initial learning rate is set to 0.005. The batch size is set to 8 because of memory limitations. We also normalize each sample independently (i.e., instance normalization), that is, subtracting the mean and dividing by the standard deviation of the sample.

4.2 Frame-level experiments

We evaluate our algorithm’s pixel-level performance with state-of-the-art methods on a dataset of the unusual crowd activity from the University of Minnesota [42]. A database created by the University of Minnesota, consisting of 11 videos, includes normal and abnormal videos. The beginning of each video is normal behavior, followed by an unusual behavior video sequence. The crowd’s unusual behavior mainly includes the one-way running of the crowd, scattered crowd, etc. The abnormal behavior in this database is artificially arranged. The database aims at the recognition of the general crowd behavior.

We aim to detect an abnormal event as soon as possible. The frames that almost all the pedestrians escaped are discarded in our test. We use the same video as the experiments in [42]. We use the first 300 frames in the first clip in each scene for training.

Figures 3, 4, and 5 are scene 1, scene 2, and scene 3 for detecting the abnormal activities in the UMN data. Table 1 shows the qualitative results for the perception of abnormal behaviors. Our method can detect earlier than the other three traditional methods. For the first scene 1, our method has one model later than the ground truth to initiate the alarm when the crowd begins to run. While the testing results of the methods from reference [43–45] with 17, 10, and 1 frame slow alarming, respectively. For scene 2, when an unusual event occurs, our method begins to trigger the alarm after two frames, while the reference methods alarm abnormal after 14 and 8 frames,



Fig. 4 Scene 2 for the detection of the abnormal activities in the UMN data. Scene 2 is an indoor scene



Fig. 5 Scene 3 for the detection of the abnormal activities in the UMN data. Scene 1 is an outdoor scene

Table 1 Comparison of the use of our method with other classical methods for detecting abnormal activities in the UMN data

Method	Scene 1 frame no.	Scene 2 frame no.	Scene 3 frame no.
Ground truth	678	466	724
Histogram of oriented gradients [45]	696	482	741
Histogram of oriented optic flow [44]	689	478	734
Violence flows ViF [43]	680	466	728
Ours	679	468	727

respectively. The results of Ref. [44] outperform our algorithm and the other two algorithms. For scene 3, our process triggers the alarm when the crowd becomes abnormal after three frames, methods in the reference alarmed at 14, 7, and 1 frame late. We implement the designed method on Tensorflow. Adam is used to optimize the networks with a learning rate of 0.00002 and a batch size of 4. Table 1 shows that our detection accuracy is superior to the other three methods.

At the training stage, the decoder is to train to minimize. Then, the decoder and the discriminator are to be trained in a competing way to decrease and. The tan h and the sigmoid were used as activation functions in the decoder and the discriminator. Besides the raw video, we use the HOG and HOF to represent the motion features of a sequence. We implemented HOG and optical flow in OpenCV package. We performed experiments on Intel Xeon E5-2630 v4 processor and NVIDIA GTX Titan X GPU.

Table 2 Performance comparison on UMN Dataset. EER and AUC are measured

Method	EER	AUC (%)
Chaotic invariants [47]	5.3	99.4
Sabokrou et al. [46]	2.5	99.6
Sparse [42]	2.8	99.6
Sligrama et al. [48]	3.4	99.5
Li et al. [27]	3.7	99.5
Ours	4.1	99.8

4.3 Performance evaluations

Because the UMN dataset does not provide the pixel-level ground truth, EER and AUC in frames are used to evaluate our method. Table 1 shows the performance of our way and the benchmark methods [46]. We use the EER and AUC in frame-level to evaluate our process. The EER and AUC results are shown in Table 2. We only use the global detector to detect an anomaly. Previous approaches performed reasonably well on this dataset. The AUC of our application is comparable with the otherwise best result, and the EER of our practice is the second better among the best previous ways.

