Survey of Video Based Small Target Detection

Ying Liu[1,2], Luyao Geng[1], Weidong Zhang[1,2], and Yanchao Gong[1,2]
[1] Center for Image and Information Processing, Xi'an University of Posts and Telecommunications, Xi'an, China
[2] Key Laboratory of Electronic Information Application Technology for Scene Investigation, Ministry of Public Security, Xi'an, China
Email: {liuying_ciip, gengluyao_625}@163.com, chluzhre@126.com, gongyanchao@xupt.edu.cn

Zhijie Xu
School of Computing and Engineering, University of Huddersfield, Queens Gate, Huddersfield, UK
Email: z.xu@hud.ac.uk

***Abstract*—Video based target detection is an important research content in intelligent video surveillance, which extracts foreground objects from background images in video sequences. Video based target detection has developed rapidly in recent years. In practical applications, however, detection of small and medium-sized objects in video remains a challenging task as small and medium-sized objects occupy too few pixels, and the obtained information is limited. The demands in aerospace, criminal investigation, face recognition, intelligent transportation and other fields have proved the research value of video based small target detection. This paper first briefly introduces the traditional video based target detection algorithms and the improvements for small target detection. Second, deep learning based models for small target detection in video are summarized in detail, which are categorized into one-stage models and two-stage models according to the detection stages. The network structures and plug-in modules for video based small target detection are also explained. In addition, this paper summarizes the common databases with evaluation criteria. Finally, applications and future research direction in this area are analyzed.**

***Index Terms*—video based small target detection, deep learning, one-stage method, two-stage method**

I. Introduction

As an important research content in the field of computer vision, target detection in video always receives wide attention from scholars. Video based target detection is challenging when the target is too small, while the demand for detecting small targets in video is increasing. It is an important approach to obtain information by processing the videos collected by satellites and Unmanned Aerial Vehicles (UAVs) in aero-space field. With the techniques of video based small target detection, the small targets can be captured in time, which can be applied for military reconnaissance and maintenance of national security. Video based small target detection is widely used in various scenarios. It can be employed for small traffic sign recognition in intelligent transportation; In the field of medicine is able to assist doctors in disease screening. In the field of criminal investigation, discovering small key targets of suspects such as clothing decorations and vehicle accessories from video surveillance in time is crucial for criminal case detection.

As shown in Fig. 1, small targets refer to those smaller than 32×32 pixels in image, or targets smaller than 10% of the image size [1]. This means that the information from pixels is limited, and the texture features normally used for image detection is inapplicable [2].

(a)20*28 Absolute small target (b)A relatively small target with a size smaller than 0.1 times the original image

Figure 1. Small target examples.

Video based target detection techniques are to detect moving targets in video sequences. The mainstream video based target detection techniques are developing rapidly, but research on video based small target detection is rare. This field is still in the early stage of development, and systematic and comprehensive survey on video based small target detection is absent.

Before 2012, video based target detection relied on traditional algorithms, including Lucas-Kanade optical flow [3], frame difference [4] and background difference [5], [6]. The Lucas-Kanade optical flow based approach models the image pixels of adjacent frames in the video and forms a vector field based on the intensity changes between time and to obtain the information of the target movement. The frame difference based approach raised [4]. Is to gain information of moving targets through the margin calculation on adjacent frames. The general idea of the background difference based approach suggested by I. Haritaoglu [5] and Y. Ivanov [6] is to obtain the foreground target by subtracting the temporal frame with the pre-stored or real-time-captured background images.

Manuscript received March 26, 2021; revised July 12, 2021.

doi: 10.18178/joig.9.4.122-134

Convolutional Neural Networks (CNN) [7] have been widely used in target detection since 2012. Target detection techniques based on deep learning developed dramatically and received great progress in moving target detection and image target detection. Different from traditional methods that exploit inter-frame relevance, deep learning based small target detection methods in video is able to process the key frames directly. This paper categorizes the existing deep learning based algorithms of small target detection in video into two types in terms of the algorithm procedure: two-stage small target detection and single-stage small target detection. The two-stage model first generates candidate regions and then employs CNN for classification, which is by now the most accurate target detection algorithm. The single-stage model however directly performs regression on images, which is faster than the former one and is suitable for real-time target detection. This paper introduces Region-CNN (R-CNN) series algorithms [8] as the representative algorithms of the two-stage models. You Only Look Once (YOLO) [9] and Single Shot multibox Detector (SSD) [10] series methods are introduced as the representative algorithms of the single-stage models.

The scholars conducted optimization on the network structure as well. He [2] proposed Residual Neural network (ResNet) which eliminates the impact of gradient explosion and gradient disappearance to a certain extent through jump connections under the premise of ensuring the depth of the network. The hourglasses network proposed by Newell A [11] determines the localization of the target by predicting the key points without change of data resolution, which shows advantage in detection of small targets. Li Z [12] specifically designed a backbone network DetNet-59 for target detection, and the embedded Feature Pyramid Networks (FPN) module [13] can construct high-level semantic feature maps of multiple scales.

Besides, there are also plug-in modules similar to FPN, which can be inserted into various models to enhance scale invariance of the models. For instance, SPP-net [14] breaks the limit of fixed input size of CNN networks. The Extended Feature Pyramid Network (EFPN) with FTT module embedded in the FPN framework is dedicated to small target detection.

The structure of this paper is shown as below: Section II introduces the principles of traditional video based small target detection methods with traditional small target detection algorithms, and compares their advantages and disadvantages. Then deep network structures applicable for target detection are discussed, and two types of deep learning based small target detection methods in video and three kinds of additional modules for improving small-sized target detection performance are introduced. Section III illustrates the video datasets for small target detection. Section IV gives a prospect of applications and future development of video based small target detection and conclude in Section V.

II. Development of Video Based Small Target Detection

The approaches for video based small target detection can be categorized as traditional method based approaches and deep learning based approaches.

There are three kinds of traditional small target detection methods, which are optical flow, frame difference and background difference separately. Most of the video based small target detection algorithms are extensions of those three mainstream algorithms. These algorithms have experienced continuous modification in the past decades and obtained considerable improvements, they still lack robustness to the objects with extremely small sizes in video.

CNN won the first place in the ImageNet challenge [7] in 2012, making CNN and deep learning a hot research topic and more widely applied in computer vision. It shows significant advantage in improving the real-time and accuracy of video based small target detection.

A. Traditional Algorithms Based Small Target Detection in Video

1) Optical flow based approaches

Lucas-Kanade optical flow (LK) [3] was first proposed as a kind of image mosaic algorithm, which by global computation on two adjacent frames, the vector field that describes the variation of the image pixel intensity is able to manifest the variation and movement of image intensity over time, with a shortcoming of the huge calculation amount. Afterwards, scholars adopted integrated SIFT features [15], GPU parallelization [16], additional characteristic corners as tracking points [17], [18] and sparse subset of samples [19] to reduce computational complexity while improve robustness and accuracy.

There are also improvements for small target detection based on traditional optical flow methods. Liu [20] proposed a modified LK method based on Hough transform for small ball tracking, which takes the target center as the tracking corner point to improve tracking accuracy. Yuan [21] proposed a matching based approach that selects candidate pixels. It can handle small target detection with the size of 5-10 pixels by employing the motion features of small targets and combining inter-frame hierarchically processing with Gaussian pyramid denoising model.

