AGVS: A New Change Detection Dataset for Airport Ground Video Surveillance

Xiang Zhang, Chang Shu, Shuai Li, *Member, IEEE*, Celimuge Wu, *Senior Member, IEEE*, and Zhi Liu, *Senior Member, IEEE*

***Abstract*— Change detection is the foundation of intelligent video surveillance of the airport ground. However, experiments have shown that change detection algorithms with good performance on traditional datasets (e.g., CDnet2014) perform poorly in airport ground surveillance. The reason is that traditional datasets focus on the diversity of scenarios, while the practical application requires robustness against various changes in a single scene. We posit that the solution to this problem is to establish a unique dataset for airport ground surveillance and develop specific algorithms for this scenario. In this paper, we present an Airport Ground Video Surveillance benchmark (AGVS) for change detection of the airport ground. AGVS includes 25 long videos, amounting to about 100000 frames and accurate ground truth for all frames. Each video contains multiple challenges specific to the airport ground (e.g., haze, camouflage, strip shape, shadow and illumination change, simultaneous multi-scale objects) and various appearance changes of the aircraft). Change detection ground truth is generated by manual annotation. The AGVS benchmark can be downloaded from www.agvs-caac.com. Furthermore, we conduct a simple review of current change detection algorithms, both unsupervised or supervised, and then 21 state-of-the-art algorithms are tested and analyzed on the AGVS benchmark. Finally, we conclude with algorithm design principles of change detection for airport ground surveillance.**

***Index Terms*— Benchmark, change detection, background subtraction, airport ground, video surveillance.**

I. Introduction

THE airport ground, as a typical transportation scene, is becoming busier as air passenger and cargo volumes continue to grow at a rapid pace. The airport ground needs to become more advanced to meet the operational challenges, especially for capacity-constrained airports. Change detection, also known as background subtraction or foreground/background modeling/detection/segmentation, aims to segment the moving object from the scene background. Change detection is the foundation of many intelligent video surveillance tasks of the airport ground [1], such as aircraft tracking, visual conflict alerts, and visual docking guidance.

Four state-of-the-art change detection algorithms are tested in airport ground videos, as shown in Fig. 1. We can see that the detection results are poor. Our further experiments indicate that the performance of almost all change detection algorithms is significantly reduced from fundamental research to the airport ground, and the average reduction of detection accuracy is about 25%. That is, there is a gap between change detection in real application and that in fundamental research. This phenomenon has also been noticed by other scholars, such as Garcia-Garcia *et al.* [2].

The study of change detection in fundamental research is based on CDnet [3] and similar datasets, which include multiple scenes but with limited samples for a single scene. In this case, excellent algorithms in fundamental research imply applicability to various scenarios, that is, good generalization. However, in practice, even a single scene may have complex changes, as in Fig. 1. Accordingly, what the real application needs is robustness of change detection against various changes in a specific scene, so generalization is not the primary concern. In other words, the real application and fundamental research put forward different requirements for change detection. This is why state-of-the-art algorithms in fundamental research fail in airport ground surveillance. To bridge the gap between fundamental research and real application, the best solution is to design datasets specific to a single scene. Such a dataset should cover the wide variety of challenges in the scene. Then, new change detection algorithms for the challenges in the scenario can be designed. Although such an algorithm may be not generalized, it is the most practical algorithm for the scene.

In this paper, a new change detection benchmark called the Airport Ground Video Surveillance benchmark (AGVS) is presented. There are 25 long videos, amounting to about 100000 frames, in AGVS. All videos contain multiple challenges specific to the airport ground, such as haze, camouflage, shadow, strip shape, simultaneous multi-scales, non-uniform illumination change, various shape and color changes of aircraft, and different weather conditions. For the given videos, we adopt manual annotation to generate the pixel-wise ground

Manuscript received 24 May 2021; revised 31 October 2021, 9 February 2022, and 18 April 2022; accepted 8 June 2022. Date of publication 27 June 2022; date of current version 7 November 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB1600500 and Grant 2018YFE0203900, in part by the National Natural Science Foundation of China under Grant U19A2052 and Grant U1733111, and in part by the Chengdu Science and Technology Project under Grant 2021-JB00-00025-GX. The Associate Editor for this article was Z. Duric. *(Corresponding author: Chang Shu.)*

Xiang Zhang and Chang Shu are with the School of Information and Communication Engineering, Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: uestchero@uestc.edu.cn; changshu@uestc.edu.cn).

Shuai Li is with the School of Control Science and Engineering, Shandong University, Jinan 250100, China (e-mail: shuaili@sdu.edu.cn).

Celimuge Wu and Zhi Liu are with the Department of Computer and Network Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan (e-mail: celimuge@uec.ac.jp; liuzhi@uec.ac.jp).

Digital Object Identifier 10.1109/TITS.2022.3184978

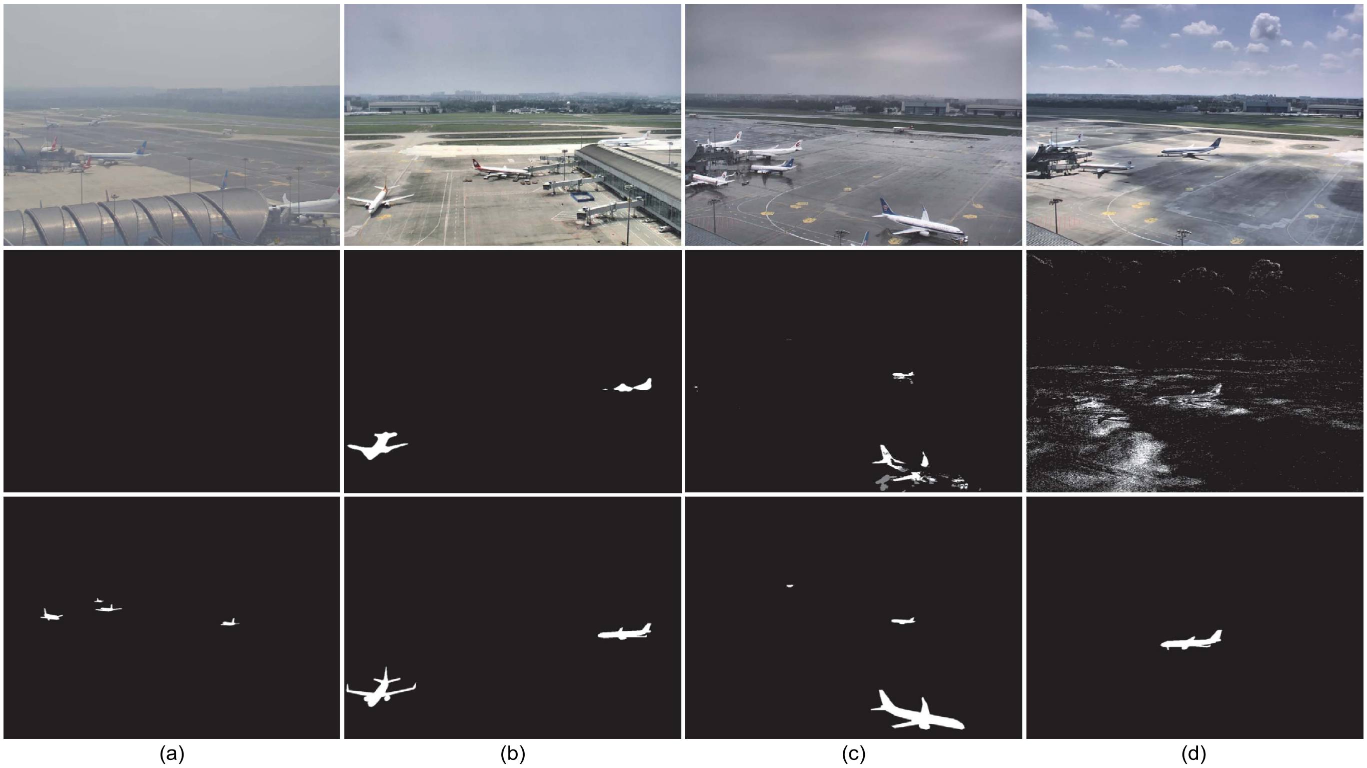


Fig. 1. Top row: typical frames with haze, camouflage, simultaneous multi-scales, shadow, and non-uniform illumination, respectively. Middle row: change detection results by SuBSENSE [20], RGMP [41], KNN [26] and FGMM [27], respectively. Bottom row: ground truth.

truth. Some extreme cases like nighttime are not included in AGVS, but six such videos without ground truth are also available on the AGVS website. Furthermore, we conduct a simple review of change detection algorithms based on whether they are supervised or not, and then 21 classic algorithms are selected and tested on AGVS. Finally, we include an additional discussion about how to develop change detection algorithms for airport ground surveillance.

The rest of this paper is organized as follows. Related work is reviewed in Section II. The AGVS benchmark is introduced in Section III. Section IV presents details of the experiments and discussions, and the conclusion is given in Section V.

