2022 IEEE/SICE International Symposium on System Integration (SII)
January 9-12, 2022. Narvik, Norway

2022 IEEE/SICE International Symposium on System Integration (SII) | 978-1-6654-4540-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/SII52469.2022.9708745

Unsupervised Video Anomaly Detection in Traffic and Crowded Scenes

Satoshi Hashimoto[1], Alessandro Moro[2], Kenichi Kudo[1], Takayuki Takahashi[3], and Kazunori Umeda[1]

***Abstract*— In this paper, we propose a scene-independent robust unsupervised video anomaly detection method based on future frame prediction as a breakthrough and better video anomaly detection technique. Most conventional methods evaluate and develop a static camera and dashcam approach as independent tasks, and no method has been proposed that is independent of the capture conditions. The proposed method introduces a frame-wide future prediction–based spatio-temporal adversarial networks that can handle arbitrary series lengths to cope with various scenes. The noise in the prediction error caused by constant background changes is improved by weighting the regions of interest for the discriminator of the generative adversarial networks (GANs). This framework can be applied to all cases regardless of the scene environment. Experiments on public datasets of general traffic scenes and crowded scenes confirm the superiority of the proposed method over current state-of-the-art methods.**

I. Introduction

Research on advanced driver assistance systems for the establishment of automated driving technology is being actively conducted. Automated driving is an important technology that provides a safe and comfortable driving experience. On the other hand, the development of technology to avoid accidents associated with independent driving is inevitable in establishing a safe system. Since there are an infinite number of scenarios of abnormalities that can occur in everyday life, including traffic scenes, a framework that can detect all kinds of abnormalities should ideally be established. In recent years, several in-vehicle accident warning systems have been tried to reduce the risk, but their detection has been limited to certain modes [1]. On the other hand, a few data-driven methods that use dashcam video images, which can be installed inexpensively, have also appeared [1–3]. Chan et al. [2] proposed a Dynamic-Spatial-Attention Recurrent Neural Network (DSA-RNN) as a supervised approach for dashcam video images. However, the supervised approach requires annotations on a huge number of datasets, and the patterns of anomalies that can be identified are limited. Yao et al. [3] proposed an unsupervised approach that is independent of specific anomaly modes and trained the model using only normal data that can be collected in large amounts. Specifically, anomaly detection is performed based on whether there is a discrepancy between the prediction results of a model trained only on normal data and the actual observed data. In all fields of anomaly detection, the above approach is often taken due to the difficulty of collecting supervised data. The advantage of this approach is that the pattern of anomalies that can be detected does not depend on the supervised data, as compared to supervised methods. Unfortunately, their method [3] is not sufficiently practical in terms of accuracy.

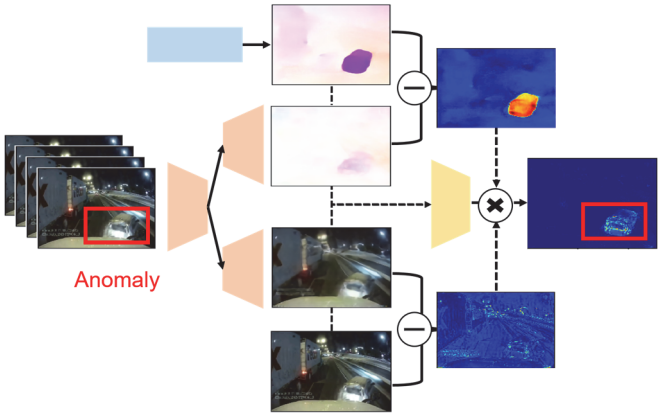

Fig. 1 An overview of our method. The model predicts the future image one frame ahead of the input video and the optical flow. Anomalies are detected by calculating the difference between the prediction and the ground truth. Since the model is trained only with normal data, the error for abnormal regions is large.

On the other hand, there has been a lot of research on video anomaly detection in surveillance camera images in daily life [4–18]. These studies all pertain to crowded scenes captured by surveillance cameras for statics. An unsupervised approach using the reconstruction (prediction) error of the video image has been successful here, and Luo et al. [8] proposed a reconstruction-based method using an Autoencoder-like spatio-temporal network for video images. Ravanbakhsh et al. [13] proposed a method that compresses the spatio-temporal information between neighboring image frames in a video into an optical flow and attributes it to image-based anomaly detection. Liu et al. [16] provided a new baseline for predicting future image frames for video inputs. However, these methods implicitly assume that the data are captured by a static camera, which is highly unsuitable for dashcam videos where the entire background of the image changes constantly over time. In particular, the constant background change has a negative noise effect on the prediction error of the entire image, which hinders the detection of abnormal scenes.

In summary, most conventional methods evaluate and develop a static camera and dashcam approach as independent tasks, and no robust video anomaly detection method has been proposed that is independent of the capture conditions. In this paper, we propose a scene-independent robust unsupervised video anomaly detection method based on future frame prediction as a breakthrough and better video anomaly

1 Course of Precision Engineering, School of Science and Engineering, Chuo University, Tokyo, Japan
(corresponding author: umeda@mech.chuo-u.ac.jp).