4.4 Video-level experiments

In our experiment, we demonstrate the effectiveness of the proposed method on the Violence-in-Crowd dataset [43]. The crowd database created by the Open University of Israel focuses on crowd violence. It consists of 246 videos, all of which are downloaded from YouTube, and the video source is real violence video. The database is designed to provide a test basis for testing violence/Non Violence classification and violence standards. In the video, the shortest clip duration is 1.04 s, and the longest clip is 6.52 s, while the average length of the video clip is 3.60 s.

Because the UMN dataset does not provide the pixel-level ground truth, EER and AUC in frames are used to evaluate our method. Table 1 shows the performance of our way and the benchmark methods [46]. We use the EER and AUC in frame-level to evaluate our process. The EER and AUC results are shown in Table 2. We only use the global detector to detect an anomaly. Previous approaches performed reasonably well on this dataset. The AUC of our application is comparable with the otherwise best result, and the EER of our practice is the second better among the best previous ways.

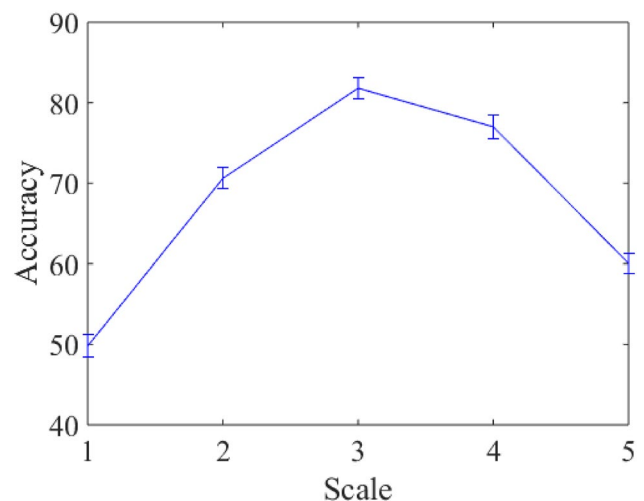
4.5 Ablation experiments

Backbone In the encoder, the scales are verified, and the optimum number of the scale is selected according to Fig. 6. In the experimental results, $n=3$ scales achieve the best performance. If more scales are applied, the nonlinear relation between the input and the latent vector is modeled better. Nevertheless, the overfitting may be caused by simply increasing the complexity of the network. Therefore, we used the encoder with three scales for effectiveness and efficiency comparison in the following experiments.

Backbone In the encoder, the scales are verified, and the optimum number of the scale is selected according to Fig. 6. In the experimental results, $n=3$ scales achieve the best performance. If more scales are applied, the nonlinear relation between the input and the latent vector is modeled better. Nevertheless, the overfitting may be caused by simply increasing the complexity of the network. Therefore, we used the encoder with three scales for effectiveness and efficiency comparison in the following experiments (Fig. 7).

Efficiency We evaluate the use of different backbones for the network, including MobileNetV2 and ResNet-101, because MobileNetV2 has relatively low computational complexity and is widely used for real-time applications ResNet-101 yields high accuracy by increasing the complexity. To fully exploit the potential in acceleration, all the pointwise convolution in the backbone is replaced by GPC.

We compare the variance of the Jaccard index and floating-point operations per second (FLOPs) according to change in the group number provided in Table 3. In this table, it can be found that the FLOPs could be further improved when the group number is increased from 1 to 3 in

**Fig. 6** Evaluation of the number of scales on the violence-in-crowd dataset

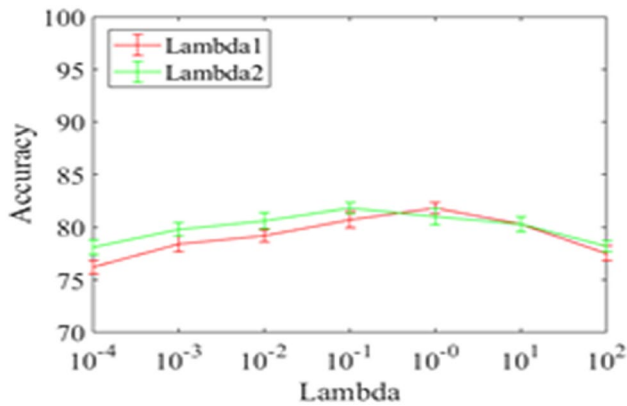


Fig. 7 Evaluation of the parameters lambda1 and lambda2 on the violence-in-crowd dataset

Table 3 Performance comparison on violence-in-crowd dataset

Method	Accuracy (%)
Histogram of oriented gradients [45]	57.3
Histogram of oriented optic flow [44]	58.5
Violence Flows ViF [43]	81.3
Tracklet-based commotion measure [49]	81.6
Analyzing tracklets for the detection [50]	78.3
Ours	81.8

grouped pointwise convolution with only limited accuracy loss.

