Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Unsupervised object detection with scene-adaptive concept learning*

Shiliang PU[1], Wei ZHAO[1], Weijie CHEN[1], Shicai YANG[1], Di XIE†‡[1], Yunhe PAN[2]

[1]*Hikvision Research Institute, Hangzhou 310051, China*

[2]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

†E-mail: xiedi@hikvision.com

Received Oct. 20, 2020; Revision accepted Feb. 24, 2021; Crosschecked Apr. 1, 2021

Abstract: Object detection is one of the hottest research directions in computer vision, has already made impressive progress in academia, and has many valuable applications in the industry. However, the mainstream detection methods still have two shortcomings: (1) even a model that is well trained using large amounts of data still cannot generally be used across different kinds of scenes; (2) once a model is deployed, it cannot autonomously evolve along with the accumulated unlabeled scene data. To address these problems, and inspired by visual knowledge theory, we propose a novel scene-adaptive evolution unsupervised video object detection algorithm that can decrease the impact of scene changes through the concept of object groups. We first extract a large number of object proposals from unlabeled data through a pre-trained detection model. Second, we build the visual knowledge dictionary of object concepts by clustering the proposals, in which each cluster center represents an object prototype. Third, we look into the relations between different clusters and the object information of different groups, and propose a graph-based group information propagation strategy to determine the category of an object concept, which can effectively distinguish positive and negative proposals. With these pseudo labels, we can easily fine-tune the pre-trained model. The effectiveness of the proposed method is verified by performing different experiments, and the significant improvements are achieved.

Key words: Visual knowledge; Unsupervised video object detection; Scene-adaptive learning
https://doi.org/10.1631/FITEE.2000567 **CLC number:** TP391

1 Introduction

Video object detection is a very important research direction in computer vision and has been widely used in many practical applications, e.g., robotics, autonomous driving, and video surveillance. Nowadays, object detection algorithms focus mainly on image detection algorithms. Existing state-of-the-art image object detection methods are mainly based on convolutional neural networks (CNNs), such as faster R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016), FPN (Lin et al., 2017a), YOLO (Redmon et al., 2016), and SSD (Liu et al., 2016). Following these works, a large number of deep learning based video detection algorithms have been proposed (Han et al., 2016; Kang et al., 2018; Xiao and Lee, 2018; Guo et al., 2019; Wang SY et al., 2019; Deng et al., 2020). For example, Kang et al. (2018) introduced a faster R-CNN based video detection model. Xiao and Lee (2018) proposed a novel video detection model combining R-FCN and RNN, called the spatial-temporal memory module. Wang SY et al. (2019) also used R-FCN as the base model to detect videos.

However, these methods cannot autonomously generalize to different scenes. In different scenes,

‡ Corresponding author

* Project supported by the National Key R&D Program of China (No. 2020AAA010400X) and the Hikvision Open Fund, China

ORCID: Shiliang PU, https://orcid.org/0000-0001-5269-7821; Di XIE, https://orcid.org/0000-0001-8065-5901

these methods require different training data, which makes each algorithm expensive. In this case, some researchers introduced unsupervised learning strategies to video detection (Tang et al., 2012; Kwak et al., 2015; Li D et al., 2016; Xiao and Lee, 2016; Croitoru et al., 2017; Ma et al., 2017; Yu et al., 2018; Chen et al., 2019; Lahiri et al., 2019). For example, Chen et al. (2019) designed an unsupervised domain adaptation based video recognition algorithm. These methods still cannot evolve adaptively with changing scenes.

The other shortcoming of existing methods is that they analyze only the information of a single object in detection. For example, Seq-NMS (Han et al., 2016) is a post-processing step that selects the high-score bounding box sequence of each object to improve the detection results. Optical flow (Zhu XZ et al., 2017), multi-branch feature map fusion (Zhu ML and Liu, 2018; Subramaniam et al., 2019), and tracking (Htike and Hogg, 2014) have also been used to propagate the object information, which further increases the complexity of the algorithms.

To address the above limitations and inspired by the fundamental concept of visual knowledge (Pan, 2020), we propose a scene-adaptive evolution unsupervised video object detection algorithm based on the visual concept (prototype and category (Pan, 2016, 2019)), which can weaken the impact of the scene change through the object group concepts. The proposed method consists of a prototype dictionary generation (PDG) strategy and a graph-based group information propagation (G-GIP) strategy. Through these strategies, the proposed method can autonomously generalize objects into different scenes and adaptively evolve with changing scenes. The designed PDG is used to build a dictionary of object prototypes. It first uses a pre-trained model to extract the object proposals and feature representation from the new scene. Similar proposals are clustered to build the dictionary of object prototypes which can limit the influence of the scene change on object detection. Different prototypes represent different object groups. Then we propose G-GIP, which combines group information and a graph convolutional network to create the category of each prototype. In this way, we can generate the positive samples and the corresponding pseudo labels. Finally, we use positive samples to fine-tune the detector, which can be used to detect new scene datasets. Experimental results show that the proposed unsupervised method can self-adaptively evolve with changing scenes.

In summary, the contributions of this paper include the following: We propose a scene-adaptive evolution unsupervised video object detection method that can evolve with changing scenes. We introduce the concept of object groups to weaken the influence of the scene change on object detection. We propose a graph-based group information propagation strategy to create the prototype category, which can mine the positive samples with pseudo labels from each object group.

2 Related works

2.1 Video object detection

The state-of-the-art object detection methods for images are divided mainly into two groups: two-stage methods and one-stage methods. The typical two-stage algorithms are R-CNN (Girshick et al., 2014), fast R-CNN (Girshick, 2015), faster R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016), FPN (Lin et al., 2017a), and Libra R-CNN (Pang et al., 2019), which consist of a region proposal, region recognition, and location. One-stage methods transform detection into a regression problem. SSD (Liu et al., 2016), YOLO (Redmon et al., 2016), RetinaNet (Lin et al., 2017b), CornerNet (Law and Deng, 2018), and FreeAnchor (Zhang et al., 2019) directly predict the location or corner point of objects and object classes. In this study, YOLO-v2 (Redmon and Farhadi, 2017) acts as the base detector.

Video object detection is similar to image object detection. The only difference is that temporal information is used in the object detection task. The existing methods can be divided into two categories based on their application of temporal information: feature-level learning (Feichtenhofer et al., 2017; Kang et al., 2017; Li JN et al., 2018; Wang SG et al., 2018; Wang SY et al., 2018; Xiao and Lee, 2018; Guo et al., 2019; Shvets et al., 2019; Li NJ et al., 2020) and post-processing strategy (Han et al., 2016; Kang et al., 2016, 2018).

For feature-level learning, Kang et al. (2017) proposed a new tubelet network and introduced an long short-term memory (LSTM) network to incorporate temporal information. PSLA (Guo et al., 2019) is a recursive feature updating subnet and

a dense feature transforming subnet to integrate features of sparse key frames and enhance features of non-key frames.