2 RITECS Inc., Tokyo, Japan
3 Prima Meat Packers, Ltd., Tokyo, Japan

detection technique. The proposed method predicts the future of the entire frame image and detects anomalies based on errors. A spatio-temporal adversarial networks is introduced to model the complex spatio-temporal information from a moving camera. The noise in the prediction error caused by constant background changes is improved by weighting the regions of interest for the discriminator of the generative adversarial networks (GANs) [19]. This framework can be applied to all cases regardless of the scene environment. The contributions of this paper are as follows.

- We propose an unsupervised video anomaly detection method that is robust to various video scenes.
- We propose a framework for focusing on anomalous regions of an image by efficiently utilizing the discriminator of GANs.

We evaluated the proposed method on a large traffic scene dataset [1] and a crowded scene dataset [7,18] of dashcams installed on trucks. As a result, we confirmed the superiority of the proposed method as compared to current state-of-the-art methods.

In the following, we first present related studies. Next, we present the framework of the proposed method. In addition, we show the validation experiments. Finally, we give the conclusion and future prospects. A schematic diagram of the proposed method is shown in Fig. 1.

II. Related Work

Video anomaly detection has attracted a great deal of attention in the fields of computer vision and robotics [20]. Prior research has been developed independently on static surveillance video for crowded scenes and dashcam-based methods for traffic scenes.

A. Video Anomaly Detection in Static Cameras

In recent years, most of static cameras train the entire frame image using deep learning models and use the reconstruction and prediction errors for unsupervised anomaly detection. The approaches are mainly classified into two categories: (1) STN-based methods [8–12] and (2) appearance and motion methods [13–17].

STN-based methods [8–12]. Spatio-temporal networks (STNs) have an encoder–decoder type of structure for reconstructing video images. STNs are trained to minimize the reconstruction errors between input and output using only normal data. Therefore, in the inference phase, the reconstruction error for normal data input is small, but for abnormal data input, the error is large because the model does not fit the normal parameters acquired during training. Luo et al. [8] proposed a method using temporally coherent stacked Recurrent Neural Network (sRNN). In recent years, GANs [19] in which the generator and discriminator learn adversarially from each other to generate high-quality images have been proposed, and they have been used in anomaly detection. Lee et al. [10] proposed spatio-temporal adversarial networks (STAN), an extension of STNs for adversarial models.

Appearance and motion method [13–17]. In contrast to STNs, methods that compress the spatio-temporal information of video images into an optical flow and attribute it to image-based anomaly detection settings have been particularly successful. In this paper, these are called appearance and motion methods. Ravanbakhsh et al. [13] uses Pix2Pix [21], an extension of GANs, to bi-directionally model the relationship between optical flow and frame images and detect anomalies based on prediction errors. Liu et al. [16] applied Pix2Pix to frame prediction, where the generator is applied as image converter to predict the future frame for the input video image. Nguyen et al. [17] also learned the transformation between frame images and optical flow but introduced the inception module [22] to help the convolutional layer learn more useful features.

B. Traffic Accident Detection in Dashcams

A few studies exist regarding traffic accident detection using dashcam data. Chan et al. [2] proposed a supervised approach using the Dynamic-Spatial-Attention Recurrent Neural Network (DSA-RNN) for the entire frame image, the appearance features of the objects obtained using the detector, and the dynamic features of the frame. However, the supervised approach has a problem, in that the number of anomaly modes that can be detected is limited. Yao et al. [3] proposed an unsupervised approach that uses only a large amount of normal data for training. They also argued that it is not necessary to accurately predict all of the information in a frame, so their approach predicts the trajectory of a specific observed object and detects anomalies based on the errors. However, limiting objects by detector may lead to a decrease in detection accuracy for unexpected anomaly scenarios, and their method does not provide practical accuracy in evaluation using datasets. Haresh et al. [1] compared and validated reconstruction-based and one-class classification–based methods and reported the effectiveness and shortcomings of the reconstruction-based method. The important point of their method is that it takes into account the relative positions of objects using graph convolutional networks (GCNs) [23]. Unfortunately, the practicality of their method has not been sufficiently confirmed in the validation.

III. Proposed Method

The goal of this paper is to develop a robust video anomaly detection method that is independent of the scene. First, we organize the issues of conventional methods by task.

Static camera method [4–18]. The static camera method implicitly assumes that the background of the video scene is stationary and simple. This assumption reduces the robustness of the method. When applying previous work to dashcam videos in Naïve, the constant changes and complexity of the entire frame may amplify the noise in the prediction error. In addition, the state-of-the-art method [17] uses only the optical flow between neighboring frames as spatio-temporal information, which is difficult to model because it cannot take into account the temporal information of the medium and long terms. In fact, the evaluation of the state-of-the-art method [16] for dashcam videos is not good [1]. In addition, most of the objects that are subject to anomaly detection in static datasets are simple objects with large visual deviations from normal. However, in view of real-world applications in society, it is necessary to achieve results not only for simple scenes but also for complex and difficult data with dynamic backgrounds.