However, when the group number continues to increase (from 3 to 5), the computational complexity decreases. At the same time, a lot of the fine information is lost, because the relation between non-local channels (in different groups) is omitted in this approach. Our result shows the best accuracy comparing with all previous methods, as reported in Table 4.

Table 4 Comparison of effectiveness and efficiency of pointwise convolution and our grouped pointwise convolution on the validation set

Approach	Backbone	FLOPs	Jaccard Index
Pointwise (G=1)	MobileNetV2	3.51B	76.4
GPC (G=3)	MobileNetV2	3.07B	76.3
GPC (G=5)	MobileNetV2	2.99B	76.0
Pointwise (G=1)	ResNet-101	45.63B	86.4
GPC (G=3)	ResNet-101	38.16B	86.3
GPC (G=5)	ResNet-101	36.47B	85.9

5 Conclusion

Our experimental results show that our method is both effective and efficient. Specifically, we present a semi-supervised method to detect abnormal events. We also use the grouped pointwise convolution which can accelerate the processing with only limited accuracy loss. Future work on anomaly detection could simplify the double-encoder architecture to decrease the complexity of the network. The potential application of the proposed method is safety surveillance and disaster avoidance.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grant 61972351, in part by the Natural Science Foundation of Zhejiang Province under Grant LY19F030005 and Grant LY18F020008, in part by the Opening Foundation of State Key Laboratory of Virtual Reality Technology and System of Beihang University under Grant VRLAB2020B15.

References

- Danelljan, M., Bhat, G., Gladh, S., Khan, F.S., Felsberg, M.: Deep motion and appearance cues for visual tracking. *Pattern Recogn. Lett.* **124**, 74–81 (2018)
- Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. [arXiv:1511.06434v2](https://arxiv.org/abs/1511.06434v2)
- Akay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: semi-supervised anomaly detection via adversarial training. In: *Asian Conference on Computer Vision (ACCV)*, pp. 622–637 (2019)
- Runsheng, Y., Zhenyu, S., Qiongxiang, M., Laiyun, Q.: Predictive learning: using future representation learning variational autoencoder for human action prediction. [arXiv:1711.09265](https://arxiv.org/abs/1711.09265) (2016)
- Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017, pp. 5967–5976 (2017)
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. [arXiv:1606.03498](https://arxiv.org/abs/1606.03498)
- Ihaddadene, N., Djeraba, C.: Real-time crowd motion analysis. In: *International Conference on Pattern Recognition*, pp. 1–4 (2008)
- Zhang, X., Zhang, Q., Hu, S., Guo, C., Yu, H.: Energy level-based abnormal crowd behavior detection. *Sensors*, **18**, 423 (2018)
- Saligrama, V., Chen, Z.: Video anomaly detection based on local statistical aggregates. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2112–2119 (2012)
- Ali, S., Waqas, M., Chen, N., Chen, D., Han, Y., Boateng, B., Xiong, J., Han, J., He, W.: Three-dimensional twisted fiber composite as high-loading cathode support for lithium sulfur batteries. *Compos. B Eng.* **174**, 107025 (2019)
- Tian, Y., Cheng, G., Gelernter, J., Yu, S., Song, C., Yang, B.: Joint temporal context exploitation and active learning for video segmentation. *Pattern Recogn.* **100**, 107158 (2020)
- Tian, Y., Gelernter, J., Wang, X., Li, J., Yu, Y.: Traffic sign detection using a multi-scale recurrent attention network. *IEEE Trans. Intell. Transp. Syst.* **20**, 4466–4475 (2019)

