A Novel Framework to Generate Synthetic Video for Foreground Detection in Highway Surveillance Scenarios

Xuan Li, Haibin Duan, *Senior Member, IEEE*, Bingzi Liu, Xiao Wang, *Member, IEEE*, and Fei-Yue Wang, *Fellow, IEEE*

***Abstract*— Foreground detection (FD) plays an important role in the domain of video surveillance for highway. The design of advanced FD algorithms requires large-scale and diverse video dataset. However, collecting and labeling real dataset is still time-consuming, labor-intensive, and highly subjective. To address this issue, we first use computer graphics (CG) to clone real highway scenarios (HS) and generate synthetic multi-challenge video datasets, called "Synthetic-HS (CG)", automatically labeled with accurate pixel-level ground truth. The Synthetic-HS (CG) dataset contains eight imaging condition sequences for computer vision research. Then, we design an image translation (IT) model that translates source domain (Synthetic-HS (CG)) to target domain (real). This model uses skip connections and attention module to generate realistic synthetic images "Synthetic-HS (IT)". We use publicly available Synthetic-HS in combination with the corresponding real video sequence to conduct experiments. The experiment results suggest that: 1) The Synthetic-HS (CG) dataset enables us to provide precise quantitative evaluation of the drawbacks of foreground detection methods 2) The realistic Synthetic-HS (IT) images can be used to promote the visual perception in highway video surveillance.**

***Index Terms*— Foreground detection, highway surveillance, parallel intelligent, synthetic dataset, scenarios engineering.**

I. Introduction

NOWADAYS, intelligent traffic surveillance has developed rapidly, which is widely used in intelligent transportation [1], [2] and unmanned aerial vehicles (UAV) [3], [4]. In parallel, the number of surveillance video on highways is growing exponentially. Scientists [5], [6] used different computer vision tasks to locate the objects in surveillance video. In fact, foreground detection (FD) is an important early task in this field, and it aims to distinguish between foreground and background areas in video sequences utilizing background models.

Currently, more research activities have focused on the modeling of foreground detection algorithms, which can be divided, under a broad categorization, into mathematical [7], [8], [9], [10], [11], machine learning [12], [13], [14], [15], [16], [17] and signal processing [18], [19] approaches. The mathematical methods [7], [8], [9], [10], [11] used for foreground detection include fuzzy and Dempster-Shafer concepts, which rely on a background model of the video to segment the moving pixels. The machine learning methods consist of three branches, such as subspace learning, neural networks and deep learning. These methods [12], [13], [14], [15], [16], [17] usually train Deep Convolutional Neural Networks (DCNN) to predict the foreground. The signal processing methods [18], [19] try to extend the classical concepts of digital signal processing to signals supported on foreground detection. Moreover, the large-scale datasets [20], [21], [22] can be useful in designing visual models. For example, the model training helps focus on moving objects in specific scenarios. While the model evaluation identifies the overall performance of models and quantitatively evaluate the shortcomings under complex conditions. Although its importance, literature lacks of comprehensive training and evaluation of recent FD methods in traffic scenarios for two reasons. First reason, labeling pixel-level real dataset is cumbersome and error-prone. Another reason is the huge effort involved in collecting the complex imaging conditions of natural video sequences. As a result, researchers can only use limited labeling frames to design, train or evaluate models, which brings serious problems to the intelligent traffic surveillance. To our knowledge, computer vision models rely on diverse datasets to ensure performance. We need to find new ways to enable intelligent traffic surveillance can perform well in complex and challenging scenarios.

In intelligent traffic systems, the advantage of virtual reality technology in supporting simulations has already been recognized by many researchers. Some new frameworks called scenarios engineering and parallel intelligent are pro-

Manuscript received 5 January 2022; revised 21 May 2022 and 15 October 2022; accepted 20 February 2023. Date of publication 28 March 2023; date of current version 31 May 2023. This work was supported in part by the Science and Technology Innovation 2030-Key Project of "New Generation Artificial Intelligence" under Grant 2018AAA0102303; in part by the National Natural Science Foundation of China under Grant 62203250, Grant U20B2071, Grant U19B2033, Grant U1913602, Grant 91948204, and Grant U2121003; in part by the Young Elite Scientists Sponsorship Program of China Association of Science and Technology under Grant YESS20210289; in part by the China Post-Doctoral Science Foundation under Grant 2020TQ1057 and Grant 2020M682823; and in part by the Basic and Frontier Research Project of the Peng Cheng Laboratory. The Associate Editor for this article was C. Guo. *(Corresponding author: Haibin Duan.)*

Xuan Li and Bingzi Liu are with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: lix05@pcl.ac.cn; liubz@pcl.ac.cn).

Haibin Duan is with the Key Laboratory of Virtual Reality Technology and Systems, School of Automation Science and Electrical Engineering, Beihang University, Beijing 100083, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: hbduan@buaa.edu.cn).

Xiao Wang and Fei-Yue Wang are with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: x.wang@ia.ac.cn; feiyue@ieee.org).