Dashcam method [1–3]. In the dashcam method, an approach [3] is proposed to predict the trajectory of a particular object using an object detector. However, since there are innumerable anomalous scenarios, such limitation of observation targets may miss the anomalies to be identified. On the other hand, video reconstruction-based methods [1] have also been proposed, but the framework for incorporating object-based positioning has not been sufficiently successful and is not applicable to various video scenes.

Proposed method. We propose a scene-independent robust unsupervised video anomaly detection method based on future frame prediction. The proposed method introduces a frame-wide future prediction–based spatio-temporal adversarial networks that can handle arbitrary series lengths to cope with various scenes. We believe that the small amount of temporal information in the previous state-of-the-art work [17] is the reason for the lack of robustness, so we introduce a flexible design for temporal information. We also propose to weight the regions of interest of the network discriminator. This is because the discriminator is a model that discriminates between "true" or "false" in GANs and is expected to focus on anomalous regions. Since this framework can focus on anomalous regions regardless of the video characteristics, it is expected to reduce the noise in the prediction error caused by motion and background complexity. Therefore, this strategy can overcome the drawbacks of approaches that model the entire frame image. Note that since the proposed method is trained unsupervised on all normal data only, future frames in normal events can be accurately predicted, and inaccurate prediction implies an anomaly.

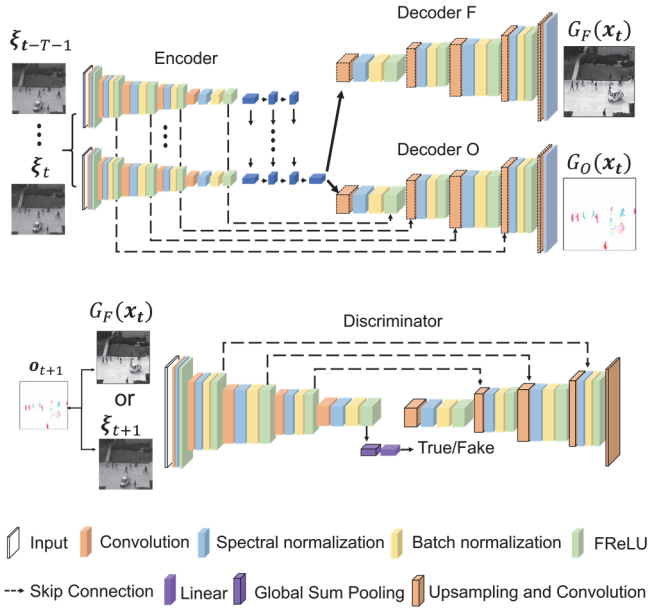

Fig. 2 Overview of our model. The upper figure shows the generator, and the lower figure shows the U-Net discriminator.

A. Our Model

The details of the model are shown in Fig. 2. Our model consists of a deep learning model with convolutional layers and four components: the Encoder, Decoder O, Decoder F, and the Discriminator. Encoder and Decoder O are defined as Generator O (G_O). Encoder and Decoder F are defined as Generator F (G_F). G_O and G_F are trained only on normal video scenes to help detect objects that deviate from normal.

Generator O (G_O). G_O is an STNs consisting of the Encoder and Decoder O. The Encoder extracts spatio-temporal features from the input video image using convolutional layers and Convolutional-LSTM [24]. Decoder O uses the features extracted by the encoder to predict the future optical flow of one frame of the video image using a deconvolution layer. Since G_O can be regarded as a kind of image-transformation task, it has a skip structure similar to that of Pix2Pix to share multi-level features between the encoder and decoder. We use optical flow for anomaly detection because we believe that the spatial information of the flow is useful and, intuitively, the prediction error of the flow forms a "blob" of anomalous regions. To the best of our knowledge, this is the first application of STNs to the task of transforming optical flows in video anomaly detection.

Generator F (G_F). G_F is an STNs consisting of the Encoder and Decoder F, which predicts the image of one frame in the future. We can choose to use U-Net for G_F instead of RNN and other spatio-temporal modules, as in Liu et al. [16]. However, we are concerned about the possibility of propagating anomalous information in the input due to the skip structure of U-Net [25]. We also believe that incorporating a module designed to handle spatio-temporal information will contribute to better modeling of video images. Therefore, we refer to the STNs model adopted by Lee et al. [10] and extend it to a predictive model. Both G_O and G_F incorporate Conv-LSTM in their encoder to help model spatio-temporal information. We introduce the frame prediction model because we believe that since anomaly detection is the identification of unexpected events, it is natural to predict future video frames from past video frames and to compare the predicted values with the ground truths for anomaly detection [16].