II. Related Work

There are already some datasets related to airports, such as the FGVC-Aircraft dataset [4] and ALERT dataset [5]. The FGVC-Aircraft dataset is for recognition, which contains pictures of hundreds of aircraft and the corresponding category labels. The ALERT dataset is used for pedestrian re-identification, and all samples are collected in an airport terminal. Therefore, the above datasets are not change detection datasets. There have been some change detection datasets and a large number of change detection algorithms, which are discussed briefly below.

A. Change Detection Datasets

The website of the book by Bouwmans *et al.* [6] links to many video-surveillance-type datasets, some of which are change detection datasets. These datasets are broadly divided into two categories, for humans and for animals, but there are no airport-themed datasets. We note that some datasets do not provide pixel-wise ground truth, so they cannot be used as a change detection benchmark.

Wallflower [7] was the first change detection dataset. It contains seven short clips; each represents a single challenge, like moved objects, time of day, light switch, waving trees, camouflage, bootstrapping, and foreground aperture. The pixel-wise ground truth is provided for only one frame of each clip. The TUVD dataset [8] is one of the latest datasets, which includes 55 videos and ground truth for all frames under degraded atmospheric weather conditions, such as fog, dust, and poor illumination. The famous PETS series [9], which starts from year 2000, was designed with the primary goal of tracking and recognition, and only the bounding-box-type ground truth is provided. Therefore, such datasets cannot be used as the change detection benchmark.

There are also some synthetic or semi-synthetic change detection datasets. The advantage of synthetic datasets is that the ground truth can be generated automatically by a computer without time-consuming manual annotation. Brutzer *et al.* [10] presented a synthetic dataset rendered by Mental Ray, and ground truth data were generated by Maya Vector. All sequences in this dataset show only one scene, a street corner, but cover different challenges, such as dynamic background, bootstrapping, darkening, noisy night, shadow, camouflage, and video compression. There are 29 outdoor video clips in BMC [11], most of which are synthetic. Complete pixel-wise ground truth are provided for nine real videos and ten synthetic videos. This dataset focuses on outdoor environments such as wind, sun or rain. Although the computer-generated ground truth is precise, the data diversity of synthetic data is relatively insufficient, as the natural scenes are diverse and full of changes.

Currently, CDnet2012 and CDnet2014 [3] are the most widely used change detection benchmarks. All clips in the two datasets are real videos, and the complete pixel-wise ground truth is available. The CDnet2012 has 31 videos and a total of about 70000 frames. This dataset covers such challenges as indoor and outdoor situations, dynamic background, camera jitter, shadow, thermal, and intermittent object motion. Based on CDnet2012, the CDnet2014 adds another 22 real videos, which have about 70000 frames. The new challenges shown in CDnet2014 are bad weather, a low frame rate, night, PTZ, and air turbulence. The scenarios are diverse in CDnet series, such as a corridor, park, lakeside, bus station, street, highway, library, office, and blizzard. In terms of the data diversity and high-quality ground truth of the CDnet series, they are frequently used to evaluate the generalization of change detection algorithms.

All the above datasets are visible spectrum datasets. There are also a few change detection datasets based on non-visible spectral sensors, such as infrared sensors and RGBD sensors. The RSIR dataset [12] was captured by a medium-wave infrared sensor and focuses on remote scenes. Pure infrared data have some disadvantages such as the lack of object details, so the fusion of visible video and infrared video is a more effective approach. Such a dataset is presented in [13], which includes 25 aligned grayscale-thermal video pairs with high diversity. The SBM-RGBD dataset [14] was captured by the RGBD sensor, where the depth data of the scene are provided in addition to color data. The depth data represent the distance from the device to the objects in the scene. Although the additional depth information is beneficial for change detection, the depth data have several problems, such as depth camouflage, specular materials, near objects, and imaging distance [15]. Therefore, RGBD data are limited to indoor environments. In addition, there are also some datasets for other detection problems, such as the EVA dataset [16] for human detection and the MAR dataset for boat detection [17].

B. Change Detection Algorithms

Change detection algorithms can be classified as unsupervised or supervised. The unsupervised methods do not need ground-truth images for training, while ground-truth images are required for training in supervised methods.

1) Unsupervised Methods: As ground-truth samples are not required in unsupervised methods, they can be easily adapted to various scenarios. The most widely used unsupervised strategy is statistical modeling [18], including background modeling and bilayer modeling.

Background modeling only models the background scene, which can be divided into two classes: non-parametric models and parametric models. These approaches model each pixel as a random variable with an associated Probability Density Function (PDF). Kernel Density Estimation (KDE) [19] is the first non-parametric model, where the PDF is estimated from samples using KDE, without any assumptions about the underlying distributions. For a new observation, its PDF is compared with a fixed threshold for classification. Besides color features, spatiotemporal binary features are adopted in SuBSENSE [20] to detect subtle local changes. Sample counting is used in ViBe [21] instead of KDE, making this method extremely fast. The classification threshold in PABS [22] is adaptive to meet the background dynamics. The HSV color space is used in SOBS [23], and each pixel is enlarged 3 × 3 times in models to account for the spatial relationship between pixels. GMM [24] is a typical parametric model, where the PDF is specified as a combination of a fixed number of Gaussian components. The probability estimation may be inaccurate if the specified Gaussian distribution cannot fit the underlying distribution [25], so scholars have made many improvements. A recursion strategy is used in KNN [26] to select the appropriate number of Gaussian components. The parameters in FGMM [27] vary in an interval with uniform possibilities instead of fixed values. Consequently, the likelihood probability of FGMM is an interval rather than a precise number to account for the GMMs' uncertainty. Bodids [28] is a bilayer method, where both the foreground and background are non-parametrically modeled and used to competitively classify new observations.

Another unsupervised strategy is Principal Component Analysis (PCA)-based or Robust-PCA (RPCA)-based modeling that relies on low-rank and sparse decomposition. The low-rank component and sparse component of videos correspond to background and foreground, respectively. This type of method is well known for robustness in the presence of illumination changes. The EigenSpace model [29] is formed by PCA on background images. The input image is projected onto the space expanded by the EigenSpace model, and the foreground is detected by thresholding the Euclidean distance between the input image and projected image. RPCA methods can be divided into two classes: batch RPCA and incremental RPCA. Batch RPCA methods, such as GoDec [30], process video frames in batches, so they are not real-time and mostly work offline. Incremental RPCA is the online modification of previous methods by incrementally computing only one frame at a time. The GRASTA [31] employs the based-L_1 norm loss function for each frame to encode the sparse foreground component, while a more complex loss term is introduced in GOSUS [32] to represent the structured sparsity of the foreground. The MEDRoP [33] is the incremental extension of the basic RPCA algorithm. The incremental learning step in IMTSL [34] is based on tensor representation. In addition to RPCA and statistical modeling, there are other types of change detection algorithms. In CodeBook [35], each background pixel is modeled with codebooks, enabling the capture of the structural information of the background.

2) Supervised Methods: Supervised change detection algorithms are essentially video object segmentation algorithms. The task of video object segmentation is to manually specify the object to be segmented in the first frame, and then automatically segment the specified object from all subsequent frames. If the object to be segmented is specified as the moving object, the video object segmentation algorithms can be adapted to change detection. The main trend toward supervised change detection is based on deep learning, which can be divided into image-type and video-type methods.

The temporal correlation between frames is not considered in image-type methods. The encoder–decoder network is widely used in image-type methods. Lim and Keles [36] proposed FgSegNet for change detection, which contains a triplet convolutional neural network (CNN) for feature encoding and a transposed convolutional network for decoding. In order to obtain richer features, the triplet CNN operates in three different scales in parallel for feature encoding of the same input image. This method is further extended in FgSegNet2 [37] by introducing feature fusions into the network to enhance multi-scale features.

The cascade structure of multiple networks is also considered in image-type methods. In CascadedCNN [38], a multi-scale CNN similar to the encoding part of FgSegNet is used as the first level of the cascade framework. Then, output foreground probability maps of the first level are concatenated with the original frames and fed to the second level, another multi-scale CNN model, to refine the foreground probability maps. As stated in [38], such a cascade structure can be used to enforce the spatial coherence constraint so that better results can be obtained, with more cascaded levels.

Video-type methods are mainly based on the 3D convolutional neural network (3D CNN), long short-term memory (LSTM), or two-stream network, which explicitly use temporal cues in videos for classification. For example, the two-stream network consists of two parallel branches, namely appearance stream and motion stream, where the motion stream is generally based on the optical flow between frames. The two-stream network is employed in SegFlow [39] for object segmentation, where the appearance and motion streams are based on the fully convolutional network and FlowNetS [40], respectively. To enable communications between the two branches, SegFlow propagates feature maps between the two streams bidirectionally during the upsampling stage.

The encoder–decoder structure is also considered in video-type methods. For example, RGMP [41] exploits a Siamese encoder–decoder framework to carry out object segmentation. The Siamese encoder includes two parameter-shared branches: a reference branch and a target branch. Inputs to the reference branch are a reference image, generally the first frame of a video, and the corresponding ground truth mask, which specifies the targets to be detected. For the target branch, inputs include the current frame and the mask of the previous frame. Therefore, the function of RGMP is to propagate the target mask from the previous frame to the current frame by referencing the ground-truth information of the first frame.