For the post-processing strategy, Han et al. (2016) proposed a simple overlap criterion, Seq-NMS, which has been designed to select boxes to maximize a sequence score. It is then used to suppress overlapping boxes to improve detection performance in video sequences. Kang et al. (2016) proposed a temporal convolutional network, which integrates the detection scores, tracking scores, and anchor offsets to improve detection performance in videos. Kang et al. (2018) designed a multi-context false sample suppression and temporal tubelet rescoring strategy to improve detection precision. In this study, YOLO acts as the base detector, and a novel post-processing algorithm based on group object information is proposed to improve detection precision.

2.2 Unsupervised video object detection

Most of the video detection methods cannot directly detect the object in a new scene. These detectors should be retrained using new datasets. In this case, some researchers proposed unsupervised video detection algorithms. The main strategies of these methods are unsupervised feature learning, object discovery, matching, tracking, and domain adaptation (Tang et al., 2012; Kwak et al., 2015; Li D et al., 2016; Xiao and Lee, 2016; Croitoru et al., 2017; Ma et al., 2017; Yu et al., 2018; Lahiri et al., 2019).

For example, Croitoru et al. (2017) proposed an unsupervised object discovery strategy based on video principal component analysis (PCA) to generate soft masks to train the detector. A tracking strategy (Yu et al., 2018) was proposed to obtain the pseudo labels for training a CNN-based object detector in video streaming. Tang et al. (2012) combined self-paced domain adaptation and a score trajectory tracking strategy to automatically mine target domain samples for unsupervised video object detection. Our method is a scene-adaptive evolution unsupervised video object detection algorithm.

2.3 Information propagation

Information propagation is an important process in video detection. Most video detection methods design a temporal information propagation strategy to improve detection results, such as optical flow strategies, deep learning models, and tracking strategies, which consume a lot of space and computing resources. For example, Zhu XZ et al. (2017) designed an optical flow based warping function to extract the features of neighboring frames to improve the detection of the current frame. Zhu ML and Liu (2018) proposed a bottleneck-LSTM layer to propagate temporal features. COSAM (Subramaniam et al., 2019) propagates temporal information through information propagation between feature maps and a weight sharing strategy. Htike and Hogg (2014) used a tracking strategy to propagate temporal information. Different from these methods, we propose a graph-based group information propagation strategy that uses the graph convolutional network to propagate the information between different object proposal groups.

2.4 Graph neural network

A large number of graph neural network models have been proposed in recent years (Kipf and Welling, 2017; Veličković et al., 2018; Wang HW and Leskovec, 2019; Wu et al., 2019). Most of them are inspired by the convolutional neural network model, such as GCN (Kipf and Welling, 2017) and GAT (Veličković et al., 2018). Taking GCN for example, the layer-wise feature propagation is

$$\boldsymbol{H}^{l+1}=\sigma(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{H}^{l}\boldsymbol{W}^{l}), \tag{1}$$

where $\boldsymbol{A}$ is the adjacency matrix, $\boldsymbol{D}$ the degree matrix, $\boldsymbol{W}^l$ the learnable weight in the l^{th} layer, $\sigma(\cdot)$ an activation function, and $\boldsymbol{H}^l=[h_1^l,h_2^l,\cdots,h_n^l]$ the l^{th} layer node feature. Therefore, these methods can be viewed as a message passing algorithm and can be applied to graph node classification. In this case, each object proposal group acts as a graph node, and the GCN model is introduced to propagate the information between different groups and identify each group for mining the training dataset in a new scene.

3 Methodology

3.1 Overview

In this study, we propose a scene-adaptive evolution unsupervised video object detection method. The proposed method can be used to detect objects in a new scene through scene-adaptive evolution.

Video datasets consist of the original scene data and the new scene data. These datasets contain similar objects and different scenes. The original scene data acts as the pre-trained data and contains manual annotations. The new scene data, which is unlabeled in the training process, acts as the detection data.

The framework of the proposed method is illustrated in Fig. 1. It consists of PDG, G-GIP, and detection. At first, original data is used to train the initial detection model. Second, the new scene data is used to establish the visual knowledge representation of objects through the PDG strategy. Each frame of new scene data is sent to the initial detection model for extracting object proposals and the corresponding features. All extracted features combined with box scale information are clustered by the k-means algorithm to build a dictionary of object prototypes. Each prototype is the center of the corresponding object group. Third, the graph-based group information propagation model is proposed to create the category of positive samples, and then to mine the new training samples with pseudo labels from the new scene dataset. Finally, the mined dataset is used to fine-tune the detection model for new scene detection.

3.2 PDG

In this study, the original and new data is sampled from different scenes. However, the classes and objects appear to be similar in the original and new data. Therefore, a PDG is designed to establish a dictionary of prototype objects to represent object groups in different scenes.

3.2.1 Pre-training

In this study, YOLO-v2 acts as the baseline model. We first send each frame of original data to the YOLO-v2 model to train the initial feature extraction model. Then the learned feature representation model is applied to represent the new scene data.

3.2.2 Similar sample group clustering

The objects appear to be similar in the original and new data. The object discriminant feature representation in the original data can be obtained