Discriminator. We use U-Net as the discriminator to identify at the pixel level whether the image is a "true" image sampled from the dataset or a "false" frame image predicted by the generator. Note that the Discriminator only identifies the prediction result and ground truth of the frame image. For the bottleneck of U-Net, the Discriminator is connected to all of the coupling layers and identifies true or false at the frame level. The introduction of the U-Net Discriminator was inspired by the success of the U-Net GAN proposed by Schonfeld et al. [26]. The major advantages of the U-Net Discriminator are the introduction of consistency regularization (discussed below) and the improvement of the prediction quality of the model in the framework of pixel-level truth discrimination. The choice of a better GANs model contributes to better modeling of complex and difficult datasets such as those of dashcams and other traffic scenes. In addition, since the Discriminator is trained in the framework of conditional generative adversarial networks (cGAN), following Pix2Pix, it can theoretically avoid mode collapse, which is a problem for GANs [17].

B. Training Our Model

Our spatio-temporal adversarial networks is optimized by alternately minimizing the following two losses: L_G and L_D.

$$L_G = \lambda_f L_{frame} + \lambda_o L_{opt} + L^G_{D_{enc}} + L^G_{D_{dec}} \quad (1)$$

$$L_D = L_{D_{enc}} + L_{D_{dec}} + \lambda_c L_{consist} \quad (2)$$

Here, G represents the Generator, and D_{enc} and D_{dec} represent the Encoder and Decoder modules of the Discriminator, respectively. L_{frame} and L_{opt} are the prediction loss of the frame image and the optical flow between the ground truth and the prediction result, respectively. λ_f and λ_o are the weighting constants for the prediction loss. λ_c is the weighting constant for the consistency regularization. Each term of L_G is as follows:

$$L_{frame} = \|\boldsymbol{\xi}_{t+1} - G_F(\boldsymbol{x}_t)\|_1 \quad (3)$$

$$L_{opt} = \|\boldsymbol{o}_{t+1} - G_O(\boldsymbol{x}_t)\|_1 \quad (4)$$

$$L^G_{D_{enc}} = -\mathbb{E}_{\boldsymbol{\xi}\sim\boldsymbol{p}_{\boldsymbol{\xi}}}\big[log\big(1 - D_{enc}([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}])\big)\big] - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{p}_{\boldsymbol{x}}}[log(D_{enc}([G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}]))] \quad (5)$$

$$L^G_{D_{dec}} = -\mathbb{E}_{\boldsymbol{\xi}\sim\boldsymbol{p}_{\boldsymbol{\xi}}} \sum_{i,j} log[\,1 - D_{dec}([\boldsymbol{\xi}_t, \boldsymbol{o}_{t+1}])]_{i,j}] - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{p}_{\boldsymbol{x}}} \sum_{i,j} log\,[D_{dec}([G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}])]_{i,j}] \quad (6)$$

The input $\boldsymbol{x}_t$ consists of partial time series $\boldsymbol{\xi}_{t-T-1}, \boldsymbol{\xi}_{t-T-2}, \ldots, \boldsymbol{\xi}_t$ with fixed length T at some time t. $\boldsymbol{\xi}$ is the image of each frame. $\boldsymbol{o}_{t+1}$ is the optical flow at time t+1. $[D_{dec}(\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1})]_{i,j}$ and $[D_{dec}(G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1})]_{i,j}$ represent the output result of the Discriminator at pixel (i,j) . $[\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]$ means the combination of $\boldsymbol{\xi}_{t+1}$ and $\boldsymbol{O}_{t+1}$ in the channel direction. We adopt the L1 norm for the loss L_{frame} and L_{opt} to approximate the prediction and ground truth of the frame image and optical flow. In addition, each term in L_D is as follows:

$$L_{D_{enc}} = -\mathbb{E}_{\boldsymbol{\xi}\sim\boldsymbol{p}_{\boldsymbol{\xi}}}\,[\,log(D_{enc}([\boldsymbol{\xi}_t, \boldsymbol{o}_{t+1}]))] - \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}[log(1 - D_{enc}([G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}])] \quad (7)$$

$$L_{D_{dec}} = -\mathbb{E}_{\boldsymbol{\xi}\sim\boldsymbol{p}_{\boldsymbol{\xi}}}[\sum_{i,j} log\,[D_{dec}([\boldsymbol{\xi}_t, \boldsymbol{o}_{t+1}])]_{i,j}] - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{p}_{\boldsymbol{x}}}[\sum_{i,j} log\,[1 - D_{dec}([G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}])]_{i,j}] \quad (8)$$

$$L_{consist} = ||D_{dec}\big(mix([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}], G_F(\boldsymbol{x}_t), M)\big) - mix\big(D_{dec}([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]), D_{dec}\big(G_F(\boldsymbol{x}_t)\big), M\big)||^2, \quad (9)$$

where $L_{consist}$ denotes the Cutmix [27]-based consistency regularization introduced in [26]. This regularization is based on the idea that the output from a well-trained discriminator should be equal across image classes and domain transformations, and its introduction has been shown to greatly improve the quality of generation. The mix can be calculated by Equation (10).