We note that the semi-supervised approach has been proposed for change detection in recent years and achieved good results. For example, a semi-supervised method called GraphBGS-TV was presented in [42] to detect moving objects by minimizing the total variation of graph signals. Another graph learning based method was introduced in [43]; it requires less labeled data than deep learning methods but has demonstrated competitive results on both static and moving cameras. In addition to the above methods, there are many excellent change detection algorithms. See [44] and [45] for a comprehensive review of unsupervised algorithms, and [46], [47] for a review of supervised algorithms.

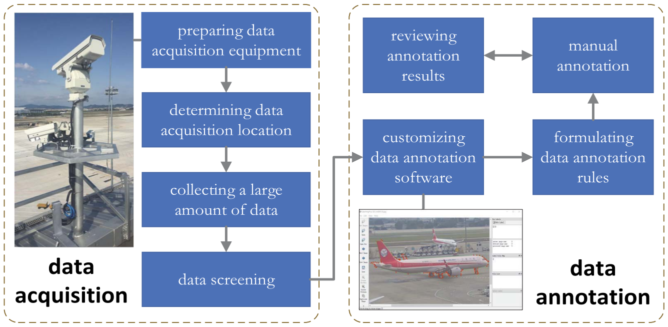

preparing data acquisition equipment
determining data acquisition location
collecting a large amount of data
data screening
data acquisition
reviewing annotation results
manual annotation
customizing data annotation software
formulating data annotation rules
data annotation

Fig. 2. The whole construction process of AGVS.

III. AGVS Benchmark

Previously compiled change detection datasets, such as the CDnet series, generally focus on data diversity. In such datasets, there is usually less data representing a single scene and less data reflecting the particularity of the scene. On the contrary, in practical application, the algorithm requires robustness against various changes in a specific scenario. This inspires us to build a dataset covering a wide range of changes for a single scene and to design unique algorithms for the scene. Based on this consideration, a new change detection dataset, AGVS, is presented for airport ground video surveillance. The whole construction process of AGVS is shown in Fig. 2, which includes two steps: data acquisition and data annotation.

A. Data Acquisition

Because the airport ground is a semi-militarized area, the authority of data acquisition is obtained through cooperation with the Civil Aviation Administration of China (CAAC). The data collection site is a large international airport in Southwest China. The data acquisition equipment includes four fixed cameras and a PTZ camera, with resolutions of 1280*960 and 1280*720, respectively. Unlike most fundamental research datasets composed of short videos, long videos are collected for AGVS. Short videos containing a single challenge are convenient for experimental analysis but are not in line with the real application, where the scene may change and multiple challenges often coexist. This is why the performance of fundamental research algorithms decreases in airport ground surveillance. We collect hundreds of long videos within a few months and finally select 25 video clips for AGVS (*S1*∼*S25*). The first 22 videos (*S1*∼*S22*) are captured by the fixed camera from different viewing angles or focal lengths, while the last three videos (*S23*∼*S25*) are captured by the PTZ camera. The total number of frames is about 100000.

Typical frames of some videos in AGVS are shown in Fig. 3. The first nine videos (*S1*∼*S9*) are captured with the same viewing angle and focal length but different weather conditions, such as moderate overcast (*S7*), heavy overcast (*S8*), in the rain (*S5*, after the rain *S9*), sunny day (*S6*), sunny days with sheet thin clouds (*S1*), lead clouds (*S3*), flocculent

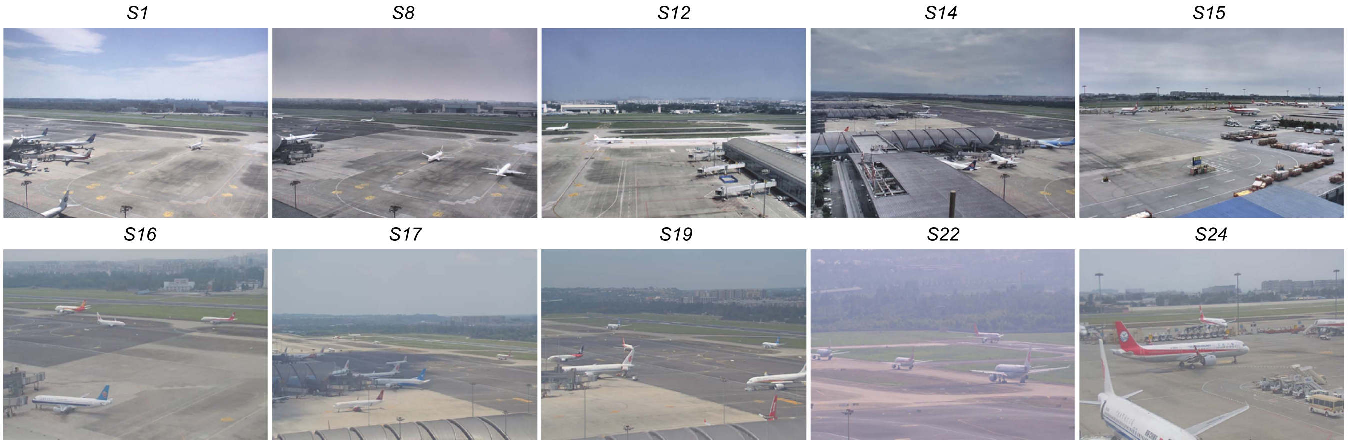


Fig. 3. Example pictures of some videos in AGVS.

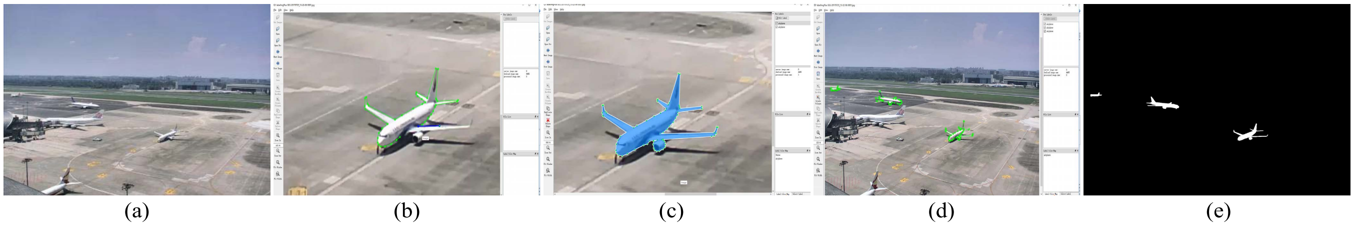


Fig. 4. Illustration of the ground-truth generation process.

clouds (*S4*), or mist (*S2*). Moreover, *S10*, *S11*, and *S13* have the same viewing angle and focal length but different weather conditions. The three videos mainly focus on the camouflage problem. *S16*, *S17*, *S18*, *S19*, and *S21* have the same viewing angle but different focal lengths, such as a long focal length (*S16*), short focal length (*S17*), and normal focal length (*S19*). Further, *S14*, *S15*, *S21*, and *S22* also have different viewing angles. The PTZ camera shows different motion patterns in *S23*, *S24*, and *S25*, such as switching between moving and stopping (*S23*), or continuing to move (*S24*, *S25*).

We have introduced AGVS from the perspective of the viewing angle and focal length. We do not further categorize the videos in AGVS based on various challenges in the scene. This is because the premise of categorization based on challenges is to obtain videos containing only a single challenge, while all videos in AGVS actually contain multiple challenges. For example, haze can be seen together with different focal lengths and viewing angles (*S16*∼*S22*). Moreover, the non-uniform illumination change can be seen in many sequences, such as *S4* and *S17*. The camouflage frequently appears in different video segments. In addition, some special challenges in AGVS do not exist independently. For instance, the shape and color variation of aircraft is reflected by the comparison of multiple targets in all videos, and the strip shape exists any time the aircraft targets appear. Because the airport ground is a broad scene, the challenge of the simultaneous multi-scales exists in any video of AGVS. The reason for this phenomenon is that we cannot access the airport ground to set the scene background and control the moving aircraft owing to the semi-militarized nature of the airport ground. Therefore, we can only obtain the real data of the uncontrollable airport ground scene. In fact, the data of all real scenes, including the airport ground, always coexist with multiple challenges. Although data containing a single challenge cannot be collected, we attempt to select such videos where a certain challenge can dominate the whole sequence or a segment of the sequence to a certain extent for AGVS. The description of various challenges in AGVS is given at the end of this section.

B. Data Annotation

The manually annotated pixel-wise ground truth is generated for all frames in AGVS, which is a time-consuming task. The change-detection ground truth for each frame is a binary mask, where one value represents the moving object, and the other represents the background. The annotation tool is developed based on LabelImg, which has been used in some detection algorithms such as R-CNN and YOLO to generate training samples. The annotation process is illustrated in Fig. 4. First, we select an aircraft and zoom in on the area where the aircraft is located. Second, we use the mouse to draw a polygon along the outline of the aircraft (Fig. 4(b)). In order to fit the contour of the aircraft as accurately as possible, we draw more lines in curved parts, such as the nose and engine of the aircraft, as shown in Fig. 4(c). Next, we annotate each target in turn as in Fig. 4(d), and finally get the complete ground-truth mask in Fig. 4(e). The above process is repeated for each frame in AGVS. Because the AGVS dataset has a total of about 100000 frames, more than ten students participated in the annotation work part-time, which took about 10 months.