2) Frame difference based approaches

The frame different algorithm [4] is a widely used small target detection algorithm in video, which detects the motion according to the inter-frame pixel intensity variation. Then the obtained image is segmented with threshold and the moving targets are extracted. The frame difference based approaches are simple in principle, fast, and easy to implement on hardware.

The frame difference based approaches might lose the targets while detecting small targets, hence jointly detection on multiple frames is a common solution [22]. For the problem, Sun [23] proposed a small moving target detection algorithm based on the three-frame difference theory, which integrates three adjacent frames

to improve the inter-frame relevance of the target, and then excludes other moving targets by applying color threshold. The feasibility of this algorithm has been proved with experiments.

3) Background difference based approaches

The background difference based approaches first store the background image and use the foreground image to subtract the background image [24], [25]. Normally, due to the distinctive difference between the intensity of background and the moving targets, a difference image is able to retain a larger intensity at the location of moving objects, and the foreground objects can be distinguished from background with appropriate threshold. The difficulty in such methods lies in the acquisition and updating of the background. Commonly used methods include manual acquisition of background and statistics of grayscale values.

4) Limitation and advantages & disadvantages of traditional algorithms

The optical flow based methods have the advantage in detecting the key pixels of small targets, but there are flaws such as sensitivity to preprocessing, complexity in calculating the optical flow field, and high requirements for hardware facilities. Therefore, there are still many problems in real-time monitoring and automatic tracking of targets by optical flow [5]. The frame difference based methods detect the targets by the difference between adjacent frames, there are mainly two approaches for improving detection of small targets based on frame difference: increasing the number of adjacent frames [22] [23] and incorporating other algorithms [26]. The simple frame difference based method is only applicable for videos with a static background, such as in surveillance videos. The multi-frame difference based method is also empirical in threshold setting [27]. The background difference based methods require the difference between the key frame and background image, however, as shown in Table I, even the surveillance video sequences with a fixed background are very sensitive to environmental changes in practice. Real-time update of the background also increases the complexity of the algorithm [28].

The traditional video based small target detection algorithms improve the accuracy for small targets in terms of target scale invariance. The optimization of the algorithms lasts half a century and is close to maturity. However, traditional algorithms have insurmountable limitations, and the space for improvement is limited. While the target detection algorithms based on deep learning have developed rapidly and can improve the detection accuracy of small targets [29].

B. Deep Learning Based Small Target Detection

Taking surveillance video as an example, real-time surveillance by manual work is costly and inefficient, and the precision of traditional video based small target detection methods cannot meet the requirements. In recent years, video surveillance is developing towards intelligence. It is in urgent demand to replace human with computers for video surveillance. Meanwhile, deep learning based video detection algorithms have shown incomparable superiority against traditional algorithms, which is helpful for detecting the extremely small targets to some extent.

The biggest difference between deep learning based methods and traditional methods for video based small target detection is the localization of the target, i.e., the traditional algorithms detect the location information of the small target through the target association between frames, while deep learning based methods directly operate on the key frames by producing bounding boxes around the targets on the key frames for detection.

This section first introduces the applicable network structure based on deep learning for small target detection. Then the deep learning based algorithms for small target detection is introduced together with the superior plug-in modules for improving robustness to scale. This paper introduces two types of methods according to the stages of the deep learning based small target detection algorithms, i.e., one-stage approach and two-stage approach.

TABLE I. COMPARISON BETWEEN ADVANTAGE & DISADVANTAGE OF THE TRADITIONAL ALGORITHMS

Method	Algorithm complexity	Advantage	Disadvantage
Optical flow	high	Suitable for dynamic and static backgrounds without too many pixels, fully utilizing the underlying features of small targets	High computational complexity, sensitive to light and noise
Frame difference	low	Simple and easy to implement	Losing small targets, sensitive to interference
Background difference	medium	Real-time implementation, immune to the hole effect of frame difference based methods	Difficult in complicated background modelling, limited mobility of equipment

One-stage model: There is no candidate boxes and the regression is directly applied to the input images with a CNN to directly obtain the localization and category information of targets. Hence, the methods are referred as regression based target detection algorithms as well. These methods are usually fast. The representative algorithms are YOLO series algorithms and SSD series algorithms, etc.

Two-stage model: These methods are also referred as classification based target detection algorithms. The procedure of these methods includes: Dividing the images into 2000 candidate boxes to generate ‘region proposals’; Applying CNN to classify the candidate regions based on features that represent color, texture, size and shape.

These methods are also known as region based target detection as the localization of the targets are obtained by

selecting the candidate boxes. These methods are of high precision. The representative algorithms are R-CNN series algorithms.

1) Network structures for small target detection

To handle targets that are extremely small, it is necessary to capture as much information from limited pixels as possible. Increasing the depth of the network can obtain more information and richer features. However, the side effect of very deep networks is that the improvement of the inter-layer learning rate will be greater than the inter-layer information delivery, and this may cause gradient explosion and vanishing gradient. The problem has greater impact on small targets. ResNet and Hourglasses networks can solve this problem to a certain extent.

ResNet: ResNet [2] uses short-circuit mechanism as the residual unit. With the shortcut connections as shown in Fig. 2, the deep network does not need to learn the entire output, but copies the residual of the shallow network, i.e., identity mapping, and learns the residuals of the previous output.

The integrity of the information is protected, so that the entire network only needs to learn the difference between input and output and the mapping is more sensitive to changes in output with residuals.

Hourglasses network: Hourglasses network [11] was originally designed to capture the spatial localization information of human skeletons for human pose estimation. The network structure is similar to an hourglass as showed in Fig. 3. The multi-layer residual modules extract high-level features and meanwhile retains the information of the original levels. Without change of the data size, it only alters the depth of data and repeatedly executes bottom-up and top-down inference. Because of the branches at each resolution, it can capture information of each scale and obtain localization information based on key points.

Darknet59 network: Although there are more and more CNN- based target detectors, they are modified from classification networks. Considering that the designing principle of image classification is not beneficial to localization tasks, Li Z [12] designed Darknet59 specifically for target detection. Darknet59 is a specialized target detection network. It retains the first four layers of ResNet-50 and modifies the last layer. A FPN module is added to deal with targets of different scale, which maintains high resolution as well as large receptive field. It obtained better results than ResNet-50 and ResNet-101.

LSTS: The existing video based target detection is to transmit the detectors from images to videos, which is a difficult task. The quality of frames will also be reduced due to occlusion, pose and motion blur. Furthermore, assembling features from adjacent frames to improve precision greatly increase the computational complexity.

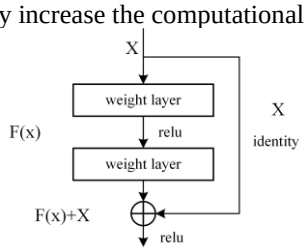

Figure 2. Illustration of shortcut connections [2].