$$mix([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}], [G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}], M) = M \circledcirc [\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}] + (\mathbf{1} - M) \circledcirc [G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}], \quad (10)$$

where $M \in \{0,1\}^{W\times H}$ is a binary mask indicating whether the pixel (i,j) is a true image (1) or (0), $\mathbf{1}$ is a binary mask filled with 1, and $\circledcirc$ is an element-wise multiplication. Note that AdaBelief [28] is used as the optimization method for learning, and FReLU [29] is used as the activation function for the Generator and Discriminator layers. We also introduce spectral normalization [30] in each layer of the Generator and Discriminator to ensure stable learning. The ground truth of the optical flow is estimated by FlowNet2 [31], following the procedure used by Nguyen et al. [17].

C. Anomaly Detection

A model that has been trained by following the steps in the previous subsection with only normal events can accurately predict normal events. Therefore, the difference between predicted frames $G_F(\boldsymbol{x}_t)$ and $G_O(\boldsymbol{x}_t)$ and their ground truths $\boldsymbol{\xi}_{t+1}$ and $\boldsymbol{O}_{t+1}$ can be used for anomaly detection. However, in Naïve, using the difference of the entire frame can cause a lot of noise in traffic scenes with dynamic backgrounds, as described above. Therefore, we focus on the Discriminator, which has been ignored in previous studies [13,15–17], and propose to utilize it efficiently. This is because the Discriminator is a model that discriminates between "true" and "false" images, and it is expected to focus on anomalous regions as "false" during inference. Specifically, the middle layer feature $F_N([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}])$ of the Nth layer of the Discriminator for input $[\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]$, which consists of "true" images and uses the difference between the feature $F_N([G_F(\boldsymbol{x}_t), G_{\boldsymbol{o}}(\boldsymbol{o}_{t+1})])$ for the input $[G_F(\boldsymbol{x}_t), G_{\boldsymbol{o}}(\boldsymbol{o}_{t+1})]$, which consists of "false" images. Our framework incorporates a flexible design that allows us to arbitrarily choose which middle layer to use and to fuse multi-level features. The fusion of middle layer features enables anomaly detection by focusing on anomalous regions in any scene, including those from dashcam videos and static camera videos. The anomaly degree $a(\boldsymbol{x}_t)$ for input $\boldsymbol{x}_t$ to the model at each frame t is defined as in Equation (11):

$$a(\boldsymbol{x}_t) = |||\boldsymbol{\xi}_{t+1} - G_F(\boldsymbol{x}_t)| \circledcirc |\boldsymbol{o}_{t+1} - G_O(\boldsymbol{x}_t)| \circledcirc D_{dec}([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}])\boldsymbol{map}||_1 \quad (11)$$

$$\boldsymbol{map}(N, \ldots, K, L) = |F_N([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]) - F_N([G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}])| \circledcirc \cdots \circledcirc |F_M([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]) - F_M([G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}]) \circledcirc F_L([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}])|, \quad (12)$$

where the input $\boldsymbol{x}_t$ consists of partial time series $\boldsymbol{\xi}_{t-T-1}, \ldots, \boldsymbol{\xi}_t$ with fixed length T at a certain time t, where $\boldsymbol{\xi}$ and $\boldsymbol{o}$ are the image and optical flow of each frame, respectively. $\boldsymbol{map}$ is the difference between the middle layer features $F_N, .., F_K$ of any $N, \ldots, K$ layer of the U-Net Discriminator and resized to the size of the input. Note that L means the output of the last layer of the Discriminator; its size

is equal to the difference between images. Therefore, no resizing is performed. The final score, $S(\boldsymbol{x}_t)$, used in the anomaly calculation is obtained by normalizing using Equation (13), following the example of Nguyen et al. [17]:

$$S(\boldsymbol{x}_t) = \frac{a(\boldsymbol{x}_t)}{max(a(\boldsymbol{x}_1), a(\boldsymbol{x}_2), \dots, a(\boldsymbol{x}_m))}, \tag{13}$$

where m is the total number of test data. This is used for anomaly detection.

IV. Experiments

We used a dataset of crowded scenes captured by a static camera and a traffic scene captured by a dashcam with a complex dynamic background for evaluation. For evaluating congested scenes, we used UCSDped2 [7], which is a common public dataset for video anomaly detection, and Avenue [18]. An example dataset is shown in Fig. 3. UCSDped2 consists of 16 clips of training data and 12 clips of test data. As shown on the left side of Fig. 3, the normal data are from walking at a normal speed. On the other hand, the anomaly data include bicycle riding and car intrusion. Avenue consists of 16 clips of training data and 21 clips of test data. As shown in the middle of Fig. 7, the normal data include walking at a normal speed. On the other hand, anomalous data include deviations from the norm, such as running or throwing a package. Avenue has some issues, such as blurring, and outliers occur during the shooting.