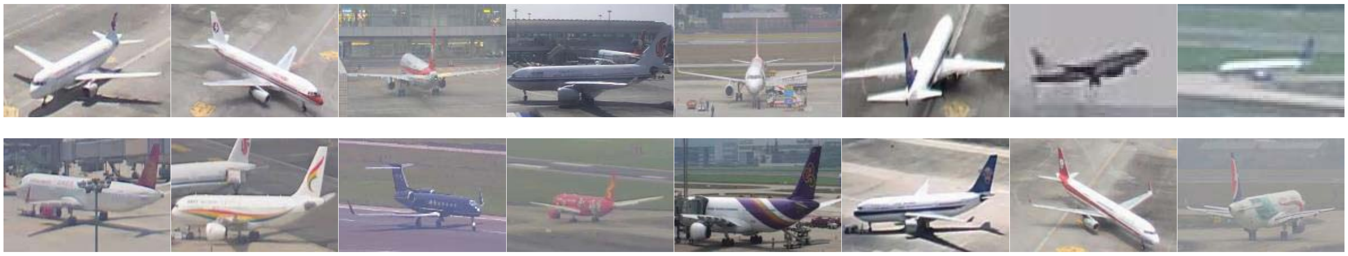


Fig. 5. Different shape and color patterns of aircraft in AGVS.

There are some important considerations in the above annotation process. First, before labeling a target, it is necessary to verify whether the target is in motion or at rest. In change detection, change is defined as the moving object, so only the moving object should be labeled, and the stationary object should not be marked as foreground. Because there are many intermittent aircraft in AGVS, the motion state of the aircraft in each frame must be carefully judged. In our work, this is achieved through repeated playback and visual observation of the whole video before annotation. Second, as most targets in AGVS are relatively small, we set a rule that if the target is too small to see an aircraft from the outline, it is not marked. The target area at this point is about 100 pixels. Finally, owing to the fatigue that occurs when people perform repetitive work, sometimes, the labeled data are not accurate enough. At this point, there must be people to check the annotation results. If the mask is not accurate enough, it must be re-annotated. In our work, two people independently review the annotation results frame by frame. Only when two people agree is the annotation qualified.

The original LabelImg was not designed for change detection, so it is not applicable for use in our work. In order to facilitate ground-truth annotation, we added some new functions in LabelImg. For example, it is allowed to load the mask of the previous frame and superimpose it on the current frame. Then, the mask of the current frame can be obtained by adjusting the polygon of the loaded mask. Considering the correlation between adjacent frames, this strategy can accelerate the manual annotation. This customized annotation tool can also be downloaded from the website of AGVS.

C. Challenges

Some challenges in AGVS are unique to the airport scenario, while others are common to other scenes but have special manifestations on the airport ground. Although all videos in AGVS have multiple challenges, some challenges can be highlighted in certain videos or video clips. Next, we describe each challenge and its distribution in the AGVS dataset.

1) Haze: Fig. 1(a) shows a type of disastrous weather, haze. Haze leads to a decrease in the image contrast, which makes it difficult to distinguish the moving object from the background. Further, the influence of haze on the image contrast increases with the increase of the monitoring distance. Because the monitoring distance of the airport ground can be as far as kilometers away, even mild haze has a serious impact on change detection. As shown in Fig. 1(a), all moving targets are misdetected under mild haze. However, this degree of haze has little effect on close surveillance. The mild or moderate haze can be seen in several sequences in AGVS, such as from *S16* to *S22*. More severe haze is not included in AGVS, because the image contrast drops to the point where the foreground and background cannot be distinguished by the human eye. Thus, haze is one of the biggest challenges for airport ground surveillance.

2) Camouflage: Camouflage is when the moving target has similar colors as the occluded background area. In this case, it is difficult to completely detect the moving object, because the foreground and background are too close in the feature space. Camouflage is widespread in all scenarios, but it has a special manifestation in airport ground surveillance. All over the world, airport grounds made of cement are gray-white, and the main color of most civil aircraft is also white, as shown in Fig. 1(b). In such cases, the camouflage problem is unavoidable, and the detection defect is serious. Camouflage of varying degrees can be seen in all videos of AGVS.

3) Simultaneous Multiscale Detection: Although the scale change can be seen in many datasets, it is unique in airport ground surveillance. In common surveillance scenarios, the challenges caused by scale change are small-scale target detection and large-scale target detection. Small targets are easily missed or submerged in the noise, while large targets are prone to detection defects and fractures. This phenomenon can also be seen in airport ground surveillance. However, in airport ground surveillance, the challenge of scale change lies not only in small and large targets but also in the simultaneous existence of multi-scale targets. As shown in Fig. 1(c), both targets within dozens of meters are visible in details, and far beyond a kilometer and even blurred in outline, they can be seen in the same camera field of view at the same time. Such concurrent multi-scales are due to the huge size of airport grounds and can be seen in all videos of AGVS.

The solutions to small and large target detection are different. For simultaneous multi-scale targets, these solutions may conflict with each other. As shown in Fig. 1(c), post-processing is used in KNN [26] to remove the shadow area under the large-scale aircraft (shadow pixels are indicated in gray). However, the smallest aircraft is removed at the same time. Therefore, how to detect multi-scale targets simultaneously, including small and large target detection, is a special challenge for airport ground surveillance.

4) Shadow and Non-Uniform Illumination Change: Shadow and illumination change can be seen in many datasets. As the airport ground is an outdoor scene, shadow and illumination

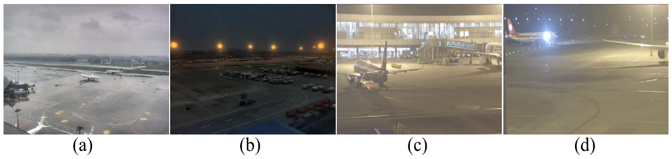


Fig. 6. (a) to (d): Water mist, midnight, self-luminescence, and strong reflection, respectively.

must be considered in AGVS. Even in weak illumination or haze, the aircraft casts shadows owing to its huge size. Both soft and strong shadows are visible in almost all sequences of AGVS. However, there is something special about the illumination change in the airport ground. As shown in Fig. 1(d), there is large-area non-uniform illumination change caused by cloud movement. This is because the airport ground is too broad. The illumination change in AGVS is almost non-uniform.

5) Shape and Color Variation: The shape of a plane looks different from different viewing angles and distances. Head-looking, side-looking, tail-looking, close-range, and long-range aircraft can be seen in AGVS. The color of aircraft also varies greatly in AGVS, such as all-white, all-red, all-blue, and white-based aircraft mixed with various strips or patterns of other colors. Some shape and color patterns in AGVS are shown in Fig. 5. Because there are all types of planes, AGVS cannot cover all aircraft appearance variations. Shape and color variation is a key challenge for supervised change detection, which generally expects that the training samples are capable of covering all appearance changes. However, no dataset can meet this requirement.

6) Strip Shape: The fuselage and wing of aircraft have strip shapes. It is difficult to detect such objects completely. For example, the fuselage often encounters detection defects and fractures, while the thin-strip wing sometimes cannot be detected at all. We consider the strip shape as an independent challenge, because the detection integrity of the fuselage and wing is important for some applications in airport ground surveillance. For instance, for visual conflict alerts and visual docking guidance, the wingtips of aircraft wings must be accurately located before further processing.

7) PTZ Camera: The video acquisition equipment from *S23* to *S25* is the PTZ camera. In such cases, both the foreground and background are moving. The PTZ video is a special challenge for unsupervised methods, as they generally assume that the background is stationary or approximately stationary to facilitate pixel-wise background modeling. However, this assumption cannot be satisfied in PTZ videos.

8) Other Challenges: Some challenges, such as different weather conditions, viewing angles, and focal length changes, have been introduced in Section III-A. There are also some interesting challenges in AGVS, such as the water mist stirring by the aircraft engine, shown in Fig. 6(a). Some extreme cases like midnight, self-luminescence, and strong reflection (Fig. 6(b) to Fig. 6(d)) are not included in AGVS, because they are too difficult for current change detection methods. However, six such videos without ground truth (*V1*∼*V6*) are also available on the AGVS website for interested readers.

IV. Experiments

In this section, we test the algorithms introduced in Section II on AGVS and discuss how to develop change detection algorithms for airport ground surveillance. A total of 21 algorithms with public codes are tested: KDE [19], SuBSENSE [20], ViBe [21], PBAS [22], SOBS [23], GMM [25], KNN [26], FGMM [27], Bodids [28], EigenSpace [29], GoDec [30], GRASTA [31], GOSUS [32], MEDRoP [33], IMTSL [34], CodeBook [35], FgSegNet [36], FgSegNet2 [37], CascadedCNN [38], SegFlow [39], and RGMP [41]. The first 16 algorithms are unsupervised, and the last five are supervised.