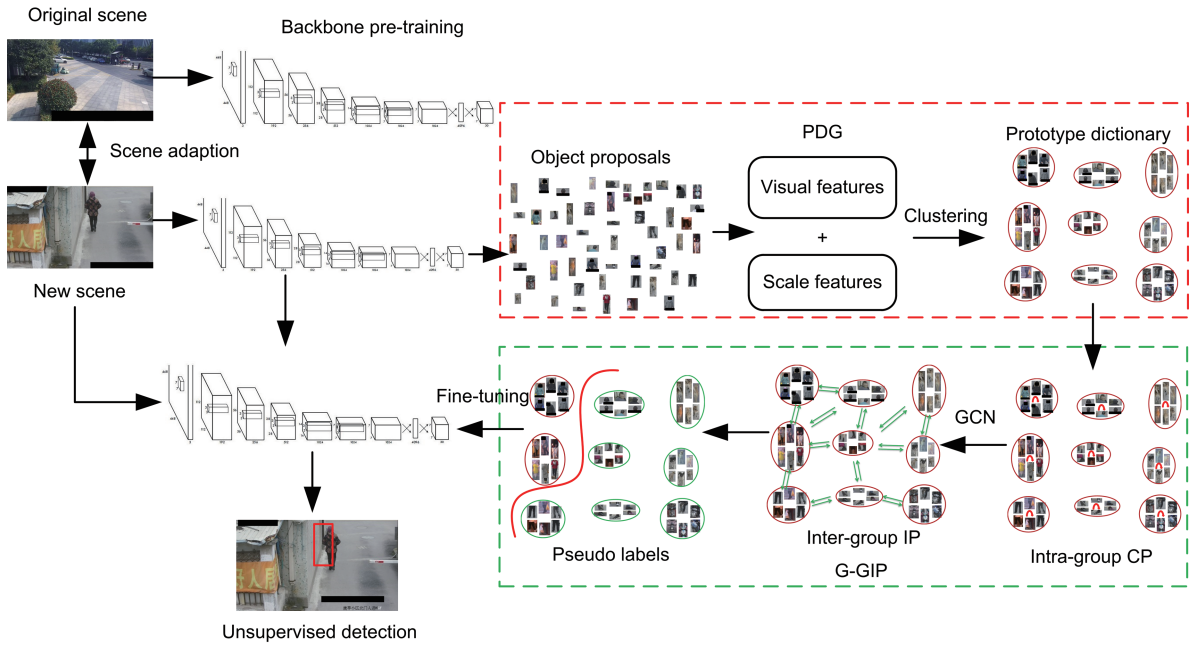

Fig. 1 Framework of the proposed method

YOLO acts as the backbone in this framework. First, the original scene is used to train the backbone network, and the new scene data is sent to the pre-trained model to generate object proposals. Then, the object proposals are used to construct the prototype dictionary of objects through PDG strategy. After that, the G-GIP strategy is used to integrate the group information of each object prototype and to generate positive samples with pseudo labels. Finally, the new scene data is detected by the detection network which is fine-tuned by the generated positive samples

by training the YOLO-v2 model. In this case, the feature representation of the object in the new scene can be obtained by sending the new scene data to the trained YOLO-v2 model. The obtained features can be used to identify the object. In different datasets, the objects have different representations even though these datasets are similar. Therefore, a similar sample group clustering strategy is proposed to build a dictionary of prototypes to represent each object group, which can better distinguish different objects.

At first, each frame of the new scene dataset is sent into the trained detection model to generate object proposals. Fig. 2a shows the obtained object proposals, which include positive and negative samples, such as pedestrian and other objects.

At the same time, the features of each object proposal at the penultimate layer of darknet19 are preserved. Then, the initial clustering centers of the object proposals are randomly generated. At last, the clustering centers are generated by the k-means algorithm. The function is designed as

$$\min \sum_{n=1}^{N} \sum_{i=1}^{K} \|\boldsymbol{C}_n - f(\boldsymbol{x}_i)\|^2, \tag{2}$$

where N is the number of clustering centers, K is the number of object proposals, $\boldsymbol{C}_n$ represents the n^{th} clustering center, $\boldsymbol{x}_i$ is the i^{th} object proposal, and $f(\boldsymbol{x}_i)$ represents the feature of object proposal. The number of initial clustering centers is set to 100. After iterations, the clustering centers can be acted as the prototype of the corresponding objects. Meanwhile, these prototypes can be represented as a group of similar samples. As shown in Fig. 2b, the object proposals of new data are divided into different groups, each of which contains similar objects. To obtain more stable clustering groups, the scale information of each object proposal is combined with the visual feature representation. The scale information is defined as follows:

$$f_{\text{s}} = \Big[\log w, \log h, \frac{w}{w_{\text{im}}}, \frac{h}{h_{\text{im}}}, \frac{w}{h}, \frac{h}{w}, \\ \mathrm{e}^{-\frac{w}{h}}, \mathrm{e}^{-\frac{h}{w}}, \log(wh), \frac{wh}{w_{\text{im}} h_{\text{im}}}\Big], \tag{3}$$

where w and h are the width and height of each object proposal respectively, and w_{im} and h_{im} are the width and height of each image, respectively. Then function (2) is changed to

$$\min \sum_{n=1}^{N} \sum_{i=1}^{K} \|\boldsymbol{C}_n - [f(\boldsymbol{x}_i), f_{\text{s}}(\boldsymbol{x}_i)]\|^2 . \tag{4}$$

Finally, visual object concepts can be used to better express new scene data through the proposed clustering process.

3.3 G-GIP

The core of the proposed algorithm is to mine the positive samples with pseudo labels from the unlabeled new scene data to train the detector. In Section 3.2.2, all object proposals are divided into different groups. However, not all these groups are used to represent positive samples. Negative samples may exist in groups. In this case, a graph-based group information propagation strategy is proposed

Fig. 2 Object proposals of a new scene dataset: (a) independent object proposals; (b) groups of object proposals

to create the category of positive samples. The G-GIP consists of intra-group confidence propagation and graph-based inter-group information propagation, as shown in Fig. 3. Intra-group confidence propagation is used to initially define object groups. Graph-based inter-group information propagation is used to improve the definition precision of object groups.

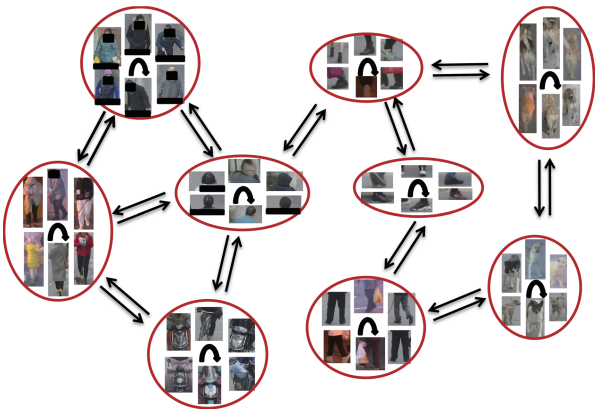

Fig. 3 Intra-group confidence propagation and inter-group information propagation of object proposals

The confidence of each object proposal can be obtained by the initial detection model. The greater the confidence value, the more likely the object proposal to be a positive sample. Each group's mean value and standard deviation are calculated for further analysis. The confidence distribution is shown in Fig. 4. As shown in Fig. 4, the confidence of the group in the upper right corner has a large mean value and small standard deviation, which indicates that this group includes large numbers of positive samples. In this case, intra-group confidence propagation is proposed to select the initial positive sample group. At first, the mean value of confidences of each object is applied to represent each group. Then, the high confidence groups are defined as the initial positive sample groups.

However, by observing the examples (in red circles), we find that some groups representing the positive objects are not distributed in the upper right corner of Fig. 4. This indicates that these samples have similar features but different confidence in this group, and that these groups will not be selected as positive groups. In this case, graph-based inter-group information propagation is proposed to mitigate this limitation and split the positive and negative groups.