Digital Object Identifier 10.1109/TITS.2023.3253919

posed [23], [24]. In scenarios engineering, events happening [25], [26] in both worlds will affect each other and form a closed self-boosting loop. In this state, intelligent traffic surveillance systems can be easily tested under different conditions in synthetic world, which can be difficult to achieve in the real world because some scenarios rarely happen. Besides, Bainbridge and Subrahmanian [27], [28] have pointed out that synthetic worlds have significant potential as laboratories for research in computer science, social sciences, economics, social and behavioral sciences. For instance, the UnrealStereo [29] and Virtual KITTI [30] show that synthetic traffic datasets can effectively control hazardous factors and test algorithms by varying the types and degrees of the hazard. Although computer graphics algorithms can easily build synthetic worlds, domain shift is a problem that deserves special attention. The variants of generative adversarial networks [31], [32] can translate images from source domain to any target domain. These techniques may further explored and applied to foreground detection community.

In light of that, computer graphics (CG) technology and generative adversarial networks (GAN) inspire the design simulation laboratory for intelligent traffic surveillance. In this paper, we propose a challenging synthetic highway scene to promote the visual intelligence algorithms of FD. In synthetic scenes, we can easily change hazardous factors and automatically acquire accurate pixel-level labels. Therefore, the Synthetic-HS (CG) dataset successfully addresses the problem of ground truth annotation and diversified data acquisition. Besides, we present an image translation (IT) framework that translate between domains with unpaired datasets. The realistic Synthetic-HS (IT) dataset can effectively improve the performance of deep learning models. To sum up, the Synthetic-HS dataset consists of Synthetic-HS (CG) and Synthetic-HS (IT). These Synthetic-HS datasets can not only automatically generate annotations under various conditions, but also capture realistic details and textures. Therefore, the Synthetic-HS dataset can be used to effectively train and quantitatively evaluate different FD models.

The main contributions in this paper are three-fold.

First, we use a novel cloning method to generate synthetic highway scenes. In synthetic scenes, advanced computer graphics techniques can help us to generate synthetic datasets with auto-labeled ground truth (as shown in Fig. 1), called "Synthetic-HS (CG)". This makes it possible to perform foreground detection experiments using synthetic data.

Second, the Synthetic-HS (CG) dataset consist of eight video sequences that are more diverse and challenging than the corresponding real scenes. We make full use of the Synthetic-HS (CG) dataset to verify the potential impact of external challenges on FD detectors. These models are quantitatively analyzed under different imaging conditions, including dynamic background, illumination variation, bad weather, etc.

Third, we propose an image translation framework consisting of the U-Net and GAN. The U-Net adopts skip connections to reconstruct the semantic space with more features. In addition, the GAN model uses attention module, cycle-consistency constraint and adversarial training to focus on local features, thus generating photo-realistic synthetic images (Synthetic-HS (IT)). Case studies are carried out to demonstrate the practicability of above Synthetic-HS datasets in real scenarios. For instance, the supervised FD algorithms obtain more detailed features, which is beneficial to promote the visual perception in highway surveillance.

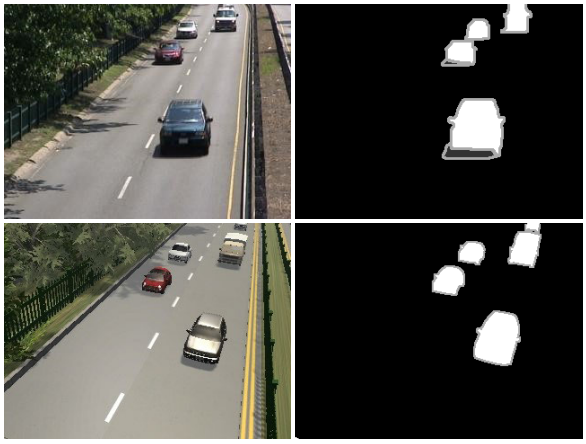

Fig. 1. Video-Frames from real CDnet2014 and rendered synthetic clones. Top: real image (left) and real annotation (right). Bottom: Synthetic-HS (CG) image (left) and Synthetic-HS (CG) annotation (right). Source data are provided with this paper. The CDnet2014 dataset can be downloaded from: http://www.changedetection.net/. The Synthetic-HS dataset is available at: https://github.com/PC-Lab-Virtual-Reality/A-Novel-Framework-to-Generate-Synthetic-Video-for-Change-Detection-in-Highway-Surveillance-Scenarios.

The overall framework of the proposed method is summarized in Fig. 2. The remainder of this article is organized as follows. In Section II, we present a brief introduction to related works of foreground detection datasets and image generation methods. The technical details of the Synthetic-HS (CG) dataset are proposed in Section III. Section IV provides the frameworks of image translation networks. Section V details experimental procedures and results on real and Synthetic-HS datasets. Finally, the conclusion is drawn in Section VI.

II. Related Works

In this part, we present the publicly available foreground detection datasets, and the existing image generation methods.