A. Experimental Settings

Because unsupervised and supervised change detection are two different solutions, it is unfair to compare them together. For unsupervised algorithms, if the algorithm parameters are fixed, the experimental results are also fixed. For supervised algorithms, using different training samples results in completely different experimental results. Therefore, our strategy is to test unsupervised and supervised methods separately for each challenging problem and then discuss the two solutions together at the end of Section IV.

For each unsupervised method, we use the recommended parameters in public codes, and then the algorithm is applied to all videos in AGVS without parameter tuning. In particular, because PCA/RPCA methods need to load multiple frames simultaneously for matrix decomposition, we first reduce the image resolution before testing. For GoDec [30], GRASTA [31], and GOSUS [32], it is reduced to $320 * 240$; for MEDRoP [33], it is $128 * 96$; and for IMTSL [34], it is $160 * 120$. As for the parameter setting of the five supervised algorithms, we find that the performance when using the original supervised method directly on AGVS is poor. Therefore, for each sequence in AGVS, we fine-tune the original supervised model based on the corresponding ground truth to obtain a scene-specific model before testing on the same sequence. We choose one ground truth frame every 20 frames for the fine-tuning operation. Note that such scene-specific models essentially are over-fitting. The training process of the five supervised algorithms is briefly described as follows.

The train stage of FgSegNet [36] and FgSegNet2 [37] is the same: The network is first initialized with the weights trained on CDnet2014 [3], and then in the process of fine-tuning on AGVS, the RMSProP optimizer is used for updating the parameters with a small learning rate of 1e-4. For cascaded-CNN [38], we also use the pre-trained model on CDnet2014 and then use the Adadelta optimization method to update the weights with an initial learning rate of 1e-2. As for SegFlow [39], owing to the infeasibility of creating optical flow ground truths, offline training is only applied on the segmentation branch. After initializing the two branches utilizing the weights from ResNet101 [48] and FlowNetS [40], respectively, we update the weights of the segmentation branch using AGVS, with the optical flow branch frozen at the same time. Moreover, the network is trained with the SGD optimizer,

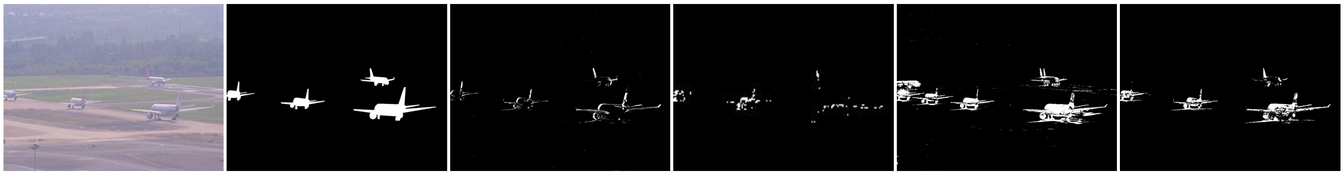


Fig. 7. Left to right: a frame in *S22*, ground truth, GMM [25], Bodids [28], CodeBook [35], EigenSpace [29], respectively.

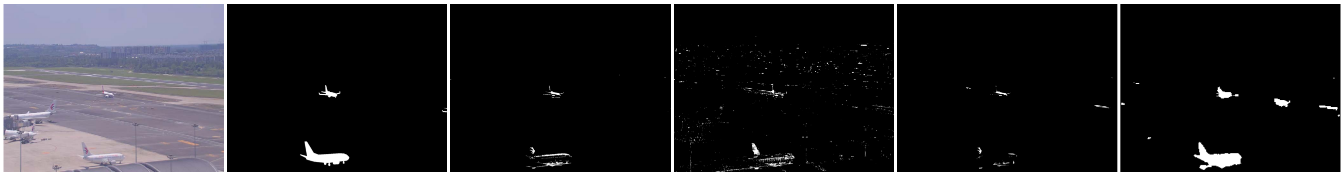


Fig. 8. Left to right: a frame in *S20*, ground truth, ViBe [21], KDE [19], SOBS [23], GoDec [30], respectively.

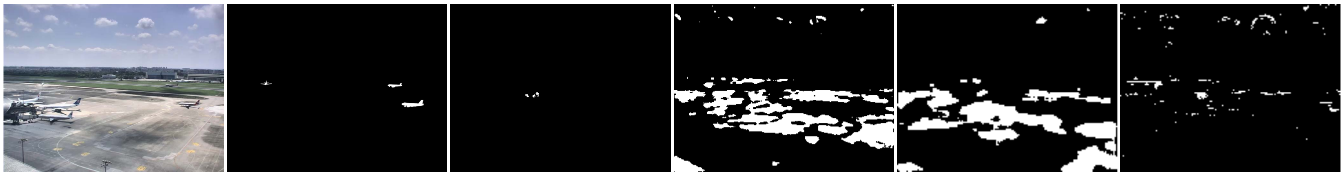


Fig. 9. Left to right: a frame in *S4*, ground truth, GRASTA [31], GOSUS [32], MEDRoP [33], IMTSL [34], respectively.

starting from the learning rate 1e-8. As for the optimization of RGMP [41], we fine-tune the network with AGVS after pre-training on DAVIS-2017 [49] dataset, but without the first stage of training on the simulated samples described in [41]. During the experiments, we use the SGD optimizer for all sequences in AGVS with a fixed learning rate of 1e-5.

B. Qualitative Analysis

The detection results for each challenge are shown in this section. As all videos in AGVS contain multiple challenges, we try to choose video frames dominated by a certain challenge for the experiment. However, although the detection performance mainly depends on the dominant challenge, it is actually the result of multiple challenges acting together to varying degrees. We cannot show the visual results of all algorithms at the same time, so only representative detection results are chosen for demonstration.

1) Unsupervised Methods: The detection result of unsupervised change detection for *haze* is shown in Fig. 1(a), where all moving objects are misdetected by SuBSENSE [20]. Note that in addition to haze, the small target is also a major challenge, as shown in Fig. 1(a). Another example is shown in Fig. 7, where haze is the most dominant challenge. It can be seen that all four algorithms, GMM [25], Bodids [28], CodeBook [35], and EigenSpace [29], have serious detection defects. The experimental results of other unsupervised algorithms on haze videos are similar to those shown in Fig. 7. Fig. 8 shows a typical *camouflage* scenario where two white aircraft are moving on the gray-white ground. We can see that there are significant detection defects by ViBe [21], KDE [19], SOBS [23], and GoDec [30]. Other unsupervised algorithms have similar results. Note that PCA/RPCA-based methods, that is, EigenSpace [29] and GoDec [30], have better detection integrity than statistical methods in the two figures, which is the advantage of this type of algorithm. However, the shortcoming of PCA/RPCA methods is also obvious; that is, stationary aircraft and ghost are also detected as the foreground. In contrast, the ability to distinguish between motion and stillness is the advantage of statistical modeling.

The detection result of unsupervised change detection for *simultaneous multi-scale detection* is shown in Fig. 1(c), where the nearest aircraft is split, and the farthest aircraft is removed as shadow. It is difficult to take into account the objects of all scales simultaneously. For example, if post-processing such as morphology is used to deal with the split of the large aircraft, the distant target may be severely deformed at the same time. Simultaneous multi-scale objects can be seen in most videos in AGVS, and this often appears at the same time as other challenges, such as in Fig. 1(a), with haze, and in Fig. 8, with camouflage. *Shadow* and *non-uniform illumination change* are shown in Fig. 1(d), and there is a lot of detection noise by FGMM [27]. The performance of other statistical methods is similar to that of FGMM. The detection results of PCA/RPCA methods under *non-uniform illumination change* are shown in Fig. 9. Although such methods are known for their robustness against illumination, we can see that non-uniform illumination change is still a severe challenge.

Shape and color variation of aircraft can be seen in all the above figures. As the unsupervised method does not require training, it essentially has good generalization. Therefore, the unsupervised change detection does not need to take special consideration of shape and color variation. The above figures also show that the *strip-shaped* fuselage and wings have serious detection defects. Almost no aircraft's fuselage and

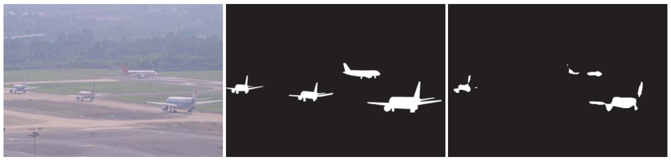


Fig. 10. Left to right: *S22*, ground truth, SegFlow [39].

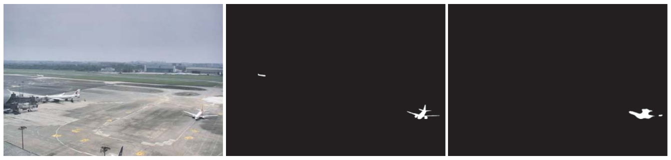


Fig. 11. Left to right: *S7*, ground truth, RGMP [41].

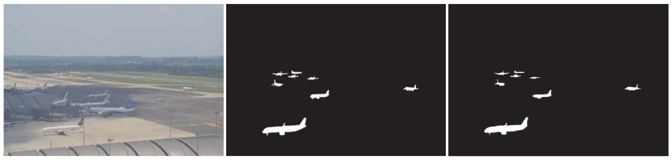


Fig. 12. Left to right: *S17*, ground truth, FgSegNet [36].

wings are completely detected by any unsupervised algorithm. One solution is to impose shape constraints on the aircraft within the unsupervised framework. However, the consequence is losing generalization and making the algorithm sensitive to shape and color variation. The detection result for *PTZ camera* is not shown, because the unsupervised algorithm is completely invalid for this problem.