Fig. 4 Confidence distribution of each group
Each patch represents one object group. References to color refer to the online version of this figure

Graph-based inter-group information propagation consists of two parts: graph construction and graph node recognition. Graph construction is proposed to build a graph to represent all groups. An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is applied to represent all groups. Each group acts as the graph node $\mathcal{V}$, and the clustering center feature of each group is applied to represent a graph node. The edge $\mathcal{E}$ is determined by the relation of one group to another. To describe the graph structure in the matrix form, an adjacency matrix $\boldsymbol{A}$ is built. A_{ij} represents the edge and the connection of node i and node j. The cosine similarity is used to generate A_{ij}:

$$A_{ij} = \frac{\boldsymbol{f}_i \cdot \boldsymbol{f}_j}{\|\boldsymbol{f}_i\| \|\boldsymbol{f}_j\|}, \tag{5}$$

where "$\cdot$" represents the vector inner product, and $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ represent the clustering center features of group i and group j, respectively. All proposals in each group are selected by the positive sample generated by the initial detection model. Therefore, there is a certain similarity between nodes. The adjacency matrix is dense, so it is difficult to separate the nodes from one another. Then, we design a row top-k strategy to improve the adjacency matrix $\boldsymbol{A}$. The detail is shown in Eq. (6):

$$A_{ij} = \begin{cases} A_{ij}, A_{ij} \in A_{i,:(\text{top3})}, & i \neq j, \\ 1, & i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

On the basis of the group graph structure, a GCN model is introduced to redefine each node. Initially, a few nodes with high and low confidence are selected as labeled nodes. Then, the GCN model is trained on the group graph structure. Finally, the positive nodes that are recognized by GCN are retained for extracting positive samples. The graph convolution process is as follows:

$$\boldsymbol{H}^{l+1}=\sigma(\hat{\boldsymbol{A}}\boldsymbol{H}^{l}\boldsymbol{W}^{l}), \tag{7}$$

where $\hat{\boldsymbol{A}}=\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{\frac{1}{2}}$, $\boldsymbol{D}$ is the degree matrix, $\boldsymbol{H}^{l}\in\mathbb{R}^{N\times m}$ is the node feature matrix (here, N is the number of nodes and m is the dimension of the node feature), $\boldsymbol{W}^{l}$ is the l^{th} layer weight, and $\sigma(\cdot)$ is the nonlinear activation function. After repeated training, we notice that the node recognition results are not stable, and are, in fact, different. However, each node represents an object group that contains a large number of objects. Obviously, this situation affects the collection of the final positive samples. To overcome this shortcoming, we add a simple subnet to modify the adjacency matrix. Through the modified adjacency matrix, we obtain a more stable graph convolution model that can generate more stable node recognition results. The subnet function is as follows:

$$A_{m(ij)}=\begin{cases}\text{leakyrelu}(\boldsymbol{w}\cdot[f_i,f_j]), & A_{ij}>0,\\ 0, & \text{otherwise},\end{cases} \tag{8}$$

where $\boldsymbol{w}$ is the network weight, "$\cdot$" is the vector inner product, and f_i and f_j are the node features. The framework is shown in Fig. 5. Finally, the object proposals belonging to all positive nodes comprise the positive samples of the new dataset for fine-tuning the detection model.

3.4 Detection

In this subsection, we introduce the detection process. Initially, the positive samples of the new scene dataset generated by G-GIP act as the new training dataset for fine-tuning the initial detection model. Then, each frame of new scene data is sent to the fine-tuned model to obtain the detection results. Finally, the results of the new scene data are obtained using the non-maximum suppression (NMS) strategy. By the above processes, the proposed method can detect the object in the new scene.

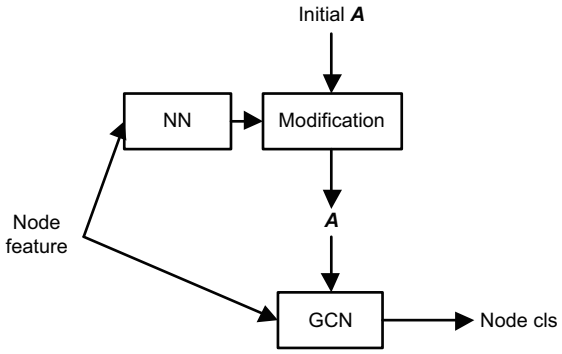

Fig. 5 Framework of the graph network (NN is the modification subnet)

4 Experiments

4.1 Dataset and setup

In this study, we propose a scene-adaptive evolution unsupervised video object detection algorithm, which is applied to detect pedestrian in scene-changed video surveillance data. In this case, we construct several video surveillance datasets named Public-Security, Residential-K, J-road, S-path, S-gate, SW-gate, and Residential-M to verify the proposed algorithm. Public-Security acts as the original scene dataset, and Residential-K , J-road, S-path, S-gate, SW-gate, and Residential-M act as new scene datasets. Public-Security includes 40k frames, Residential-K contains 58k frames of training data and 4.8k frames of test data, J-road contains 21k frames of training data and 5.7k frames of test data, S-path contains 27k frames of training data and 0.7k frames of test data, S-gate contains 11k frames of training data and 1.6k frames of test data, SW-gate contains 49k frames of training data and 5.7k frames of test data, and Residential-M contains 8.7k frames of training data and 1.1k frames of test data; the training data of new scene datasets does not contain labels. The size of all video frames is 1920×1080. All test data examples are shown in Fig. 6. These datasets include the day and night scenes.

To verify the effectiveness of the proposed algorithm, additional public datasets are introduced in the experiments. The scene-adaptive evolution unsupervised video object detection task is similar to unsupervised domain adaptation based object detection. The difference is that the scene-adaptive evolution does not use the source scene dataset to develop the model in new scene detection. Therefore, the public datasets of unsupervised domain adaptation

can be used to simulate the scene-adaptive evolution. Three public datasets are used in our experiments. KITTI (Geiger et al., 2012) contains 7481 labeled images with different scenes in a city. Sim10k (Johnson-Roberson et al., 2016) collects 10k images of different city scenes in Grand Theft Auto V. Cityscapes (Cordts et al., 2016) contains 2975 training images and 500 validation images collected from the outdoor street scenes of different cities. In this study, the KITTI and Sim10k datasets act as different source scenes and the Cityscapes dataset acts as the new scene.

The base detection model is YOLO-v2, and the backbone is darknet19. During training, Adam acts as the solver, the batch size is 20, the epoch is set to 5, the learning rate is set to 1×10^{-5}, the weight decay is 3×10^{-4}, and the NMS threshold is set to 0.3. For GCN training, the learning rate is 0.01, the weight decay is 5×10^{-4}, and there are 500 iterations. The number of cluster centers is set to 100. Finally, the Recall-FPPI0.1 (RF0.1) and average precision (AP) are used to evaluate the proposed method. All the experiments are performed using PyTorch with version 1.1.0 and PyThon running on one Tesla V100.