To evaluate the traffic scenes, we use RetroTrucks, a large public dataset proposed by Haresh et al. [1], which consists of 136,687 training frames and 36,663 test frames; this is shown on the right side of Fig. 7. The training data contain normal driving scenes, while the test data contain various driving scenes such as collisions and near misses. The training data are used to evaluate the performance of the system. Note that the training data for all datasets used in the evaluation consist only of normal data.

Fig. 3 Examples of datasets. The upper row shows an example of normal, and the lower row shows an example of abnormal datasets. The red rectangle in the bottom row is the area of abnormal events.

A. Evaluation Metrics

For these datasets, we performed a quantitative evaluation of the area under the receiver operating characteristic curve (AUROC) model for the frame-level, following the procedure of many previous studies. Since the evaluation is done by AUROC, we do not discuss the anomaly threshold.

B. Implementation Details

The hyperparameters used in the experiments are shown below. Following the state-of-the-art method [17], the learning rates of the Generator and Discriminator were set to 2e-4 and 2e-5[17], respectively, the batch size was set to 1 and the time step T was set to 4 (UCSD and Avenue) and 16 (RetroTrucks). All frame images were resized to 256×256. The prediction loss weights λ_f and λ_o were set to 100 and 200, respectively, and the consistency regularization weight λ_c was set to 10 [26]. We used NVIDIA GeForce TITAN GPUs for the computation and PyTorch, a deep learning library, for the implementation.

C. Baselines

To quantitatively confirm the effectiveness of the proposed method, we compare it with recent deep learning–based methods [1,8,13,16,17], which are evaluated using static camera data, and the method of Haresh [1], which is evaluated using dashcam data. The methods that are closest to ours are those used by Liu et al. [16] and Nguyen et al. [17]. For details of the methods, please refer to Section II.

D. Results of UCSD and Avenue

Qualitative evaluation. Figure 4 shows the input/output of the model, the difference images, the middle layer features, and the anomaly map fused with them. Looking at the UCSD example (upper), we see that the model is unable to accurately predict the ground truth of the anomalous object, "car." Looking at the difference between the ground truth and the prediction, we can see that the values around the anomalous object are larger. The noise can be caused by small noises in the dataset or limitations in the prediction quality of the model; this noise is more pronounced for data with complex backgrounds, such as that of RetroTrucks, described below. To address this issue, our proposed framework for fusing ***map*** helps to reduce the noise in Naïve difference images and obtain anomaly maps that focus on anomalous locations.

The Avenue example in Fig. 4 (below) shows the poor prediction accuracy of the model's optical flow. In particular, the difference values are large over a wide area other than the red rectangle anomaly area, forming a "blob" noise. This may be due to the fact that the UCSD data contain pedestrians moving at almost a constant speed, while the Avenue data contain more complex and difficult movements, such as people standing still for a certain period of time. This result suggests the problem of treating the optical flow as temporal information content. However, the fusion of the middle layer features of the proposed method does not significantly affect the performance of the final anomaly map, since such noise can be removed in the final map.

Quantitative evaluation. Table 1 shows the list of AUROC scores for UCSD, Avenue, and RetroTrucks. Quantitative evaluation using AUROC confirms the effectiveness of the proposed method. Our method significantly outperforms that of Nguyen et al. [17] and other state-of-the-art methods in terms of frame-level scores. Similarly, for the static camera dataset, we observed that the performance was improved by fusing the ***map***. Note that the w/o ***map*** data in the table was computed using the difference in Naïve frame images.

TABLE 1 AUROC results for each dataset as compared with those of previous studies

Method	AUROC↑		
	UCSDped2	Avenue	RetroTrucks
Luo et al. [8]	0.922	0.817	N/A
Ravanbakhsh et al. [13]	0.935	N/A	N/A
Liu et al. [16]	0.951	0.851	0.606
Nguyen et al. [17]	0.962	0.872	N/A
Haresh et al. [1]	0.700	N/A	**0.715**
Ours w/o ***map***	0.665	0.741	0.618
Ours w/ ***map***(1)	**0.965**	0.882	0.637
Ours w/ ***map***(1,*L*)	0.957	0.891	0.661
Ours w/ ***map***(3)	0.958	**0.901**	0.646

E. RetroTrucks Results

Qualitative evaluation. In the example of RetroTrucks, shown in Fig. 5, a rear-ending car is observed as an anomalous object. As we pointed out in Section III, noise other than the car exists in the background regions in the difference image, but it can be removed by ***map*** fusion. The framework of the proposed method was found to be effective for dashcam videos. However, the prediction of the frame images in the figure shows that the proposed method predicts anomalous objects well. This is not an expected result in an unsupervised anomaly detection framework. One possible reason is that the model has acquired high expressive power for complex events through training with a huge dataset.