2) Supervised Methods: Fig. 10 shows the detection result of SegFlow [39] under *haze*, and the result by RGMP [41] under *camouflage* is shown in Fig. 1(b). It can be seen that in both cases, many foreground pixels are misclassified by supervised algorithms. As for *simultaneous multi-scale detection*, the result of RGMP [41] is shown in Fig. 11. We can see that the small aircraft is missed, and some parts of the large aircraft are defective or broken, so the multi-scale detection problem also occurs in supervised change detection. As for *shadow and non-uniform illumination change*, the detection results by CascadedCNN [38], FgSegNet [36], RGMP [41], and SegFlow [39] are shown in Fig. 12. This experiment reflects that the supervised strategy is robust against shadows and illumination Change, which is one of the main advantages of supervised change detection. This benefits from the training strategy of the supervised solution. If the training samples are accurate enough (e.g., no shadow pixels), and the training samples cover the whole process of illumination change, the supervised method is robust against shadows and illumination change.

As for the *shape and color variation* of aircraft, a new experiment is conducted, as shown in Fig. 13. The first five sequences (*S1*∼*S5*) are used for training, and other sequences are used to test the supervised algorithms. With such an experimental setting, the testing images in *S9* and *S14* have new shape and color patterns that are not included in the training samples. We can see that the performance of the supervised methods in Fig. 13 is worse than that of the unsupervised method ViBe [21]. The outline of the detected aircraft by supervised algorithms is indistinguishable. This suggests a serious problem with the generalization of supervised change detection. In current literature, this is often called the unseen video problem [51], which means new patterns shown in the testing videos are unseen in the training videos. In fact, the unseen degree in Fig. 13 is not serious, because the airport ground appears in both training videos and testing videos, so it is not unseen. Even if the aircraft in the testing videos is unseen, it still has many similarities with the aircraft in the training videos. However, when the unseen degree is not serious, the performance in Fig. 13 is already poor, indicating that the unseen video problem is a severe challenge for supervised change detection. The unseen video issue has been a research hotspot. For example, data-level and algorithm-level solutions to unseen videos were presented in [52] and [53] for change detection, respectively.

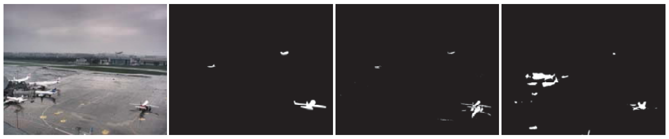


Fig. 13. Left to right: *S9*, ground truth, ViBe [21], CascadedCNN [38].

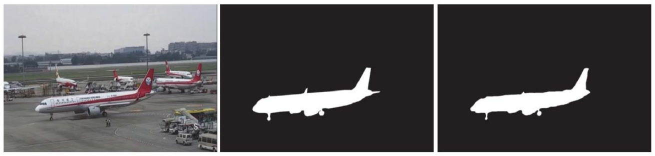


Fig. 14. Left to right: *S25*, ground truth, FgSegNet2 [37].

The above figures also show that the supervised change detection is not robust against *strip-shaped* fuselages and wings, but it is relatively better than unsupervised algorithms. The detection result for *PTZ camera* is shown in Fig. 14, which indicates that the supervised change detection is effective for this problem. Only limited detection results are shown in the above examples. In order to obtain a more intuitive impression of change detection in airport ground surveillance, we strongly recommend downloading the whole AGVS dataset and testing it with public code.

C. Quantitative Analysis

We choose *RE*, *PR* and *F-Measure* (*FM*) for quantitative comparison of detection accuracy:

$$RE = \frac{TP}{TP + FN}, \tag{1}$$

$$PR = \frac{TP}{TP + FP}, \tag{2}$$

$$FM = \frac{2 \times RE \times PR}{RE + PR}, \tag{3}$$

where *TP*, *FP* and *FN* are the numbers of true positives, false positives, and false negatives, respectively. A higher *RE* means fewer detection defects, and a higher *PR* indicates fewer detection noises. *FM* is the comprehensive result of *RE* and

TABLE I

PERFORMANCE COMPARISON OF TOP4 UNSUPERVISED ALGORITHMS AND TOP2 SUPERVISED ALGORITHMS ON AGVS

Sequences	SuBSENSE[20]			ViBe[21]			KNN[26]			CodeBook[35]			CascadedCNN[38]			FgSegNet[36]		
	RE	*PR*	*FM*	*RE*	*PR*	*FM*	*RE*	*PR*	*FM*	*RE*	*PR*	*FM*	*RE*	*PR*	*FM*	*RE*	*PR*	*FM*
S1	0.55	0.54	0.55	0.33	0.43	0.37	0.35	0.43	0.39	0.61	0.43	0.50	**0.97**	0.70	**0.82**	0.88	**0.74**	0.81
S2	0.73	0.68	0.71	0.44	0.66	0.53	0.43	0.36	0.39	0.70	0.03	0.06	**0.98**	0.87	0.92	0.96	**0.97**	**0.97**
S3	0.70	0.74	0.72	0.47	0.74	0.58	0.32	0.61	0.42	0.60	0.10	0.18	**0.99**	0.97	**0.98**	0.98	**0.98**	**0.98**
S4	0.43	0.60	0.50	0.29	0.46	0.36	0.19	0.10	0.13	0.49	0.06	0.10	**0.99**	0.95	**0.97**	0.94	**0.97**	0.96
S5	0.62	0.35	0.44	0.48	0.45	0.46	0.42	0.41	0.42	0.71	0.28	0.40	**0.99**	0.87	0.92	0.96	**0.96**	**0.96**
S6	0.69	0.63	0.66	0.57	0.58	0.58	0.52	0.58	0.55	0.74	0.51	0.61	**0.94**	0.85	0.89	0.90	**0.94**	**0.92**
S7	0.73	0.51	0.60	0.55	0.62	0.59	0.56	0.66	0.60	0.73	0.18	0.29	0.97	0.81	0.88	**0.98**	**0.92**	**0.95**
S8	0.64	0.46	0.54	0.61	0.51	0.56	0.47	0.50	0.49	0.72	0.33	0.45	**0.96**	0.71	0.82	0.89	**0.88**	**0.88**
S9	0.61	0.50	0.55	0.58	0.57	0.57	0.45	0.54	0.49	0.65	0.19	0.30	**0.97**	0.85	0.91	0.94	**0.96**	**0.95**
S10	0.52	0.74	0.61	0.53	0.63	0.58	0.57	0.62	0.60	0.77	0.43	0.55	**0.98**	0.85	0.91	0.89	**0.95**	**0.92**
S11	0.47	0.71	0.56	0.44	0.61	0.51	0.33	0.58	0.42	0.58	0.08	0.14	**0.97**	0.91	0.94	**0.97**	**0.93**	**0.95**
S12	0.44	0.57	0.49	0.50	0.50	0.50	0.58	0.51	0.54	0.73	0.41	0.53	**0.96**	0.88	**0.91**	0.88	**0.92**	0.90
S13	0.69	0.66	0.68	0.61	0.60	0.61	0.56	0.60	0.58	0.73	0.26	0.38	**0.99**	0.82	0.90	0.98	**0.89**	0.93
S14	0.35	0.46	0.40	0.43	0.53	0.47	0.27	0.57	0.36	0.44	0.42	0.43	**0.98**	0.83	0.90	0.89	**0.98**	**0.93**
S15	0.64	0.59	0.61	0.44	0.46	0.45	0.47	0.40	0.44	0.77	0.38	0.51	**0.99**	0.95	0.97	0.97	**0.98**	**0.98**
S16	0.52	0.87	0.65	0.22	0.91	0.36	0.26	0.90	0.40	0.54	0.71	0.61	**0.99**	0.93	0.96	0.97	**0.97**	**0.97**
S17	0.24	0.89	0.38	0.09	0.83	0.16	0.14	0.47	0.22	0.37	0.14	0.20	**0.97**	0.91	0.94	**0.97**	**0.96**	**0.97**
S18	0.01	0.32	0.01	0.02	0.67	0.03	0.05	0.46	0.10	0.19	0.20	0.19	**0.97**	0.77	0.86	0.94	**0.87**	**0.90**
S19	0.42	0.79	0.55	0.29	0.83	0.43	0.23	0.81	0.36	0.47	0.59	0.52	**0.99**	0.95	**0.97**	0.98	**0.97**	**0.97**
S20	0.58	0.77	0.66	0.30	0.84	0.45	0.47	0.80	0.60	0.72	0.72	0.72	**0.98**	0.96	**0.97**	0.96	**0.97**	0.96
S21	0.56	0.73	0.63	0.33	0.64	0.44	0.45	0.63	0.52	0.70	0.47	0.56	**0.99**	0.87	0.93	0.94	**0.96**	**0.95**
S22	0.58	0.65	0.62	0.14	0.84	0.24	0.22	0.81	0.35	0.45	0.63	0.52	**0.97**	0.77	0.86	0.96	**0.87**	**0.91**
S23													**0.99**	**0.99**	**0.99**	**0.99**	0.99	**0.99**
S24													**0.99**	**0.99**	**0.99**	**0.99**	0.99	**0.99**
S25													**0.99**	0.97	0.98	**0.99**	**0.99**	**0.99**