4.2 Results

Our method is compared with YOLO-v2 using the Residential-M dataset to demonstrate the effectiveness. The results are shown in Table 1 and Figs. 7–10. The RF0.1 of our method is 96.21%, which is 5.45% higher than that of YOLO. It indicates that our method can greatly improve the false positives. Our method achieves 97.13% AP, which is 3.03% higher than that of YOLO. This indicates

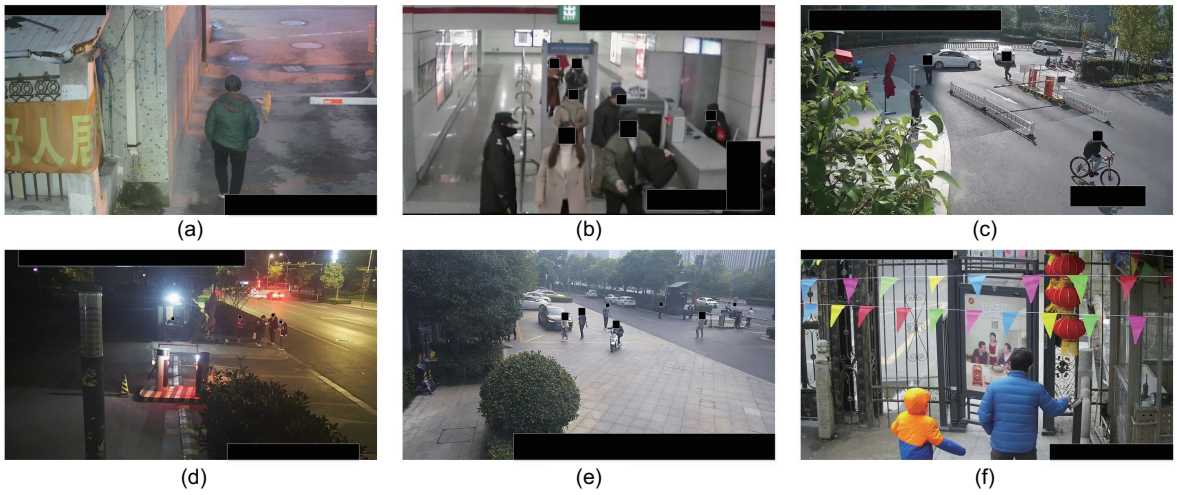

Fig. 6 Examples of all test scenes: (a) Residential-K; (b) J-road; (c) S-path; (d) S-gate; (e) SW-gate; (f) Residential-M

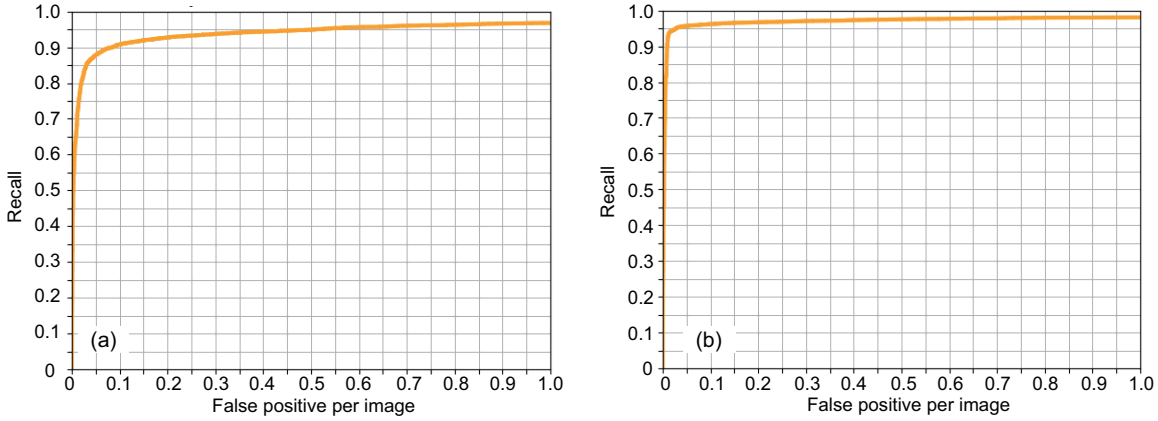

Fig. 7 Recall-FPPI curves (YOLO, class: 3, iou-thres = 0.30): (a) baseline model; (b) the proposed model

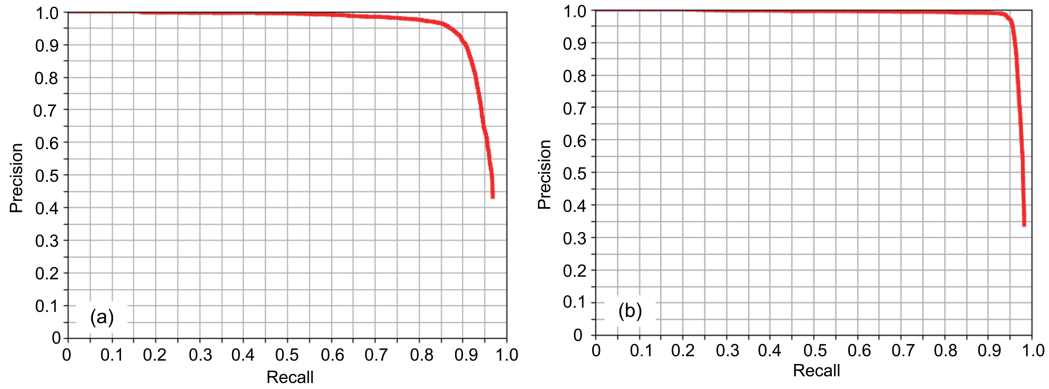

Fig. 8 Recall-precision curves: (a) baseline model (class: 3, AP = 0.94, iou-thres = 0.30); (b) the proposed model (class: 3, AP = 0.97, iou-thres = 0.30)

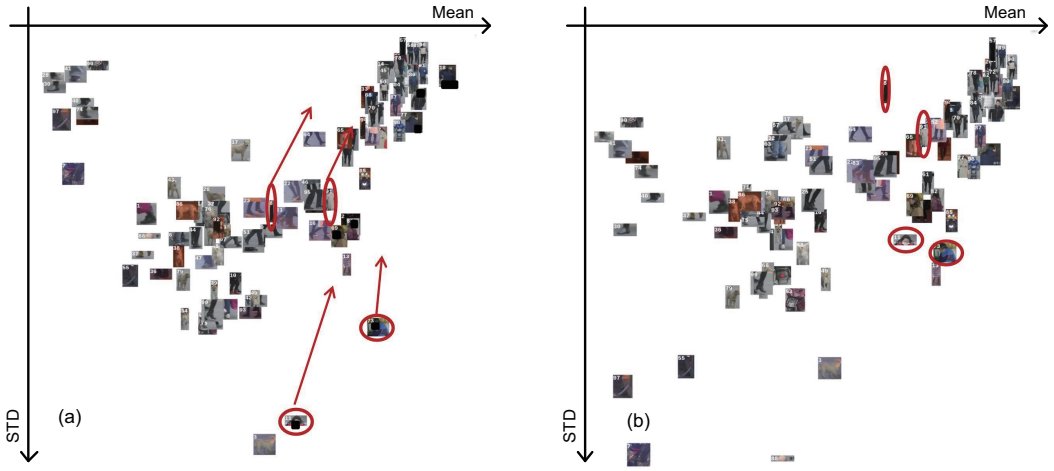

Fig. 9 Effect of graph-based group information propagation on the group distribution: (a) the distribution of proposals before G-GIP; (b) the distribution of proposals after G-GIP (References to the color refer to the online version of this figure)