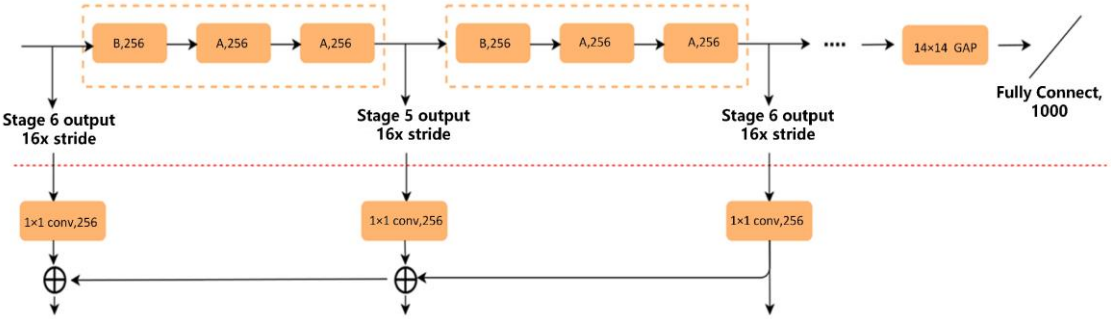

Figure 3. Feature pyramid network based on DetNet [12].

Learnable Spatio-Temporal Sampling (LSTS) [30] was proposed to precisely learn the semantic level correspondence relations between the features of frames. The LSTS module first randomly initializes the sampling position, and then iteratively updates it to gradually find a better spatial correspondence guided by the supervision from detection. In addition, as shown in Fig. 4, the Sparsely Recursive Feature Updating (SRFU) module and the Dense Feature Aggregation (DFA) module were introduced to model the temporal relationship and enhance the features of each frame separately. Experiments proved that the framework can achieve optimal performance with real-time speed and low computational complexity.

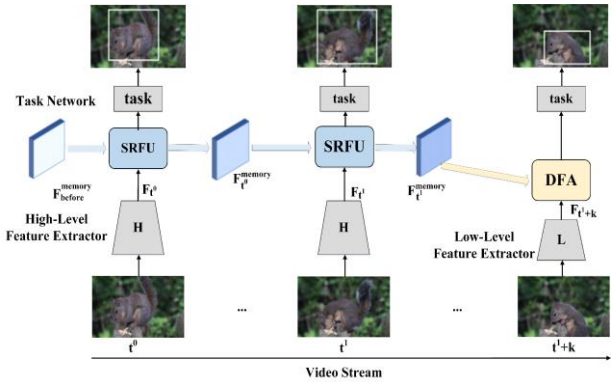

Figure 4. Framework of LSTS [30].

2) One-stage methods for small target detection

YOLO Series Methods: The YOLO algorithm is able to obtain the localization information with one-off regression, which has advantage in speed but loses accuracy due to the absence of candidate boxes. YOLOv2 [9] adopts the Darknet-19 structure for feature extraction, which greatly improve the speed. However, the precision improvements for small target detection are limited as the structure is simple. YOLOv3 based on YOLOv2 [31], applies ResNet [2] to increase the depth of feature extraction layers. The structure of Darknet-53 is applied for feature extraction, which will lose the localization information of the features in the shallow layers, and the speed is slower.

In 2019, M. Ju [32] combined the down-sampled feature maps from the output of YOLOv3 with Darkent-53, and formed a feature fusion layer with 4-times down-sampled output for target detection. Meanwhile, two residual units are added to obtain more features of small targets. In 2020, A. Bochkovskiy [33] presented YOLOv4, on the basis of YOLOv3, YOLOv4 adopts CSPDarknet_[34] with SPP-net module to ensure the speed as well as precision, and hence it is a desirable approach for small target detection.

SSD algorithm: Single Shot multibox Detector (SSD) [35] is an end-to-end target detection algorithm, and its highlight lies in the multiscale feature maps. CNN is employed to extract features of input image, and multi-scale feature maps are produced. As shown in Fig. 5, the pixels of the feature maps and the 8732 prior bounding boxes are matched and the feature maps are converted by the convolutional layers to output the best bounding box prediction. SDD has the advantages of high accuracy as R-CNN and high speed as YOLO, which enables real-time detection with high accuracy and has fine detection performance for targets with different sizes. Based on SSD, Deconvolutional Single Shot Detector (DSSD) [36] instead brought in residual blocks and used ResNet 101, which showed better performance for small target detections.

RetinaNet: Lin [37] studied the reasons that the detection accuracy of single-stage methods is lower than multi-stage methods. The authors proposed RetinaNet, which used focused loss function to replace the traditional cross entropy loss. This modification reduces the weights of background samples which are relatively simple, so that the model focuses on more difficult target samples in the learning process. The methods are effective for small target detection, and the accuracy exceeds all previous two-stage detectors.

EfficientDet: EfficientDet [38] employed EfficientNet as the backbone. With the key idea to optimize the feature pyramid, EfficientDet proposed Bi-directional Feature Pyramid Network (BiFPN) which extracts features from the 3-7 layers in the backbone network to rapidly and repeatedly proceed the multi-scale bidirectional feature fusion of bottom-up top-down inference as shown in Fig. 6.

Meanwhile, a composite resizing method was proposed to jointly resize the resolution, depth and width of all the backbone networks, feature networks and prediction networks simultaneously. Experimental results prove that EfficientDet-D7 obtained the highest precision among the one-stage models with 51.0 mAP in the COCO dataset.

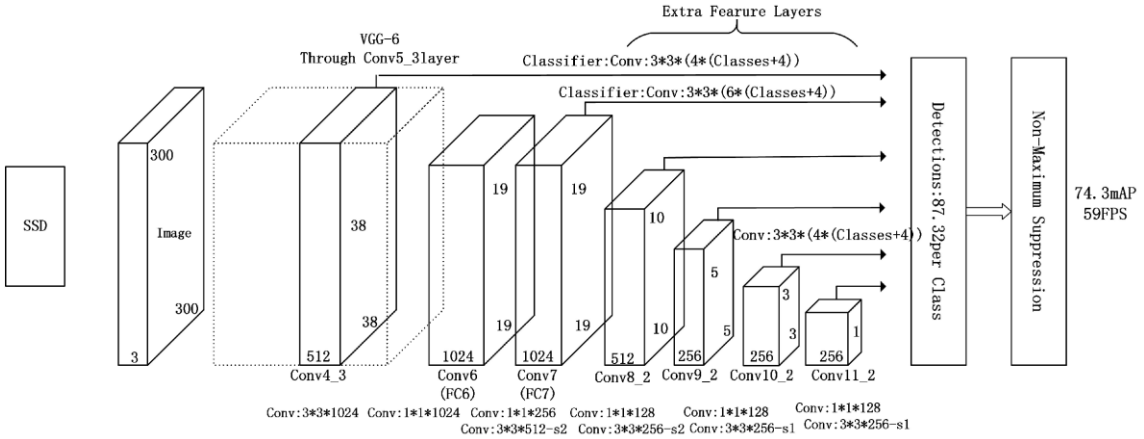

Figure 5. Network structure of SSD300 [35].

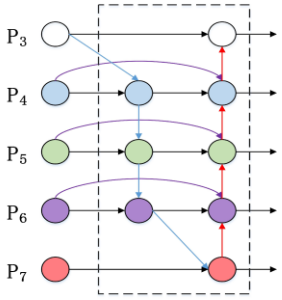

Figure 6. The structure of BiFPN [38].