Quantitative evaluation. From Table 1, we confirmed the effectiveness of the proposed method for RetroTrucks by quantitative evaluation using AUROC. Our method outperforms the state-of-the-art method in terms of frame-level scores. In particular, we have quantitatively confirmed the performance improvement by integrating the middle layer feature map of the U-Net Discriminator. However, it was lower than the score of Haresh et al. [1]. This can be attributed to the fact that the generalization performance of the model of the proposed method was high as mentioned above, and the gain of the anomaly score of the anomalous data was not sufficient. Note that the results of Liu et al. [16] were referenced from Haresh et al. [1].

F. Impact of ***map***

In UCSD, our method only improved by 0.003 over the state-of-the-art method [17] in AUROC evaluation. In addition, ***map*** can achieve a high gain of about 0.1~0.3 for Naïve differences no matter which layer is selected, but depending on the selection, the results are worse than those of Nguyen et al. [17]. This suggests that the proposed method does not provide a large gain on relatively simple datasets with little noise and variation. On the other hand, Avenue improves the score of Nguyen et al. [17] by 0.029. In terms of the choice of a ***map*** layer, we obtain a better score when we fuse a more abstract layer (the third layer). This may be due to the fact that Avenue is littered with "blob" noise from optical flow prediction errors, so the deeper and more abstract layers, or "false" semantic high-level features, were more effective. In other words, the "false" images in the discriminator are semantically close to the anomaly. An example of a middle layer feature is shown in Fig. 6. This result suggests that the proposed method is effective in denoising complex data.

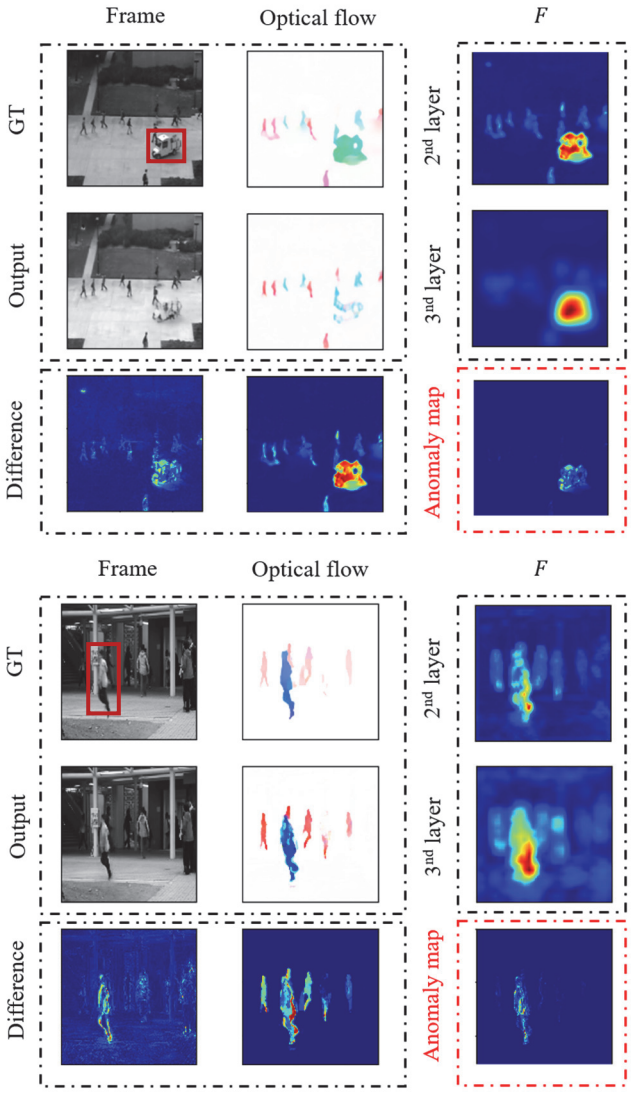

Fig. 4 Results of UCSD (upper) and Avenue (lower). The first column is the Frame image, which consists of the ground truth, the prediction result, and their difference image. The second column is the optical flow ground truth, prediction result, and their difference image. The third column shows the differences of the middle layer features of the second and third layers. The area in the red rectangle shows the final anomaly map obtained by fusing ***map***.