Fig. 10 Examples of detection results on the Residential-K dataset

The left part of each image shows the detection results of the baseline model, and the right part shows the results of the proposed method

that our method can also improve the recognition of positive samples. Compared with the Recall-FPPI curves in Fig. 7, the curve generated by our method is closer to the left upper corner, which indicates that our method leads to fewer false positives. Compared with the Recall-precision curves in Fig. 8, the curve generated by our method is closer to the right upper corner, which indicates that our method achieves higher precision and recall. The left part of each subimage in Fig. 10 shows the results of the baseline model, and the right part shows the results of the proposed method. Comparing these image pairs, the results show that the proposed method can reduce false detection and missed detection, and improve the detection results.

Table 2 shows the detection results of the positive sample selection strategy based on a confidence grid search. The RF0.1 is improved while the AP declines, which indicates that sample selection can reduce false positives, but still does not meet expectations. The experimental results are further improved when simple similar group clustering is introduced into the proposed algorithm, which is shown in Table 3.

Table 1 Detection results of the baseline model and the proposed method

Model	RF0.1 (%)	AP (%)
YOLO-v2	90.76	94.10
G-GIP	96.21	97.13

Table 2 Confidence grid search for positive sample selection

Confidence	RF0.1 (%)	AP(%)
0.70	91.13	92.94
0.75	90.60	92.78
0.80	92.00	93.12
0.85	**92.37**	**93.60**
0.90	89.40	88.81
YOLO-v2	90.76	94.10

Best results are in bold

Table 3 Group with top-k confidences for positive sample selection (without G-GIP)

Top-k	RF0.1 (%)	AP (%)
Top-35	90.30	94.62
Top-40	**91.56**	**95.19**
Top-45	91.04	95.15
Top-50	91.54	95.31
YOLO-v2	90.76	94.10

Best results are in bold

Fig. 2a shows the independent object proposals of the new scene dataset. These object proposals contain positive and negative samples, which demonstrates the effectiveness of scene transfer. Fig. 2b shows the group object proposals of the new scene dataset. This shows that the object proposals can be divided into different groups by the simple clustering strategy. Comparing Tables 2 and 4, we find that the detection method can be improved by group information in the detection of the new scene. The bold values of Tables 2–4 are the best results.

Table 5 shows the stability of the graph model in different conditions. In this table, we compare the graph node classification results of the proposed method with and without the adjacency matrix modification subnet. We conducted multiple experiments with different numbers of training nodes. P_{m} is the mean number of positive nodes classified by the G-GIP model. P_{s} is the corresponding standard deviation. When the training model is repeated 50 times, the results show that the adjacency matrix modification subnet can generate smaller standard

Table 4 Results of the final detection

Node	RF0.1 (%)	AP (%)
4	96.13	97.15
6	96.00	97.03
8	96.11	96.65
10	96.26	97.02
12	**96.21**	**97.13**
16	95.78	96.87
24	95.96	97.37
32	95.72	97.21
50	95.56	97.20
80	94.45	96.40
YOLO-v2	90.76	94.10

Best results are in bold

Table 5 Node classification results with 50 G-GIP training repetitions

Node	With subnet		Without subnet	
	P_{m}	P_{s}	P_{m}	P_{s}
4	41.60	0.95	40.80	2.15
6	38.70	1.04	41.90	11.90
8	37.00	0	38.90	1.77
10	37.00	0	40.20	8.72
12	37.10	0.27	40.30	8.65
16	37.10	0.92	39.90	2.18
24	39.20	0.88	40.70	8.79
32	39.90	0.55	42.30	11.90
50	41.80	1.21	52.80	8.00
80	49.10	1.70	59.10	12.90

P_{m} is the mean number of positive nodes classified by the G-GIP model. P_{s} is the corresponding standard deviation

deviation and more stable graph node recognition results. Fig. 11 also demonstrates that the subnet can improve the graph node recognition results.

Comparing Tables 3 and 4, the results indicate that G-GIP can improve the group distribution for obtaining more effective positive samples. In Fig. 9, we can clearly observe the change of group distribution, i.e., the groups in red circles are moving to the top right corner.

To further demonstrate the scene adaptability and generality of the proposed method, all video surveillance datasets are used to test the proposed method and YOLO-v3, as the basic detection model is combined with the proposed method. The results are shown in Tables 6 and 7. The mean RF0.1 of YOLO-v3 is 6.95% lower than that of the proposed method, and the mean AP of YOLO-v3 is 2.95% lower than that of the proposed method.

For public datasets, the proposed method is compared with the state-of-the-art unsupervised domain adaptation object detection methods (He and Zhang, 2019; Khodabandeh et al., 2019; Shen et al., 2019; Zhu XG et al., 2019). Otherwise, most of these comparative methods are based on the VGG16-based faster R-CNN detection framework. Therefore, the VGG16-based faster R-CNN (without any other tricks: multi-scale, FPN, OHSM, focal loss, and so on) acts as the basic detection model in the proposed G-GIP(F). The results are shown in Table 8; the faster R-CNN and YOLO-v3 results are also shown, which detect a new scene through the model trained by the source scene. An example of faster R-CNN detection results is shown in Fig. 12. These results demonstrate that the proposed method

Table 8 AP of the KITTI-to-Cityscapes (K2C) and Sim10k-to-Cityscapes (S2C)

Model	AP (%)	
	K2C	S2C
Khodabandeh et al. (2019)'s	43.00	42.60
Zhu XG et al. (2019)'s	42.50	43.00
Shen et al. (2019)'s	41.90	42.60
He and Zhang (2019)'s	41.00	41.10
Faster R-CNN	35.76	33.10
G-GIP(F)	42.75	37.81
YOLO-v3	47.90	53.70
G-GIP(Y)	51.50	59.40

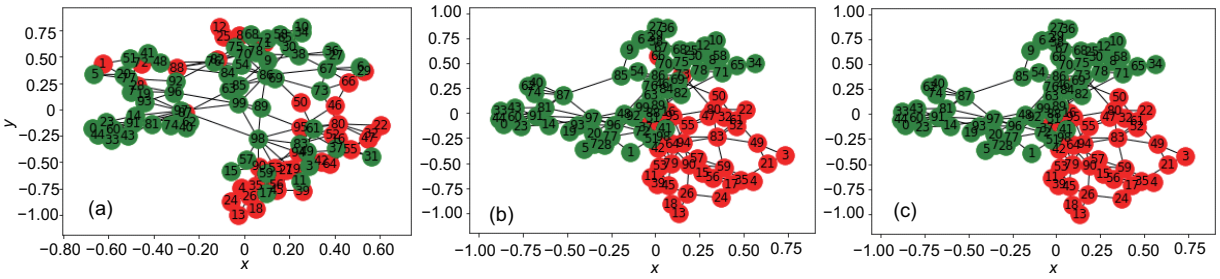

Fig. 11 Effect of graph-based group information propagation: (a) the proposed method without G-GIP; (b) G-GIP strategy without an adjacency matrix modification subnet; (c) final results