FCOS algorithm: Anchor-based algorithms are widely used, but the performance of anchor-based detection is very sensitive to the size, aspect ratios and number of anchor boxes. It is difficult to detect targets with large scale changes, especially small targets, and the predefined anchor box will also hinder the generalization ability of the detector. In addition, it also involves complex IOU frame calculations. It is not only difficult to set these hyper-parameters, but also are often sensitive to the final detection performance; in view of the shortcomings of the above-mentioned Anchor-based target detection. Z. Tian

[39] proposed an anchor-free Fully Convolutional One-Stage object detection algorithm (FCOS) based on Fully Convolutional Networks (FCN) [40], which directly predicts each target point. This method is more suitable for small targets with fewer pixels. The usage of RetinaNet as the backbone and multi-layer prediction of the FPN module has a high recall rate, which eliminates the problem of overlapping bounding boxes caused by anchor-free detection to a certain extent. At the same time, the single-layer branch, i.e., center-ness layer is added to calculate the centrality. The detection accuracy is comparable to that of the two-stage algorithms in the COCO dataset.

3) Two-stage methods for small target detection

R-CNN series methods: R. Girshick proposed region with CNN [8] (R-CNN) in 2014, which employs selective search to generate about 2000 candidate regions. The candidate regions are fed to the CNN for training, and then classification and bounding box regression are performed for target detection. This is the foundation of the R-CNN methods [41]. One year later, the authors proposed Fast-R-CNN [42] to refine the classification. First, the input image is fed to a CNN, and the candidate regions are produced via selective search. Different from the R-CNN that warps the input images for size normalization, Fast-R-CNN maps information from the candidate bounding boxes to the last layer of feature maps through ROI Pooling.

Based on the above methods, He *et al.* proposed to incorporate Region Proposal Network (RPN) and presented Faster-R-CNN [43] in 2015. In the training stage, feature extraction is first employed with CNN, and then selective searching is adopted to obtain the candidate bounding boxes. This greatly improves efficiency as there is no need to feed each candidate region to CNN. The next year, He *et al.* replaced the backbone from VGG-16 to ResNet-101 and proposed Faster-R-CNN+++ [2], which simplifies the training procedure of the deep network and the performance is improved.

SNIP algorithm: B. Singh [44] took COCO dataset as example and indicated that the difficulty in target detection is from the various scale of the targets. Considering the poor performance in detecting small targets, the authors proposed Scale Normalization for Image Pyramids (SNIP) and improved detection performance for small targets in experiments.

SNIP excludes the impact of the targets with extremely small sizes. In the pretraining process, the feedback of gradient for targets with different sizes is restricted, which produces gradient feedback for targets within a specified range of size. As shown in Fig. 7, the purple bounding box is beyond the candidate region with the specified range. Experimental results prove that the method is impressively effective for small target detection. In addition, B. Singh [45] proposed an improved algorithm SNIPER based on SNIP, which does not rely on high-resolution images, but generates fixed-size chips based on proportions. The improved SNIPER algorithm performs better in practical applications.

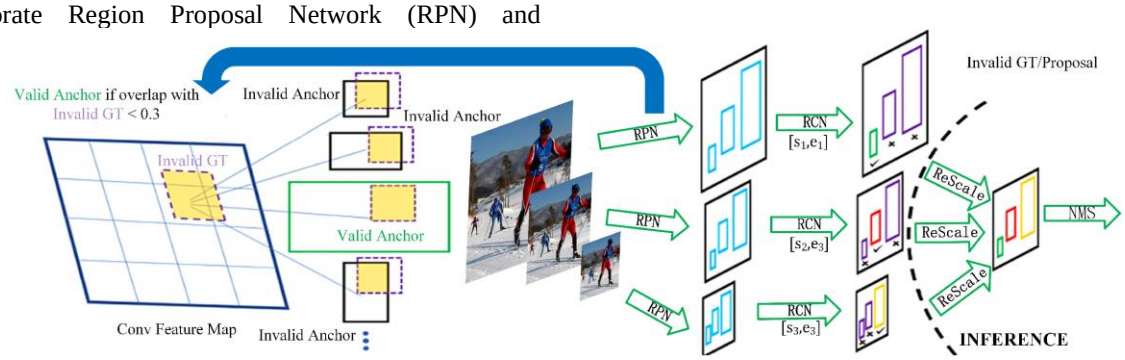

Figure 7. The inference procedure of the SNIP algorithm [44].

Mask R-CNN: He K [46] proposed Mask R-CNN following the framework of Faster R-CNN. The idea is to incorporate ResNet with FPN, and mask prediction branches are introduced to enhance the semantic information and spatial information simultaneously. The detection precision is greatly improved in multi-scale detection, especially in small target detection.

Despite the rapid development of the two-stage methods, these methods tend to focus on the improvement of precision rather than speed. The demand in specific fields that require real-time performance like video based small target detection cannot be satisfied. Therefore, regression based one-stage methods for small target detection are presented.

4) Other algorithms

Apart from the two-stage and one-stage small target detection methods, some excellent methods that do not belong to the above two types of methods have appeared in recent years.

Cascade R-CNN: For the threshold selection of the two-stage models, Z. Cai [47], [48] proposed a multi-level target detection model Cascade R-CNN. It can constantly enhance the IOU threshold to achieve a balance of the sample quantity and quality on the premise that the number of positive samples is guaranteed. This cascade approach achieves high precision in small target detection.

DETR: The existing target detection methods make prediction indirectly by performing regression on a large number of bounding box proposals. However, the predicted target localization point set and the post-processing algorithm have a great influence on the prediction of the target. In 2020, Detection Transformer (DETR) algorithm proposed by N. Carion [49], bypasses

the bounding box proposals and recognizes the problem of target detection as directly predicting the bounding box set in an end-to-end manner. As shown in Fig. 8, DETR is the combination of a CNN and a transformer framework with a group-based overall loss function. The loss function enforces unique prediction through binary matching and the transformer with encoder-decoder systematic structure, and the model directly outputs the final prediction set in parallel. Without any customized layer, it can be duplicated easily in any framework including standard CNN and transformer classes. DETR is simpler than the one-stage methods, meanwhile the prediction speed and comparable precision to the two-stage methods like Faster RCNN-R101-FPN can be satisfied.

DETR performs better on large-scale targets than small-sized targets. The designer also pointed out that strengthening the size robustness is the development direction for DETR in the future.

5) Plug-in modules

Due to the strong transferability, plug-in modules can significantly enhance the precision of object detection with a little additional inference cost [33]. Generally speaking, these plugin modules are for enhancing certain attributes in a model, such as strengthening feature integration capability, or enlarging receptive field. This paper introduces three types of plug-in modules including SPP-net, FPN and EFPN that can expand the receptive field and feature enhancement of the model.

SPP-net: The occurrence of SPP-net [14] solved the problem in CNN that the size of input image is fixed (e.g., 224×224). SPP-net feeds the feature maps from CNN to the Spatial Pyramid Pooling (SPP) layer, and then gathers features from arbitrary region to produce a feature vector of fixed length for training the detectors. This method is able to improve all the CNN based target detection methods.