13. Tian, Y., Wang, X., Wu, J., Wang, R., Yang, B.: Multi-scale hierarchical residual network for dense captioning. *J. Artif. Intell. Res.* **64**, 181–196 (2019)
14. Tian, Y., Hu, W., Jiang, H., Wu, J.: Densely connected attentional pyramid residual network for human pose estimation. *Neurocomputing* **347**, 13–23 (2019)
15. Tian, Y., Chen, T., Cheng, G., Yu, S., Li, X., Li, J., Yang, B.: Global context assisted structure-aware vehicle retrieval. *IEEE Trans. Intell. Transp. Syst.*, 1–10 (2020)
16. Wang, X., Tian, Y., Zhao, X., Yang, T., Gelernter, J., Wang, J., Cheng, G., Hu, W.: Multi-person pose estimation by mask-aware deep reinforcement learning. *ACM Trans. Multimedia Comput. Commun. Appl.*, 84–100 (2020)
17. Tian, Y., Zhang, K., Li, J., Lin, X., Yang, B.: LSTM based traffic flow prediction with missing data. *Neurocomputing* **318**, 297–305 (2018)
18. Tian, Y., Zhang, Y., Zhou, D., Cheng, G., Chen, W.-G., Wang, R.: Triple attention network for video segmentation. *Neurocomputing* **417**, 202–211 (2020)
19. Tian, Y., Jia, Y., Shi, Y., Liu, Y., Ji, H., Sigal, L.: Inferring 3D body pose using variational semi-parametric regression. In: 18th IEEE International Conference on Image Processing, pp. 29–32 (2011)
20. Yang, B., Sun, S., Li, J., Lin, X., Tian, Y.: Traffic flow prediction using LSTM with feature enhancement. *Neurocomputing* **332**, 320–327 (2019)
21. Zeng, Q., Martin, R.R., Wang, L., Quinn, J.A., Sun, Y., Tu, C.: Region-based bas-relief generation from a single image. *Graph. Models* **76**, 140–151 (2014)
22. Chen, W., Sun, T., Li, M., Jiang, H., Zhou, C.: A new image co-segmentation method using saliency detection for surveillance image of coal miners. *Comput. Electric. Eng.* **40**, 227–235 (2014)
23. Yuan, S., Zhou, W., Chen, L.: Epileptic seizure prediction using diffusion distance and bayesian linear discriminate analysis on intracranial EEG. *Int. J. Neural Syst.* **28**, 1750043 (2018)
24. Zhou, C., Liu, C.: An efficient segmentation method using saliency object detection. *Multimedia Tools Appl.* **74**, 5623–5634 (2015)
25. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 4489–4497 (2015)
26. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: 2015 IEEE Workshop on Applications of Computer Vision, pp. 148–155 (2015)
27. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 18–32 (2013)
28. Xie, H., Yang, D., Sun, N., Chen, Z., Zhang, Y.: Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recogn.* **85**, 109–119 (2019)
29. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.C.: ActionVLAD: learning spatio-temporal aggregation for action classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017, pp. 3165–3174 (2017)
30. Qi, L., Dai, P., Yu, J., Zhou, Z., Xu, Y.: Time–location–frequency–aware internet of things service selection based on historical records. *Int. J. Distrib. Sens. Netw.* **13**, 155014771668869 (2017)
31. Hou, M., Gao, Y., Liu, J., Dai, L., Kong, X., Shang, J.: Network analysis based on low-rank method for mining information on integrated data of multi-cancers. *Comput. Biol. Chem.* **78**, 468–473 (2019)
32. Zhou, C., Liu, C.: Co-segmentation of multiple similar images using saliency detection and region merging. *IET Comput. Vision* **8**, 254–261 (2014)
33. Wang, J., Liu, J., Zheng, C., Wang, Y., Kong, X., Wen, C.: A mixed-norm Laplacian regularized low-rank representation method for tumor samples clustering. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **16**, 172–182 (2019)
34. Wei, C., Wang, P., Zhang, Y.: Entropy, similarity measure of interval-valued intuitionistic fuzzy sets and their applications. *Inf. Sci.* **181**, 4273–4286 (2011)
35. Schlegl, T., Seebeck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: The International Conference on Information Processing in Medical Imaging (IPMI), pp. 146–157 (2017)
36. Dimokranitou, A.: Adversarial autoencoders for anomalous event detection in images, Doctoral Dissertation (2017)
37. Larsen, A.B.L., Sonderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML'16), vol. 48, pp. 1558–1566 (2016)
38. Li, R., Sturtivant, C., Yu, J., Cheng, X.: A novel secure and efficient data aggregation scheme for IoT. *IEEE Internet Things J.* **6**, 1551–1560 (2019)
39. Chen, W., Wilson, J.T., Tyree, S., Weinberger, K.Q., Chen, Y.: Compressing neural networks with the hashing trick. [arXiv:1504.04788](https://arxiv.org/abs/1504.04788)
40. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
41. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: FitNets: hints for thin deep nets. [arXiv:1412.6550](https://arxiv.org/abs/1412.6550)
42. Unusual Crowd Activity Dataset of the University of Minnesota. http://mha.cs.umn.edu/proj_events.shtml
43. Hassner, T., Itcher, Y., Klipergross, O.: Violent flows: real-time detection of violent crowd behavior. In: 3rd IEEE International Workshop on Computer Vision and Pattern Recognition (CVPR), pp. 1–6 (2012)
44. Tian, W., Snoussi, H.: Histograms of optical flow orientation for visual abnormal events detection. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp. 13–18 (2012)
45. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
46. Sabokrou, M., Fathy, M., Hoseini, M., Klette, R.: Real-time anomaly detection and localization in crowded scenes. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015, pp. 56–62 (2015)
47. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In: Twenty-third IEEE Conference on Computer Vision Pattern Recognition, pp. 2054–2060 (2010)
48. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3449–3456 (2011)
49. Mousavi, H., Nabi, M., Kiani, H., Perina, A., Murino, V.: Crowd motion monitoring using tracklet-based commotion measure. In: IEEE International Conference on Image Processing (ICIP), pp. 2354–2358 (2015)
50. Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., Mur, V.: Analyzing tracklets for the detection of abnormal crowd behavior.