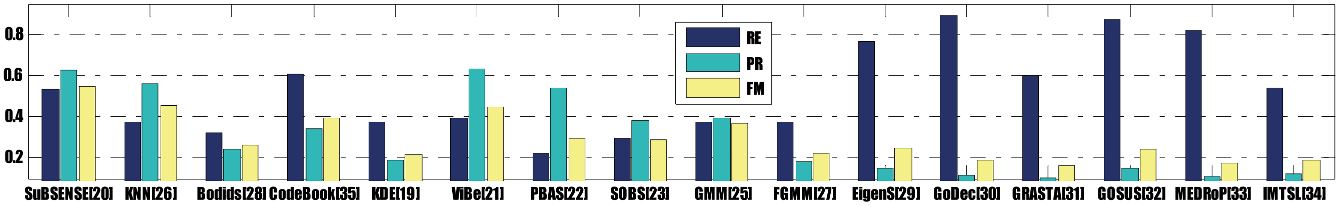


Fig. 15. Performance comparison of all unsupervised algorithms on AGVS.

PR. When both *RE* and *PR* are close to 1, *FM* is also close to 1. When either *RE* or *PR* deteriorates, *FM* also decreases.

The average *RE*, *PR*, and *FM* of all unsupervised methods on the AGVS dataset are shown in Fig. 15. PTZ videos (S23∼S25) are not considered, as unsupervised algorithms cannot deal with this problem. We can see that the performance of statistical modeling is better than that of PCA-/RPCA-based methods, and SuBSENSE [20], ViBe [21], KNN [26], CodeBook [35], and GMM [25] are the top five unsupervised algorithms in terms of *FM*. In fact, the *RE* of PCA-/RPCA-based modeling is good, but the overall performance is lowered because of the poor *PR*. The reason is that PCA-/RPCA-based methods cannot distinguish the moving and stationary targets well, resulting in a large number of false alarms and low *PR*.

For most testing algorithms in this section, the detection results on CDnet series are reported on the homepage of CDnet2014 [3]. By comparing the results on the two datasets, we find that the average detection accuracy in terms of *FM* decreases by about 25% from CDnet2014 to AGVS. Considering that CDnet2014 is the most widely used benchmark in fundamental research, we think that the gap in change detection between fundamental research and airport ground surveillance is 25% detection accuracy.

Next, we compute the average *RE*, *PR*, and *FM* of the top five unsupervised algorithms and all supervised algorithms on AGVS, as shown in Fig. 16. We can see that the performance of supervised change detection far exceeds that of unsupervised change detection. Most supervised algorithms have a *FM* above 0.8, but none of the unsupervised algorithms have a *FM* above 0.6. Furthermore, Table I shows the average *RE*, *PR*, and *FM* of the top four unsupervised algorithms and top two supervised algorithms on each sequence of the AGVS dataset. In Table I, the performance of SuBSENSE is due to the spatiotempral binary feature, and spatial coherence is considered in the model update strategy of ViBe, which makes ViBe robust to isolated detection noise. As a GMM-type model, the number of Gaussian components in KNN is adaptive, and the model in CodeBook can reflects some

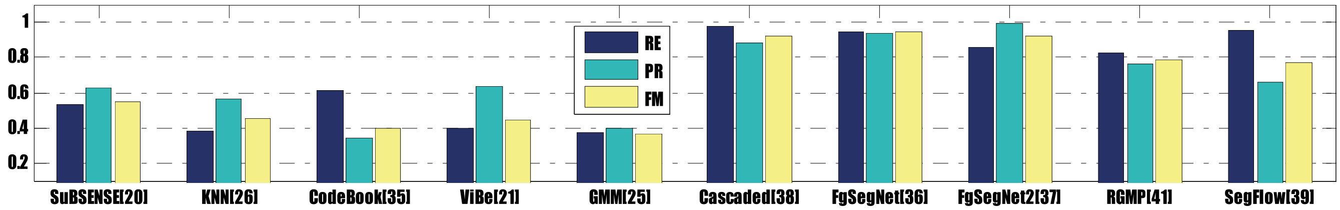


Fig. 16. Performance comparison of the top5 unsupervised algorithms and all supervised algorithms on AGVS.

TABLE II

COMPLEXITY COMPARISON IN TERMS OF FPS ON AGVS

Algorithm	SuBSENSE[20]	ViBe[21]	KNN[26]	CodeBook[35]	Bodids[28]	KDE[19]	PBAS[22]	FGMM[27]	GMM[25]	EigenS[29]
FPS	1.62	33.7	24.2	34.1	13	12.2	42.6	20.8	25.6	23.2
Algorithm	GoDec[30]	GRASTA[31]	GOSUS[32]	MEDRoP[33]	IMTSL[34]	Cascaded[38]	FgSegNet[36]	FgSegNet2[37]	RGMP[41]	SegFlow[39]
FPS	8.5	10.1	0.3	1.2	0.15	0.93	1.3	1.2	8.5	0.12

structural information. CascadedCNN has a basic model and a cascade structure. The output of the basic model is refined in the cascade structure. FgSegNet is successful because it can extract multi-scale features. The *RE* of CascadedCNN is better than that of FgSegNet, because FgSegNet is the basic model in CascadedCNN. However, the *PR* of CascadedCNN is not as good as that of FgSegNet. This may be because when the cascade structure of CascadedCNN improves the foreground detection rate, the false alarm is also enhanced.

Because of the excellent performance of the supervised algorithm, some may speculate that the unsupervised algorithm has lost its research significance. However, this is not the case for two reasons. First, as mentioned earlier, the comparison in Fig. 16 and Table I is unfair for unsupervised and supervised change detection. If we use different training strategies for supervised change detection, just as shown in Fig. 13, we obtain completely different detection results. Second, and most importantly, we cannot obtain the ground truth in advance in practical application. The ground-truth images are required for training or fine-tuning of supervised algorithms in our experiments. In practical application, especially real-time application, this training strategy is not feasible. In this case, the unsupervised change detection is still preferred in real application, at least for now.

Regarding the computation complexity, we compute the frame per second (FPS) of comparison algorithms, as shown in Table II. It can be found that the unsupervised method is much faster than the supervised method from the perspective of FPS. In our experiments, the platform for the unsupervised method is a PC with Intel i5-11600KF CPU and 32-GB RAM, while a single NVIDIA-GTX 1080Ti GPU is added for the supervised method. Therefore, the hardware cost of the supervised algorithm is much higher than that of the unsupervised algorithm. Regarding the space complexity, the supervised algorithm requires more memory to store model parameters than the unsupervised method.

D. Additional Discussion

The above experiments indicate that both unsupervised and supervised change detection encounter multiple challenges in airport ground surveillance. However, the challenges they face differ:

- Almost all issues mentioned in Section III pose severe challenges to the unsupervised change detection except for *shape and color variation*;
- Some challenges still exist for supervised change detection, such as *haze* and *camouflage*, while some other issues almost have no effect on supervised algorithms when training can be guaranteed, such as *shadows*;
- *Shape and color variation*, which relates to training, is a severe challenge for supervised change detection.

Given that the existing methods cannot solve the above problems well, how should we design application-oriented change detection algorithms? Considering that there are multiple challenges in airport ground surveillance, it is unrealistic to expect a single algorithm to solve all problems. We believe that the reasonable solution is ensemble learning [50]. Each classifier in the ensemble system tackles a special challenge, and then the decisions made by each classifier are combined by a certain combination rule. This way, all challenges are considered, and the advantages of various algorithms, such as the sensitivity to movement of statistical methods, the robustness of RPCA approaches to illumination change, and the detection integrity of the supervised algorithms, can be combined to obtain better results. There are two key issues in this strategy. One is to design an algorithm for a specific challenge, and the other is a combination rule of each algorithm.

The detection integrity of supervised change detection on seen videos is impressive. However, in our experiments, except in Fig. 13, the ground truth of the video to be tested is required for training, which cannot be satisfied in practical application. In order to make use of supervised algorithms in practical application, it is necessary to improve their generalization performance so that they can work on unseen videos. Because the background scenario is fixed in airport ground surveillance, we have a new way to improve the generalization of supervised change detection by the use of the unique prior information in airport ground. For example, the aircraft can only travel along a fixed route on the ground, the airport ground has a unique color distribution, an electronic map of the airport may be available, and there may be data sources other than videos.

Note that because there is only a single scene on the airport ground, AGVS can be used to develop application-oriented algorithms or evaluate the performance of fundamental research algorithms in real applications, but it is not suitable for training fundamental research algorithms.