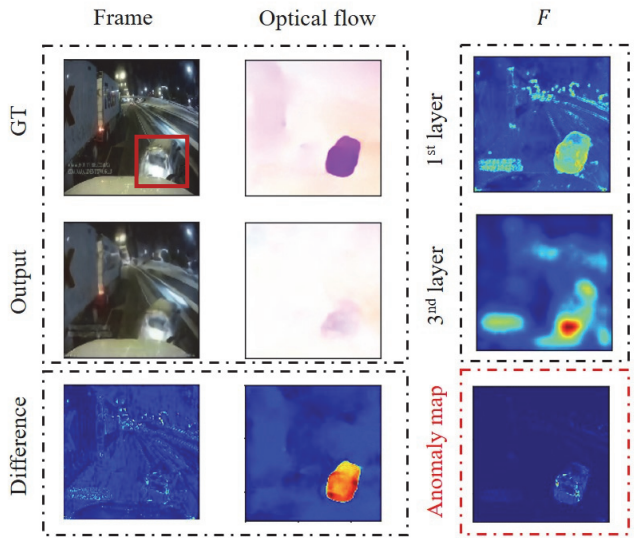

Fig. 5 RetroTrucks Results

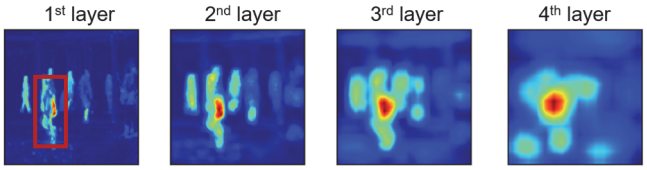

Fig. 6 Visualization of middle layer features in each layer. As the layers get deeper, the regions of interest that are "false" are extracted at a higher level.

V. Conclusion

In this paper, we proposed a scene-independent robust unsupervised video anomaly detection method based on future frame prediction. The main challenges are dealing with traffic scenes with a constantly changing background and complex data. The novelty of our method is the efficient use of generative adversarial network discriminators and the introduction of a framework to focus on anomalous regions, which improves the challenges. Experiments on public datasets of general traffic scenes and crowded scenes confirm the superiority of the proposed method over current state-of-the-art methods.

References

[1] Sanjay Haresh, et al., "Towards Anomaly Detection in Dashcam Videos," IV, 2020.
[2] Fu-Hsiang Chan, et al., "Anticipating Accidents in Dashcam Videos," ACCV, 2016.
[3] Yu Yao, et al., "Unsupervised Traffic Accident Detection in First-Person Videos," IROS, 2019.
[4] Waqas Sultani, et al., "Real-world Anomaly Detection in Surveillance Videos," CVPR, 2018.
[5] Guansong Pang, et al., "Self-trained Deep Ordinal Regression for End-to-End Video Anomaly Detection," CVPR, 2020.
[6] Mahmudul Hasan, et al., "Learning Temporal Regularity in Video Sequences," CVPR, 2016.
[7] Vijay Mahadevan, et al., "Anomaly Detection in Crowded Scenes," CVPR, 2010.
[8] Weixin Luo, et al., "A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework," ICCV, 2017.
[9] Weixin Luo, et al., "Remembering History with Convolutional LSTM for Anomaly Detection," ICME, 2017.
[10] Sangmin Lee, et al., "STAN: Spatio-Temporal Adversarial Networks for Abnormal Event Detection," ICASSP, 2018.
[11] Asim Munawar, et al., "Spatio-temporal Anomaly Detection for Industrial Robots through Prediction in Unsupervised Feature Space," WACV, 2017.
[12] Lin Wang, et al., "Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal Autoencoder," ICIP, 2017.
[13] Mahdyar Ravanbakhsh, et al., "Abnormal Event Detection in Videos using Generative Adversarial Nets," ICIP, 2017.
[14] Mahdyar Ravanbakhsh, et al., "Training Adversarial Discriminators for Cross-channel Abnormal Event Detection in Crowds," WACV, 2019.
[15] Hung Vu, et al., "Robust Anomaly Detection in Videos Using Multilevel Representations," AAAI, 2019.
[16] Wen Liu, et al., "Future Frame Prediction for Anomaly Detection–A New Baseline," CVPR, 2018.
[17] Trong Nguyen Nguyen, et al.," Anomaly Detection in Video Sequence with Appearance-Motion Correspondence," ICCV, 2019.
[18] Cewu Lu, et al., "Abnormal Event Detection at 150 FPS in MATLAB," ICCV, 2013.
[19] Ian J. Goodfellow, et al., "Generative Adversarial Networks," NeurIPS, 2014.
[20] Varun Chandola, et al., "Anomaly Detection: A Survey," ACM CSUR, 2009.
[21] Phillip Isola, et al., "Image-to-Image Translation with Conditional Adversarial Networks," CVPR, 2017.
[22] Christian Szegedy, et al., "Going Deeper with Convolutions," CVPR, 2015.
[23] Thomas N. Kipf, et al., "Semi-Supervised Classification with Graph Convolutional Networks," ICLR, 2017.
[24] Xingjian Shi, et al., "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," NeurIPS, 2015.
[25] Olaf Ronneberger, et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI, 2015.
[26] Edgar Schonfeld, et al., "A U-Net Based Discriminator for Generative Adversarial Networks," CVPR, 2020.
[27] Sangdoo Yun, et al., "Cutmix: Regularization Strategy to Train Strong Classifiers with Localizable Features," ICCV, 2019.
[28] Juntang Zhuang, et al., "AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients," NeurIPS, 2020.
[29] Ningning Ma, et al., "Funnel Activation for Visual Recognition," ECCV, 2020.
[30] Takeru Miyato, et al., "Spectral Normalization for Generative Adversarial Networks," ICLR, 2018.
[31] Eddy Ilg, et al., "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks," CVPR, 2017.