FPN: Feature Pyramid Network (FPN) [13] can be implemented on various models, which takes advantage of the inherent multi-scale pyramid hierarchical structure of deep CNNs to proceed feature enhancement by fusing features of different levels. Except for the lateral connections, FPN also integrates top-down pathway connection as shown in Fig. 9. The targets become smaller along with the corresponding features from the up-down mapping in the pyramid structure, which is applicable to construct high-level semantic feature maps in various scales. By combining FPN with Faster R-CNN, the precision of Faster R-CNN-FPN is improved with 8.0 percentage points.

EFPN: In 2020, C. Deng [10] proposed Extended Feature Pyramid Network (EFPN) with an extra high-resolution pyramid level, which was specifically designed for small target detection. The Feature Texture Transfer (FTT) module is embedded in the FPN framework to perform super-resolution and minutiae detection in confidence regions. The levels in the pyramid are expanded to capture more local details. Furthermore, it designed foreground-background-balance loss fusion to relieve the background interference, which obtained impressive accuracy on TT100K (small target dataset) and MS-COCO (general dataset).

6) Analyses and performance comparison

This section compares the existing deep learning based small target detection methods as shown in Table II. The advantages and disadvantages for each method during modification are analyzed and the performances are also compared. The COCO dataset with variable-scale and small-sized target [1] is selected as benchmark, and the evaluation metrics are the average precision (AP) and APs, which is specific for small targets with fewer than 32×32 pixels. The results show that the overall precision of the two-stage methods is higher than the one-stage methods, with the cost of loss in speed. Among the two-stage methods, the performance of SNIP and its extension SNIPER [45] are the best for small target detection. Among the one-stage methods, YOLOv4 shows the best performance.

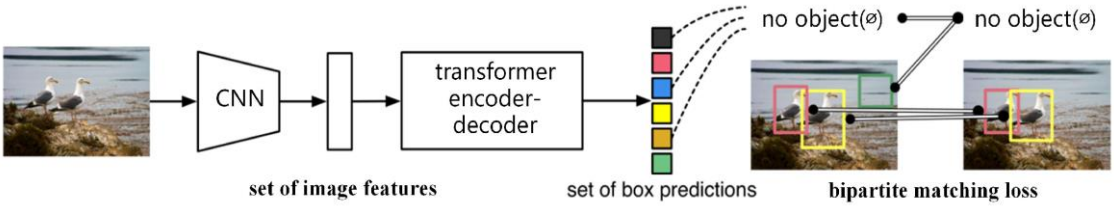

Figure 8. The network structure of DETR [49].

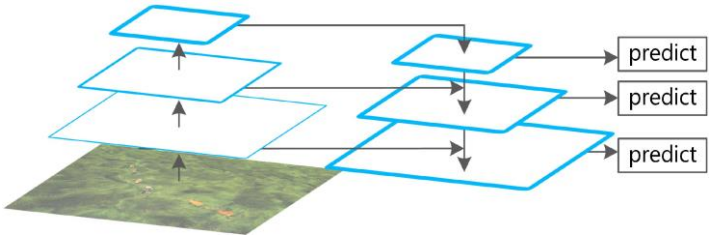

Figure 9. The pyramid structure of feature maps [13].

TABLE II. ANALYSES OF DIFFERENT METHODS AND PERFORMANCE COMPARISON ON COCO DATASET

Stage	Algorithm	Advantage	Disadvantage	AP	APs	FPS
One-stage	SSD512 [35]	Feature fusion of multiple layers	Low detection accuracy for small targets	28.8	10.9	-
	DSSD513 [36]	Improved SSD APs	Increasement of time cost	33.2	13.0	-
	YOLOv3 [31]	Multi-scale feature extraction	-	33.0	18.3	20.0
	YOLOv3+SPP [14]	Improved precision compared to YOLOv3	Increasement of time cost	36.2	20.6	20.0
	YOLOv4 [33]	Without loss in speed, improved performance in multi-scale detection of targets	-	43.5	26.7	62.0
	EfficientDet-D3 [38]	The highest precision among the one-stage model	Slower than YOLOv4	45.8	26.6	23.0
	FCOS [39]	Using key points for detection	Slow	44.7	27.6	7.0
Two-stage	Faster R-CNN+++ [2]	Usage of residual network to overcome the degradation of deep networks	-	34.9	15.6	2.4
	Faster R-CNN+FPN [13]	Feature pyramid for feature enhancement by multi-layer feature fusion	Increasement of time cost	42	25.2	20.0
	Mask R-CNN [46]	High detection precision for small targets	Mainly used in target segmentation	39.8	22.1	11.0
	SNIP [44]	Improved capabilities for multi-scale target detection	-	43.4	27.2	-
	SNIPER [45]	High detection precision	-	46.1	29.6	2.5
Other	Cascade R-CNN [48]	High precision	Slower than two-stage methods	42.8	23.7	8.0
	DETR [49]	With simple structure, fast and high precision at the same time	Low detection accuracy for small targets	44.9	23.7	28.0

III. DATASETS AND EVALUATION CRITERIA FOR VIDEO BASED SMALL TARGET DETECTION

The research of video based small target detection requires specific image and video datasets that focus on small targets. To meet the demands of video based small target detection in different application scenarios, it is required that there are diverse types of videos in the dataset and small targets which meet the standard show up in the videos. The video datasets with small targets

listed in this section contains public datasets and datasets for specific researchers as shown in Table III, including daily lives, aerial targets, human action, A sample of each dataset is shown in Fig. 10.

A. Common Datasets

ImageNet VID: This dataset [50] is used in the ImageNet large scale visual recognition challenge for video based target detection, which is also known as ILSVRC 2015-VID. It contains 3862 snippets for training, 555 snippets for validation and 937 snippets for testing with 30 different categories. It also takes into consideration multiple factors such as movements, video backgrounds and average target quantity. Each frame of a video clip has complete labels. All the videos in ImageNet VID dataset have complete annotations with boundary boxes and tracking IDs. The evaluation criterion is mAP, which is same with image target detection.

YouTube-8M: This is a large-scale video dataset [51] released by Google Inc., which contains 8 million URLs of YouTube videos with the total length of 500 thousand hours. There are 4716 types of labels, i.e., 3.4 labels for each video in average.

Kinetics dataset: Kinetics is a human action video dataset [52]. This dataset focuses on human and contains 400 human action classes with at least 400 video clips for each action. Each video lasts about 10 seconds. All the video clips are cut out from different YouTube videos.

VOT series datasets: The VOT challenge hosted every year is a testing platform that focuses on single target tracking [53]. Since 2013, the dataset is updated every year till now, and has become a mainstream dataset in target tracking. VOT dataset has its own evaluation criteria. The initial VOT2013 has six attributes such as vibrating blurring, illuminating, size variation, blurring, and non-degradation. Each frame of the sequences in the dataset is annotated with visual attribute. New evaluation criteria or improvements of existing evaluation metrics are raised every year for the VOT challenge. There are three evaluation criteria for tracking, i.e., accuracy (A), robustness (R) and Average Expected Overlapping (AEO).