V. Conclusion

Given the gap in change detection between fundamental research and airport ground surveillance, in this paper, we have presented a new change detection benchmark, AGVS, and identified the principles of how to develop specific change detection algorithms for airport ground surveillance. AGVS includes 25 long videos, amounting to about 100000 frames and accurate pixel-wise ground truth. AGVS contains various challenges on the airport ground, such as haze, camouflage, simultaneous multi-scale detection, shadows and non-uniform illumination change, shape and color variation, strip shape, PTZ camera and some other problems. Some challenges are unique to the airport ground, and some are common but have special manifestations in airport ground. After the comparison experiments of 21 state-of-the-art algorithms on AGVS, we have concluded that the solutions for change detection in airport ground are ensemble learning or making use of prior knowledge of the airport ground to improve the generalization. Furthermore, we believe that the gap between fundamental research and real application is widespread, so there should be a proprietary dataset for each application scenario, and new algorithms should be developed on this basis.

References

[1] P. Blauensteiner and M. Kampel, "Visual surveillance of an airport's apron: An overview of the AVITRACK project," in *Proc. Workshop Austral. Assoc. Pattern Recognit.*, Jun. 2004, pp. 213–220.

[2] B. Garcia-Garcia, T. Bouwmans, and A. J. R. Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Comput. Sci. Rev.*, vol. 35, Feb. 2020, Art. no. 100204.

[3] *Change Detection*. Accessed: 2014. [Online]. Available: https://www.changedetection.net

[4] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*.

[5] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, Mar. 2019.

[6] *Background Subtraction Website*. Accessed: 2014. [Online]. Available: https://sites.google.com/site/backgroundsubtraction/overview

[7] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Jun. 1999, pp. 255–261.

[8] S. D. Roy, M. K. Bhowmik, and J. Oakley, "A ground truth annotated video dataset for moving object detection in degraded atmospheric outdoor scenes," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1318–1322.

[9] L. Patino, T. Nawaz, T. Cane, and J. Ferryman, "PETS 2017: Dataset and challenge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2126–2132.

[10] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2011, pp. 1937–1944.

[11] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequievere, "A benchmark dataset for outdoor foreground/background extraction," in *Proc. Asian Conf. Comput. Vis. Workshops*, Nov. 2012, pp. 291–300.

[12] G. Yao, T. Lei, J. Zhong, P. Jiang, and W. Jia, "Comparative evaluation of background subtraction algorithms in remote scene video captured by MWIR sensors," *Sensors*, vol. 17, no. 9, pp. 1945–1975, Sep. 2017.

[13] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, "Weighted low-rank decomposition for robust grayscale-thermal foreground detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 725–738, Apr. 2017.

[14] L. Maddalena and A. Petrosino, "Background subtraction for moving object detection in RGBD data: A survey," *J. Imag.*, vol. 4, no. 5, pp. 725–738, May 2018.

[15] G. Moyà-Alcover, A. Elgammal, A. Jaume-I-Capó, and J. Varona, "Modeling depth for nonparametric foreground segmentation using RGBD devices," *Pattern Recognit. Lett.*, vol. 96, pp. 76–85, Sep. 2017.

[16] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 4–27, Mar. 2010.

[17] D. D. Bloisi, L. Iocchi, A. Pennisi, and L. Tombolini, "ARGOS-Venice boat classification," in *Proc. 12th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2015, pp. 1–6.

[18] T. Bouwmans, F. E. Baf, and B. Vachon, "Statistical background modeling for foreground detection: A survey," in *Handbook of Pattern Recognition and Computer Vision*. Singapore: World Scientific, 2010, pp. 181–199.

[19] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.

[20] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.

[21] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[22] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 38–43.

[23] L. Maddalena and A. Petrosino, "The SOBS algorithm: What are the limits?" in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 21–26.

[24] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1999, pp. 246–252.

[25] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[26] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 773–780, May 2006.

[27] F. E. Baf, T. Bouwmans, and B. Vachon, "Type-2 fuzzy mixture of Gaussian model: Application to background modeling," in *Proc. Int. Symp. Vis. Comput.*, Dec. 2008, pp. 772–781.

[28] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1778–1792, Nov. 2005.

[29] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[30] T. Zhou and D. Tao, "GoDec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2011, pp. 33–40.

[31] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1568–1575.

[32] J. Xu, V. K. Ithapu, L. Mukherjee, J. M. Rehg, and V. Singh, "GOSUS: Grassmannian online subspace updates with structured-sparsity," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3376–3383.

[33] P. Narayanamurthy and N. Vaswani, "MEDRoP: Memory-efficient dynamic robust PCA," 2017, *arXiv:1712.06061*.

[34] A. Sobral, C. G. Baker, T. Bouwmans, and E. Zahzah, "Incremental and multi-feature tensor subspace learning applied for background modeling and subtraction," in *Proc. Int. Conf. Image Anal. Recognit.*, Oct. 2014, pp. 94–103.

[35] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imag.*, vol. 11, no. 3, pp. 172–185, Jun. 2005.

[36] L. A. Lim and H. Y. Keles, "Foreground segmentation using a triplet convolutional neural network for multiscale feature encoding," *Pattern Recognit. Lett.*, vol. 112, pp. 256–262, Jan. 2018.

[37] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," 2018, *arXiv:1808.01477*.

[38] Y. Wang, Z. Luo, and P.-M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognit. Lett.*, vol. 96, pp. 66–75, Sep. 2017.

[39] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "SegFlow: Joint learning for video object segmentation and optical flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 686–695.

[40] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[41] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7376–7385.

[42] J. H. Giraldo and T. Bouwmans, "Semi-supervised background subtraction of unseen videos: Minimization of the total variation of graph signals," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3224–3228.

[43] J. H. Giraldozuluaga, S. Javed, and T. Bouwmans, "Graph moving object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2485–2503, May 2022.

[44] T. Bouwmans, F. Porikli, B. Horferlin, and A. Vacavant, *Background Modeling and Foreground Detection for Video Surveillance*. Boca Raton, FL, USA: CRC Press, 2014.

[45] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery," *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 32–55, Jul. 2018.

[46] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *J. Neural Netw.*, vol. 117, pp. 8–66, Sep. 2019.

[47] M. Mandal and S. K. Vipparthi, "An empirical review of deep learning frameworks for change detection: Model design, experimental frameworks, challenges and research needs," *IEEE Trans. Intell. Transp. Syst.*, early access, May 19, 2021, doi: 10.1109/TITS.2021.3077883.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

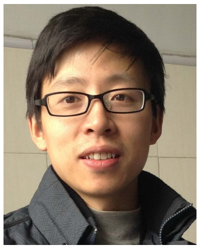
[49] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," 2017, *arXiv:1704.00675*.

[50] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 241–258, Apr. 2020.

[51] M. O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-Net: A fully-convolutional neural network for background subtraction of unseen videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2774–2783.

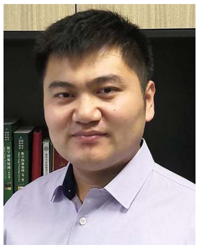
[52] M. O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-Net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction," *IEEE Access*, vol. 9, pp. 53849–53860, 2021.

[53] M. Mandal, V. Dhar, A. Mishra, S. K. Vipparthi, and M. Abdel-Mottaleb, "3DCD: Scene independent end-to-end spatiotemporal feature learning framework for change detection in unseen videos," *IEEE Trans. Image Process.*, vol. 30, pp. 546–558, 2021.



Xiang Zhang received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2003 and 2006, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2010. He is currently an Associate Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. His research interests include intelligent transportation, video analysis, and machine learning.

Chang Shu received the Ph.D. degree from Tsinghua University, in 2011. He is currently with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, as a Lecturer. His research interests include pattern recognition, biometrics, and computer vision. He was a coauthor of the Best 10% Paper Award from VCIP 2016.



Shuai Li (Member, IEEE) received the Ph.D. degree from the University of Wollongong, Australia, in 2018. From 2018 to 2010, he was with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, as an Associate Professor. He is currently with the School of Control Science and Engineering, Shandong University, China, as a Professor and a QiLu Young Scholar. His research interests include image/video coding, 3D video processing, and computer vision. He was a co-recipient of Two Best Paper Awards from the IEEE BMSB 2014 and IIH-MSP 2013, respectively.

Celimuge Wu (Senior Member, IEEE) received the Ph.D. degree from The University of Electro-Communications, Tokyo, Japan, in 2010. He is currently a Professor with The University of Electro-Communications. His research interests include vehicular networks, the Internet-of-Things, edge computing, and application of machine learning in wireless networking and computing. He serves as an Associate Editor for IEEE Transactions on Network Science and Engineering, IEEE Transactions on Green Communications and Networking, and IEEE Open Journal of the Computer Society. He was a recipient of the IEEE Communications Society Outstanding Paper Award in 2021, the IEEE Internet of Things Journal Best Paper Award in 2021, the IEEE Computer Society 2020 Best Paper Award, and the IEEE Computer Society 2019 Best Paper Award Runner-Up.

Zhi Liu (Senior Member, IEEE) received the Ph.D. degree in informatics from the National Institute of Informatics. He is currently an Associate Professor at The University of Electro-Communications. His research interests include video network transmission and mobile edge computing. He is currently an Editorial Board Member of Springer wireless networks and IEEE Open Journal of the Computer Society.