Red nodes are the positive groups and green nodes are opposite. References to the color refer to the online version of this figure

Table 6 RF0.1 of all video surveillance datasets

Model	RF0.1 (%)						Mean (%)
	J-road	S-path	S-gate	SW-gate	Residential-M	Residential-K	
YOLO-v3	68.08	63.48	40.36	77.67	74.59	94.93	69.85
G-GIP	73.64	72.17	50.07	82.63	85.64	96.46	76.80

Table 7 AP of all video surveillance datasets

Model	AP (%)						Mean (%)
	J-road	S-path	S-gate	SW-gate	Residential-M	Residential-K	
YOLO-v3	95.20	83.73	60.60	91.75	84.82	96.17	85.38
G-GIP	96.31	87.37	66.02	92.60	90.41	97.29	88.33

Fig. 12 Examples of detection results on K2C (a) and S2C (b)

The left parts of (a) and (b) show the detection results of the baseline model training on the source data, and the right parts of (a) and (b) show the results of the proposed method

can obtain competitive domain adaptation detection results without domain alignment, instance alignment, or class alignment. Moreover, the proposed method can obtain better results on the powerful baseline model.

In summary, experimental results show that the proposed method can autonomously mine positive samples in different scenes through the G-GIP strategy and detect objects in different scenes. Otherwise, the G-GIP strategy can be combined with different basic detection models. In this case, the proposed method can achieve adaptive evolution with changing scenes.

5 Conclusions

In this paper, we have proposed a scene-adaptive evolution unsupervised video object detection algorithm based on the object groups concept, which consists of a prototype dictionary generation strategy and a graph-based group information propagation strategy. The PDG has been proposed to build a dictionary of object prototypes to represent the visual knowledge of object groups. Then, the G-GIP strategy has been proposed to create the object prototype category for mining positive samples from the unlabeled new scene dataset. Finally, the new positive samples with pseudo labels act as the training data to fine-tune the detection model for detecting the new scene. The experimental results showed that the proposed method can adapt to and evolve with changing scenes, and demonstrated the effectiveness of the proposed method. Moreover, the proposed method can effectively reduce false positives.

Contributors

Shiliang PU, Di XIE, and Yunhe PAN designed the research. Wei ZHAO and Weijie CHEN conducted the experiments. Wei ZHAO drafted the manuscript. Shicai YANG and Di XIE helped organize the manuscript. Wei ZHAO and Shicai YANG revised and finalized the paper.

Compliance with ethics guidelines

Shiliang PU, Wei ZHAO, Weijie CHEN, Shicai YANG, Di XIE, and Yunhe PAN declare that they have no conflict of interest.

References

Chen MH, Kira Z, AlRegib G, et al., 2019. Temporal attentive alignment for large-scale video domain adaptation. Proc IEEE/CVF Int Conf on Computer Vision, p.6320-6329. https://doi.org/10.1109/ICCV.2019.00642

Cordts M, Omran M, Ramos S, et al., 2016. The cityscapes dataset for semantic urban scene understanding. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3213-3223. https://doi.org/10.1109/CVPR.2016.350

Croitoru I, Bogolin SV, Leordeanu M, 2017. Unsupervised learning from video to detect foreground objects in single images. Proc IEEE Int Conf on Computer Vision, p.4345-4353. https://doi.org/10.1109/ICCV.2017.465

Dai JF, Li Y, He KM, et al., 2016. R-FCN: object detection via region-based fully convolutional networks. Proc 30th Int Conf on Neural Information Processing Systems, p.379-387. https://doi.org/10.5555/3157096.3157139

Deng JJ, Pan YW, Yao T, et al., 2020. Single shot video object detector. *IEEE Trans Multim*, 23:846-858. https://doi.org/10.1109/TMM.2020.2990070

Feichtenhofer C, Pinz A, Zisserman A, 2017. Detect to track and track to detect. Proc IEEE Int Conf on Computer Vision, p.3057-3065. https://doi.org/10.1109/ICCV.2017.330

Geiger A, Lenz P, Urtasun R, 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3354-3361. https://doi.org/10.1109/CVPR.2012.6248074

Girshick R, 2015. Fast R-CNN. Proc IEEE Int Conf on Computer Vision, p.1440-1448. https://doi.org/10.1109/ICCV.2015.169

Girshick R, Donahue J, Darrell T, et al., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.580-587. https://doi.org/10.1109/CVPR.2014.81

Guo CX, Fan B, Gu J, et al., 2019. Progressive sparse local attention for video object detection. Proc IEEE/CVF Int Conf on Computer Vision, p.3908-3917. https://doi.org/10.1109/ICCV.2019.00401

Han W, Khorrami P, Le Paine T, et al., 2016. Seq-NMS for video object detection. https://arxiv.org/abs/1602.08465v1

He ZW, Zhang L, 2019. Multi-adversarial faster-RCNN for unrestricted object detection. Proc IEEE/CVF Int Conf on Computer Vision, p.6667-6676. https://doi.org/10.1109/ICCV.2019.00677

Htike KK, Hogg DC, 2014. Efficient non-iterative domain adaptation of pedestrian detectors to video scenes. Proc 22nd Int Conf on Pattern Recognition, p.654-659. https://doi.org/10.1109/ICPR.2014.123

Johnson-Roberson M, Barto C, Mehta R, et al., 2016. Driving in the matrix: can virtual worlds replace human-generated annotations for real world tasks? Proc IEEE Int Conf on Robotics and Automation, p.746-753. https://doi.org/10.1109/ICRA.2017.7989092

Kang K, Ouyang WL, Li HS, et al., 2016. Object detection from video tubelets with convolutional neural networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.817-825. https://doi.org/10.1109/CVPR.2016.95

Kang K, Li HS, Xiao T, et al., 2017. Object detection in videos with tubelet proposal networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.889-897. https://doi.org/10.1109/CVPR.2017.101

Kang K, Li HS, Yan JJ, et al., 2018. T-CNN: tubelets with convolutional neural networks for object detection from videos. *IEEE Trans Circ Syst Video Technol*, 28(10):2896-2907. https://doi.org/10.1109/TCSVT.2017.2736553

Khodabandeh M, Vahdat A, Ranjbar M, et al., 2019. A robust learning approach to domain adaptive object detection. Proc IEEE/CVF Int Conf on Computer Vision, p.480-490. https://doi.org/10.1109/ICCV.2019.00057