OTB-2015: WU [54] established OTB-2013 (OTB50) dataset for the evaluation of target tracking algorithms. It contains 50 sequences with complete annotation for tracking. For the videos in the dataset, there are 11 common tracking difficulties as the attributes for tracking evaluation, including Scale Variation (SV), Illumination Variation (IV), in-plane rotation (IPR), Occlusion (OCC), Deformation (DEF), Fast Motion (FM), Motion Blur (MB), Background Clutters (BC), Out-of-Plane Rotation (OPR), Out-of-View (OV), Low Resolution (LR). In 2015, the authors further extended the dataset to OTB-2015 with 100 video sequences

CAVIAR: The CAVIAR dataset [55] sums up 28 video sequences for result comparison in the seminar PETS04 [55]. The lengths of these sequences range from 500 to 1400 frames and there are about 26,500 frames in total composed of six different activity scenarios. The

sequences are public surveillance oriented and each frame is labelled with bounding boxes and semantic description of the activities in the frame.

VIVID: The VIVID dataset [56] is an open source video dataset. It is available for general moving target detection as well as infrared based tasks. The dataset contains 9 sequences mainly composed of vehicle videos from aerial photography. The evaluation attributes include resolution, contrast ratio, pose and occlusion levels. The dataset is of high detection difficulty due to the properties like small-sized targets and severe occlusion.

KITTI: The KITTI dataset [57] is a real image dataset sampled from urban districts, countryside and highways with up to fifteen vehicles and thirty pedestrians in each image as well as occlusions and interruptions in different degrees. The whole dataset consists of 389 stereo and optical flow pairs, stereo visual odometry sequences of 39.2 length, and over 200k 3D-annotated images for objects.

TABLE III. COMMONLY USED VIDEO BASED TARGET DETECTION DATASETS

Name	Years	Development unit	Number of videos
ImageNet VID	2015	Olga Russakovsky, Jia Deng, Hao Su, etc.	5354
YouTube-8M	2016	Google	800w
Kinectics	2017	Deepmind	50w
VOT 2019	2019	Kristan M, Matas J, etc.	60
OTB-2015	2015	Wu Y, Lim J, Yang M H	100
CAVIAR	2004	R. B. Fisher	28
VIVID	2005	Dr. Thomas Strat	9
KITTI	2011	Karlsruhe Institute of Technology, Germany & Toyota Technological Institute, USA	20w

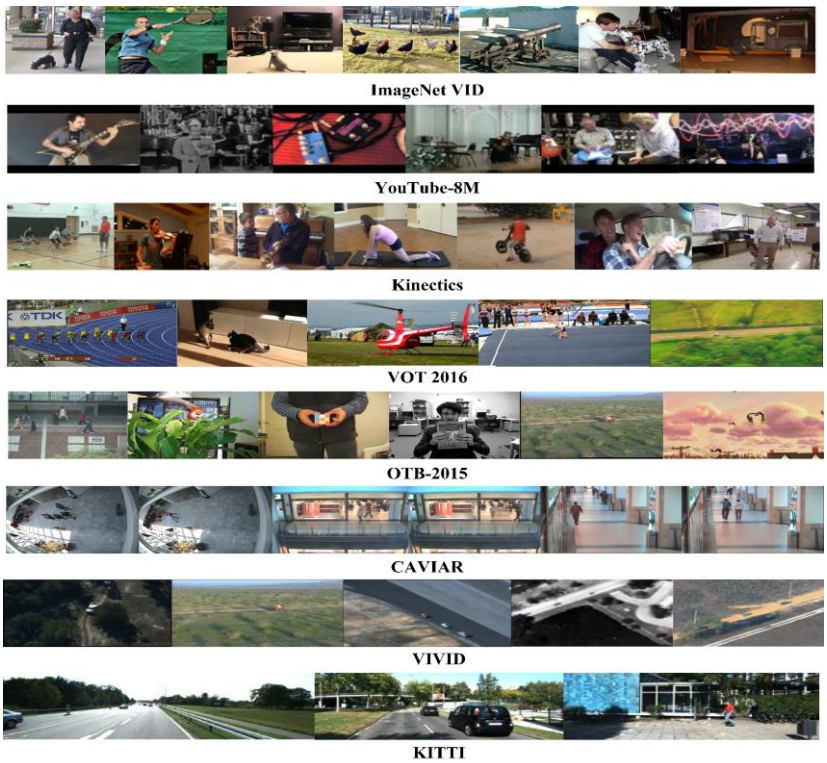

Figure 10. Samples of different datasets [50]-[57].

B. Evaluation Criteria

Both accuracy and speed are important for video based small target detection. The main criteria adopted for small target detection are the same with target detection, i.e., Average Precision (AP), mean Average Precision (mAP) and Frames per Second (FPS).

AP, *mAP*, *Aps:* Average Precision (AP) is aimed at the classification performance of one class of targets. It works in terms of a certain class in the dataset. The mAP criterion is the mean of all types of APs in the dataset in terms of the whole dataset. Generally, a better classifier obtains higher AP value. Lin T Y [1] gives the evaluation criteria of APs for small targets, which only focuses on the APs of targets with fewer than 32×32 pixels.

$$\mathrm{AP}=\frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} \tag{1}$$

where *TP* is the number of true positives; *FP* is the number of false positives; *FN* is the number of false negatives; *TN* is the number of true negatives.

FPS: Frame per Second (FPS) means the number of images processed in each second. In addition, the time for processing an image can also be used for detection speed evaluation.

IoU thresholds: The Intersection over Union (IoU) thresholds can be understood as the overlapping extent of the predicted bounding box and ground truth bounding box.

$$\text{IoU}=\frac{\text{Detection Result}\bigcap\text{Ground Truth}}{\text{Detection Result}\bigcup\text{Ground Truth}} \quad (2)$$

where Detection Result represents the predicted bounding box and Ground Truth represents ground truth bounding box.

IV. Applications and Future Research Direction

A. Applications of Video Based Small Target

The applications of video based small target detection is wide, including the fields of aerospace, remote sensing video processing [50]-[61], automatic driving, intelligent transportation [62]-[64], public security [65], [66], and face recognition [67]-69]. This is a common problem for most of the video based target detection tasks, and therefore the techniques have great development potential. Liang [50] employed video based small target detection to UAV detection & tracking. The authors proposed a detection and tracking method that focuses on small high-speed moving targets. The trackers based on kernelized correlation filters are initialized with the detection results, and the tracking results are constantly updated to avoid false-alarms. Y. Sun [66] proposed to improve the detection of small moving targets with a flight path based detection algorithm. First, to reduce mis-detection rate, the adaptive foreground extraction method is proposed by fusing the regional textural features and difference probability. Second, to reduce the false alarm rate, a flight path correlated probability model is designed to build the connection for the suspected small moving targets. Detecting and tracking small targets in surveillance videos play an important role in the field of public security.

Video based small target detection has various application fields, while how to produce improvements and novelty for various application fields is a difficult point. However, video based small target detection algorithms are still at the early stage with rising attention and vast improving space on precision and speed.

B. Future Research Direction

1) Optimization algorithms based on one-stage model

Unlike simple small target detection, video based small target detection requires high precision as well as real-time performance. Although the accuracy of one-stage models is inferior to the two-stage models, the latest one-stage models are faster than most of the two-stage models. For the task of real-time small target detection in video, the one-stage models have the advantage in speed. Starting with a fast and precise one-stage model should be priority [33].