In: IEEE Winter Conference on Applications of Computer Vision Workshop (WACV), pp. 148–155 (2015)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Qiulei Han received a Ph.D. degree from the University of Chinese Academy of Science, Changchun, China, in 2007. She is currently an associate professor in the School of Computer Science, Changchun University, China. Her current interests are machine learning and computer vision.

Haofeng Wang received a Masters Degree from the graduate school of the University of Science and Technology of China, Beijing. He is currently a researcher of computer system architect in Beijing Institute of Computer Technology and Applications, Beijing, China. His research interests include system architect, heterogeneous computing, and deep learning.

Lin Yang received the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China. He is currently a senior engineer at the Beijing Institute of Computer Technology and Applications, Beijing, China. His research interests include embedded systems, information retrieval, and machine learning.

Min Wu received a Masters Degree in Engineering in the School of Automation Science and Electrical Engineering from Beijing University of Aeronautics and Astronautics. She is currently a general design engineer at the Beijing Institute of Computer Technology and Applications, Beijing, China. Her research interests include system architect, network communication, and embedded systems.

Jinqiao Kou received a Masters Degree in Engineering in the School of Automation Science and Electrical Engineering from Beijing University of Aeronautics and Astronautics. He is currently a system architect-engineer in the Beijing Institute of Computer Technology and Applications, Beijing, China. His research interests include system architect, heterogeneous computing, and deep learning.

Qinsheng Du received a Ph.D. degree from Jilin University, Changchun, China, in 2015. He is currently an Associate Professor in the School of Computer Science, Changchun University, China. His current interests are Big Data and Image Processing.

Nianfeng Li received a Ph.D. degree from Jilin University, Changchun, China, in 2008. He is currently a Professor in the School of Computer Science, Changchun University, China. His current interests are Computer application, health care assistive technology.