Kipf TN, Welling M, 2017. Semi-supervised classification with graph convolutional networks. https://arxiv.org/abs/1609.02907

Kwak S, Cho M, Laptev I, et al., 2015. Unsupervised object discovery and tracking in video collections. Proc IEEE Int Conf on Computer Vision, p.3173-3181. https://doi.org/10.1109/ICCV.2015.363

Lahiri A, Ragireddy SC, Biswas P, et al., 2019. Unsupervised adversarial visual level domain adaptation for learning video object detectors from images. Proc IEEE Winter Conf on Applications of Computer Vision, p.1807-1815. https://doi.org/10.1109/WACV.2019.00197

Law H, Deng J, 2018. CornerNet: detecting objects as paired keypoints. Proc 15th European Conf on Computer Vision, p.765-781. https://doi.org/10.1007/978-3-030-01264-9_45

Li D, Hung WC, Huang JB, et al., 2016. Unsupervised visual representation learning by graph-based consistent constraints. Proc 14th European Conf on Computer Vision, p.678-694. https://doi.org/10.1007/978-3-319-46493-0_41

Li JN, Liang XD, Shen SM, et al., 2018. Scale-aware fast R-CNN for pedestrian detection. *IEEE Trans Multim*, 20(4):985-996. https://doi.org/10.1109/TMM.2017.2759508

Li NJ, Chang FL, Liu CS, 2020. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. *IEEE Trans Multim*, 23:203-215. https://doi.org/10.1109/TMM.2020.2984093

Lin TY, Dollár P, Girshick R, et al., 2017a. Feature pyramid networks for object detection. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.936-944. https://doi.org/10.1109/CVPR.2017.106

Lin TY, Goyal P, Girshick R, et al., 2017b. Focal loss for dense object detection. Proc IEEE Int Conf on Computer Vision, p.2999-3007. https://doi.org/10.1109/ICCV.2017.324

Liu W, Anguelov D, Erhan D, et al., 2016. SSD: single shot multibox detector. Proc 14th European Conf on Computer Vision, p.21-37. https://doi.org/10.1007/978-3-319-46448-0_2

Ma XL, Zhu XT, Gong SG, et al., 2017. Person re-identification by unsupervised video matching. *Patt Recogn*, 65:197-210. https://doi.org/10.1016/j.patcog.2016.11.018

Pan YH, 2016. Heading toward artificial intelligence 2.0. *Engineering*, 2(4):409-413. https://doi.org/10.1016/J.ENG.2016.04.018

Pan YH, 2019. On visual knowledge. *Front Inform Technol Electron Eng*, 20(8):1021-1025. https://doi.org/10.1631/FITEE.1910001

Pan YH, 2020. Miniaturized five fundamental issues about visual knowledge. *Front Inform Technol Electron Eng*, online. https://doi.org/10.1631/FITEE.2040000

Pang JM, Chen K, Shi JP, et al., 2019. Libra R-CNN: towards balanced learning for object detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.821-830. https://doi.org/10.1109/CVPR.2019.00091

Redmon J, Farhadi A, 2017. YOLO9000: better, faster, stronger. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6517-6525. https://doi.org/10.1109/CVPR.2017.690

Redmon J, Divvala S, Girshick R, et al., 2016. You only look once: unified, real-time object detection. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.779-788. https://doi.org/10.1109/CVPR.2016.91

Ren SQ, He KM, Girshick R, et al., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. Proc 28th Int Conf on Neural Information Processing Systems, p.91-99. https://doi.org/10.5555/2969239.2969250

Shen ZQ, Maheshwari H, Yao WC, et al., 2019. SCL: towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. https://arxiv.org/abs/1911.02559

Shvets M, Liu W, Berg A, 2019. Leveraging long-range temporal relationships between proposals for video object detection. Proc IEEE/CVF Int Conf on Computer Vision, p.9755-9763. https://doi.org/10.1109/ICCV.2019.00985

Subramaniam A, Nambiar A, Mittal A, 2019. Co-segmentation inspired attention networks for video-based person re-identification. Proc IEEE/CVF Int Conf on Computer Vision, p.562-572. https://doi.org/10.1109/ICCV.2019.00065

Tang K, Ramanathan V, Li FF, et al., 2012. Shifting weights: adapting object detectors from image to video. Proc 25th Int Conf on Neural Information Processing Systems, p.638-646. https://doi.org/10.5555/2999134.2999206

Veličković P, Casanova A, Lio P, et al., 2018. Graph attention networks. https://arxiv.org/abs/1710.10903

Wang HW, Leskovec J, 2019. Unifying graph convolutional neural networks and label propagation. https://arxiv.org/abs/2002.06755

Wang SG, Cheng J, Liu HJ, et al., 2018. Pedestrian detection via body part semantic and contextual information with DNN. *IEEE Trans Multim*, 20(11):3148-3159. https://doi.org/10.1109/TMM.2018.2829602

Wang SY, Zhou YC, Yan JJ, et al., 2018. Fully motion-aware network for video object detection. Proc 15th European Conf on Computer Vision, p.557-573. https://doi.org/10.1007/978-3-030-01261-8_33

Wang SY, Group A, Lu HC, et al., 2019. Fast object detection in compressed video. Proc IEEE/CVF Int Conf on Computer Vision, p.7103-7112. https://doi.org/10.1109/ICCV.2019.00720

Wu F, Souza A, Zhang TY, et al., 2019. Simplifying graph convolutional networks. Proc 36th Int Conf on Machine Learning, p.6861-6871.

Xiao FY, Lee YJ, 2016. Track and segment: an iterative unsupervised approach for video object proposals. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.933-942. https://doi.org/10.1109/CVPR.2016.107

Xiao FY, Lee YJ, 2018. Video object detection with an aligned spatial-temporal memory. Proc 15th European Conf on Computer Vision, p.494-510. https://doi.org/10.1007/978-3-030-01237-3_30

Yu HK, Guo DZ, Yan ZP, et al., 2018. Unsupervised learning for large-scale fiber detection and tracking in microscopic material images. https://arxiv.org/abs/1805.10256

Zhang XS, Wan F, Liu C, et al., 2019. FreeAnchor: learning to match anchors for visual object detection. https://arxiv.org/abs/1909.02466

Zhu ML, Liu M, 2018. Mobile video object detection with temporally-aware feature maps. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5686-5695. https://doi.org/10.1109/CVPR.2018.00596

Zhu XG, Pang JM, Yang CY, et al., 2019. Adapting object detectors via selective cross-domain alignment. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.687-696. https://doi.org/10.1109/CVPR.2019.00078

Zhu XZ, Wang YJ, Dai JF, et al., 2017. Flow-guided feature aggregation for video object detection. Proc IEEE Int Conf on Computer Vision, p.408-417. https://doi.org/10.1109/ICCV.2017.52