2) Anchor-free detector and key point detection

Despite the good results, anchor-based detector mostly relies on empirical setting of parameters. Dense anchor box is beneficial to improve the accuracy of small target detection, but a large number of redundant boxes will be generated. In recent years, frequently exploited anchor-free detector [70] breaks the constrain of anchor frames, which makes it more flexible, and even able to detect targets from the key points [71]. Therefore, anchor-free detector can recognize targets with different sizes.

3) Multi-modality fusion

The optimization on small-sized targets by single-modality target detections is still at the growth stage, but the quantities of literatures that incorporate infrared equipment are huge [72]-74]. The improvement progress of the algorithms is also comparably mature. Multi-modality fusion is always the hotspot of machine vision, and incorporating infrared information for small target detection is also one of the improvement directions in the future.

4) Resolution improvement

By introducing image super-resolution or generative adversarial networks [29], the image with small targets can be modified and reconstructed. The number of pixels and resolution of small targets will be improved with more feature information. As in Ref [69], the authors applied image super-resolution to search for tiny faces from images.

5) Enlarging the receptive field

As an important component of convolutional neural networks, receptive field has been modified and employed in small target detection by more and more researchers [75], [76]. Plug-in modules [10], [13], [14] are also proposed to enlarge the receptive field. Adjusting the balance of the expanded convolution layer and the convolution kernels will optimize the receptive field and detection efficiency.

6) Optimization of the backbone network

Optimizing the backbone networks [77], [78] can directly influence the performance of small target detection algorithms. An excellent backbone network can be applied not only to video based small target detection, but general target detection as well.

V. Conclusion

This article presents an intensive survey on the progress of video-based small target detection using deep learning, with recent techniques categorized into one-stage models and two-stage models. The network structures and plug-in modules are also described.

The experimental data prove that one-stage methods are faster and more suitable for real-time target detection, while two-stage methods can provide higher detection accuracy. In addition, integrating plug-in modules could further improve the accuracy of video-based small target detection. In future work, Transformer-based algorithms such as DETR can be further explored. This paper

provides useful information that can serve as reference for relevant researchers.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Ying Liu and Luyao Geng conducted the research; Ying Liu, Luyao Geng, Weidong Zhang, and Yanchao Gong revised and edited the paper together; Yanchao Gong and Ying Liu acquired funding; all authors have approved the final version.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China project (No. 61801381).

REFERENCES

[1] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L, Zitnick, "Microsoft COCO: Common objects in context," *Lecture Notes in Computer Science, PART 5*, pp. 740-755, 2014.
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016.
[3] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221-255, 2004.
[4] T. Reed, "Digital video processing," *The Computer Engineering Handbook*, 2001.
[5] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-Time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809-830, 2000.
[6] Y. Ivanov, A. Bobick, and J. Liu, "Fast lighting independent background subtraction," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 199-207, 2000.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2017, vol. 25.
[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Ohio, 2014, pp. 580-587.
[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016, pp. 779-788.
[10] C. Deng, M. Wang, L. Liu, and Y. Liu, "Extended feature Pyramid network for small object detection," arXiv preprint arXiv:2003.07021, 2020.
[11] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. the 14th European Conference on Computer Vision*, The Netherlands, 2016, pp. 483-499.
[12] Z. Li, P. Chao, Y. Gang, X. Zhang, and S. Jian, "DetNet: A backbone network for object detection," arXiv preprint arXiv:1804.06215, 2018.
[13] T. Y. Lin, *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017, pp. 2117-2125.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1904-1916, 2014.
[15] Y. Wu, L. Li, Z. Xiao, and S Liu, "Optical flow motion tracking algorithm based on SIFT feature," *Computer Engineering and Applications*, vol. 49, no. 15, pp. 157-161, 2013.
[16] S. Avidan, "Ensemble tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261-271. 2007.
[17] J. Shi and C. Tomasi, "Good features to track," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 1994.
[18] M. Liu, C. Wu, and Y. Zhang, "Motion vehicle tracking based on multi-resolution optical flow and multi-scale Harris corner detection," in *Proc. IEEE International Conference on Robotics and Biomimetics*, 2007.
[19] L. Bertinetto, J. F. Henriques, J. Valmadre, P. H. S Tor, and A Vedaldi, "Learning feed-forward one-shot learners," *Advances in Neural Information Processing Systems*, 2016.
[20] X. Liu, D. Shen, X. Zhang, and G. Zhang, "Improved LK method tracks mobile ball in complex background," *Computer Systems & Applications*, vol. 28, no. 7, pp. 221-227, 2019.
[21] H. W. Yuan, Y. Lu, H. C. Mao, and S. C. He, "A small moving object extraction algorithm based on optical flow," *Optics & Optoelectronic Technology*, vol. 10, no. 1, pp. 67-70, 2012.
[22] X. Han, G. Yuan, L. Zheng, Z. Zhang, and D, Niu, "Research on moving object detection algorithm based on improved three frame difference method and optical flow," in *Proc. 5th International Conference on Instrumentation and Measurement, Computer, Communication, and Control*, 2015.
[23] Q. T. Sun, X. Y. Zhao, and X. M. Zhang, "Improvement of color recognition for recognition of tiny objects by three frame difference method," *Equipment Manufacturing Technology*, pp. 135-137, 2019.
[24] Y. Ivanov, A. Bobick, and J. Liu, "Fast lighting independent background subtraction," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 199-207, 2000.
[25] M. Piccardi, "Background subtraction techniques: A review," in *Proc. IEEE International Conference on Systems, Man and Cybernetics*, 2004.
[26] Y. S. Qu, W. Tian, and Y. Li, "Detecting small moving target in image sequences using optical flow based on the discontinuous frame difference," in *Proc. Third International Symposium on Multispectral Image Processing and Pattern Recognition*, 2003.
[27] L. I. Yi, Z. X. Sun, B. Yuan, and Y. Zhang, "An improved method for motion detection by frame difference and background subtraction," *Journal of Image and Graphics*, vol. 14, no. 6, pp. 1162-1168, 2009.
[28] K. X. Dai, G. H. Li, T. Dan, and J. Yuan, "Prospects and current studies on background subtraction techniques for moving objects detection from surveillance video," *Journal of Image and Graphics*, vol. 7, pp. 919-927, 2006.
[29] Y. Liu, H. Y. Liu, J. L. Fan, Y. C. Gong, and Y. H. Li, "Survey of research and application of small object detection based on deep learning," *Acta Electronica Sinica*, vol. 48, no. 3, pp. 590-601, 2020.
[30] Z. Jiang, *et al.*, "Learning where to focus for efficient video object detection," in *Proc. European Conference on Computer Vision*. 2020, pp. 18-34.
[31] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
[32] M. Ju, H. B. Luo, Z. B, Wang, M. He, and B. Hui, "Improved YOLO V3 algorithm and its application in small target detection," *Acta Optica Sinica*, vol. 39, no. 7, pp. 253-260, 2019.
[33] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.
[34] C. Y. Wang, *et al.*, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2020.
[35] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. European Conference on Computer Vision.*, 2016, pp. 21-37.
[36] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," arXiv preprint arXiv:1701.06659, 2017.
[37] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision*, Venice, 2017, pp. 2999-3007.

[38] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Comput. Soc. Conf. Comput.*, 2020, pp. 10778-10787.
[39] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE International Conference on Computer Vision*, Seoul, October 2019.
[40] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.
[41] X. N. Liu, Z. P. Wang, Y. T. He, and Q. Liu, "Research on small target detection based on deep learning," *Tactical Missile Technology*, vol. 193, no. 1, pp. 106-113, 2019.
[42] R. Girshick, "Fast R-CNN," in *Proc. IEEE International Conference on Computer Vision*, Santiago, 2015, pp. 1440-1448.
[43] S. Ren, *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1137-1149, June 2017.
[44] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection - SNIP," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018, pp. 3578-3587.
[45] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," *Advances in Neural Information Processing Systems*, December 2018.
[46] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," in *Proc. IEEE International Conference on Computer Vision*, Venice, 2017, pp. 2961-2969.
[47] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018, pp. 6154-6162.
[48] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[49] N. Carion, F. Massa, G. Synnaeve, and N. Usunier, "End-to-End object detection with transformers," in *Proc. European Conference on Computer Vision*, Glasgow, 2020, pp. 213-229.
[50] O. Russakovsky, *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 211-252, March 2015.
[51] S. Abu-El-Haija, N. Kothari, J. Lee, P Natsev, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," arXiv preprint arXiv:1609.0867, 2016.
[52] W. Kay, J. Carreira, K. Simonyan, B. Zhang, and A. Zisserman, "The Kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.
[53] M. Kristan, *et al.*, "The visual object tracking VOT2016 challenge results," *Lecture Notes in Computer Science*, pp. 777-823, 2016.
[54] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 2013, pp. 2411-2418.
[55] R. B. Fisher, "The PETS04 surveillance ground-truth data sets," in *Proc. 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2004, pp. 1-5.
[56] R. Collins, X. Zhou, and S. K. Teh, "An open source tracking testbed and evaluation web site," in *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2005, vol. 2.
[57] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 2012, pp. 3354-3361.
[58] D. Liang, S. Gao, H. Han, and N. Z. Liu, "UAV detection in motion cameras combining kernelized correlation filters and deep learning," *Acta Aeronautica et Astronautica Sinica*, vol. 41, pp. 159-171, September 2020.
[59] X. Li, X. Xu, and J. Li, "Small target detection in remote sensing images based on aviation security," *Aero Weaponry*, vol. 27, pp. 54-61, March 2020.
[60] X. Bao, *et al.*, "Context modeling combined with motion analysis for moving ship detection in port surveillance," *Journal of Electronic Imaging*, vol. 22, April 2013.
[61] Y. Ren, C. Zhu, and S. Xiao, "Small object detection in optical remote sensing images via modified faster R-CNN," *Applied Sciences*, vol. 8, May 2018.
[62] V. Kastrinaki, M. Zervakis, and K. Kalaitzakis, "A survey of video processing techniques for traffic applications," *Image and Vision Computing*, vol. 21, pp. 359-381, April 2003.
[63] O. Bulan, R. P. Loce, W. Wu, Y. R. Wang, E. A. Bernal, and Z. Fan, "Video-Based real-time on-street parking occupancy detection system," *Journal of Electronic Imaging*, vol. 22, February 2013.
[64] R. Rios-Cabrera, T. Tuytelaars, and L. V. Gool, "Efficient multi-camera vehicle detection, tracking, and identification in a tunnel surveillance application," *Computer Vision and Image Understanding*, vol. 116, pp. 742-753, June 2012.
[65] C. Eggert, A. Winschel, Z. Dan, and R. Lienhart, "Saliency-Guided selective magnification for company logo detection," in *Proc. 23rd International Conference on Pattern Recognition*, 2016, pp. 651-656.
[66] Y. F. Sun, J. Wu, Y. Y. Huang, and G. M. Tang, "A small moving object detection algorithm based on track in video surveillance," *Journal of Electronics & Information Technology*, vol. 41, pp. 2744-2751, November 2019.
[67] Z. Yang, J. Li, W. Min, and Q. Wang, "Real-Time pre-identification and cascaded detection for tiny faces," *Applied Sciences*, vol. 9, p. 4344, October 2019.
[68] P. Hu and D. Ramanan, "Finding tiny faces supplementary materials," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017, pp. 951-959.
[69] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018, pp. 21-30.
[70] Z. Zhong, L. Sun, and Q. Huo, "An anchor-free region proposal network for faster R-CNN-based text detection approaches," *Document Analysis and Recognition*, vol. 22, March 2019.
[71] Z. Dong, G. Li, Y. Liao, F. Wang, P. Ren, and C. Qian, "CentripetalNet: Pursuing high-quality keypoint pairs for object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 2020.
[72] C. Wang and S. Qin, "Approach for moving small target detection in infrared image sequence based on reinforcement learning," *Journal of Electronic Imaging*, vol. 25, May 2016.
[73] C. Yang, J. Ma, M. Zhang, S. Zheng, and X. Tian, "Multiscale facet model for infrared small target detection," *Infrared Physics & Technology*, vol. 67, pp. 202-209, July 2014.
[74] W. Q. Tong, Y. Hua, C. C. Huang, W. Jin, and L. Yang, "Algorithm of small moving target detection based on infrared and visual image fusion," in *Proc. International Symposium on Photoelectronic Detection and Imaging*, 2008.
[75] W. F. Wang, J. Jin, and J. J. Chen, "Rapid detection algorithm for small objects based on receptive field block," *Laser & Optoelectronics Progress*, vol. 57, pp. 250-255, July 2020.
[76] J. R. Chen and L. Peng, "Detection algorithm of small target in receptive field Block," *Journal of Frontiers of Computer Science and Technology*, pp. 1-12, March 2020.
[77] M. Ju, J. Luo, P. Zhang, M. He, and H. Luo, "A simple and efficient network for small target detection," *IEEE Access*, vol. 7, pp. 85771-85781, June 2019.
[78] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. 30th IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017, pp. 1222-1230.

Ying Liu received the B.E. degree in School of Information Engineering from Xidian University, China, M.Eng. in school of Electrical Engineering from the National University of Singapore, and Ph.D. in School of Computing and Information Technology from Monash University, Australia. She is currently a full professor in the School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications (XUPT), China. Her research activities focus on image/video retrieval.

Luyao Geng received her B.E. degree in Communication Engineering from Linyi University, China. She is now pursuing the M.Sc. degree in Xi'an University of Posts and Telecommunications. Her research focuses on small target detection.

Weidong Zhang received his B.Sc. degree in Biomedical Engineering from Zhejiang University, China. He obtained his Ph.D. degree from Shandong University. He is now an associate professor in School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications. His research focuses on indoor scene understanding.

Yanchao Gong received his B.Sc. degree from Shandong University of Technology, China. He obtained his M.Sc. and Ph.D. degrees from Northwestern Polytechnical University. He is now a lecturer in School of School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications. His research focuses on video rate control.

Zhijie Xu received the B.Sc. degree in communication engineering from the Xi'an University of Science and Technology, China. He received his Ph.D. degree from the University of Derby, UK. He is now a full professor at the University of Huddersfield, UK. His research interests include visual computing, vision systems, data science and machine learning.