A. Foreground Detection Datasets

It is generally accepted that the traditional or supervised learning-based approaches rely on datasets for experimental design, training, and testing. In recent decades, diverse FD datasets have been presented to explore the impact of different challenges on algorithms. For instance, the Wallflower dataset[1] [33] allowed researchers to test background model maintenance capability. Ferryman et al. [34] provided the PETS2001 dataset[2] to evaluate the performance of FD models under illumination quick changes. In 2004, Li et al. [35] presented 10 video sequences to test illumination changes and dynamic background influence on FD algorithms. After

[1]https://www.microsoft.com/en-us/research/people/jckrumm/downloads/
[2]http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001-dataset.html

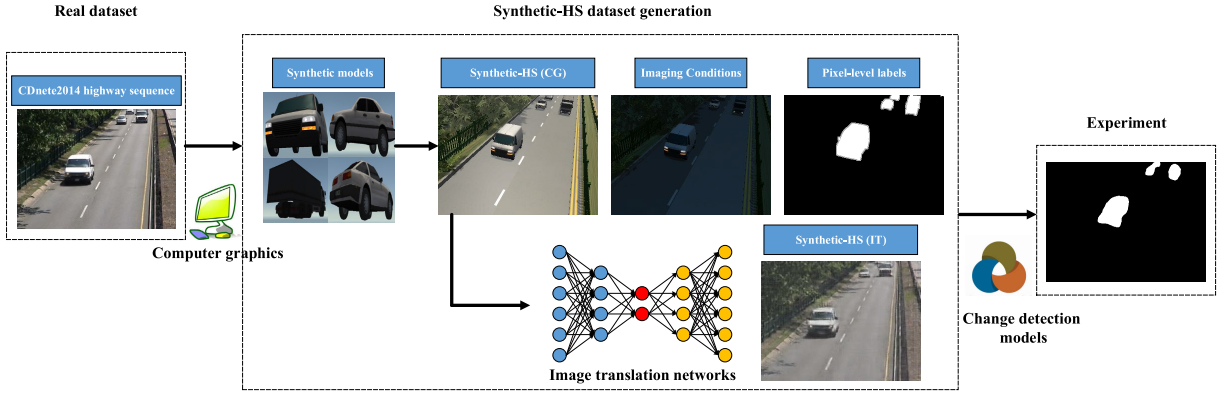

Fig. 2. The overall framework of the proposed method. The Synthetic-HS dataset consists of Synthetic-HS (CG) and Synthetic-HS (IT). The Synthetic-HS (CG) dataset denotes to the image with accurate labeling and multi-challenge generated by computer graphics platform. The Synthetic-HS (IT) dataset represents to the image with realistic textures generated by image translation network. Note that real and Synthetic-HS datasets are utilized to conduct experiments for foreground detection in highway surveillance scenarios.

that, a unique FD dataset consisting of 31 video sequences representing 6 categories (dynamic background, camera jitter, shadow, intermittent motion, etc) was proposed in 2014, namely CDNet2014[3] [36]. This dataset provides algorithm rankings on a website and is widely used as a FD benchmark.

Although above datasets are very important, effort involved in generating qualitatively dataset is still a challenging process. Hence, literature lacks of recent real FD datasets. A feasible scheme is that the synthetic data generated by synthetic worlds contributes to the rapid development of computer vision. For foreground detection, Tiburzi et al. [37] used the chroma technique to automatically obtain pixel-level segmentation masks for 15 semisynthetic traffics. Brutzer et al. [38] provided a synthetic dataset[4] to compare the performance of nine FD algorithms under six challenges. On November 5th 2012, Vacavant et al. [39] presented a benchmark dataset[5] and evaluation process built from both synthetic and real videos for the Background Models Challenge (BMC). To be specific, this dataset consists of 20 synthetic videos and 9 real videos with different scenes. In view of that fact, the synthetic world offers an alternative, a cost-effective dataset that appears realistic, and is automatically labeled for computer vision community. However, those synthetic datasets are especially suitable for algorithm's results evaluation. In fact, the synthetic dataset has great potential to promote FD research.

B. Image Generation Methods

Image generation is a hot topic, and a large number of classical methods and models have been proposed in recent years. Here, we brief introduce the related works from two aspects. With computer graphics development, it is possible to construct synthetic scenes manually based on real scene. More importantly, synthetic scenes allow us to extract key information from 3d models for automatic image annotation. For instance, Ros et al. [40] proposed a synthetic scene (SYNTHIA)[6] to automatically generate synthetic images with semantic segmentation annotations. Richter et al. [41] presented an approach to rapidly creating pixel-accurate semantic label maps for images extracted from Grand Theft Auto V (a modern computer game). The Virtual KITTI[7] [30] contains 17,000 photo-realistic synthetic images, all with automatic accurate ground truth (including object detection, tracking, semantic segmentation, etc.). Wang et al. [42] developed a data collector and labeler for crowd counting, which can automatically collect and annotate images without any manpower. Li et al. [21], [25] constructed large-scale, realistic synthetic scenes to generate challenging driving images, which can be used for quantitative analysis and testing object detection algorithms.

Above synthetic scenes and images provided major challenges of computer vision to assess detectors. Due to the advancement of deep learning, the Generative Adversarial Network (GAN) [43], [44], [45], [46], [47], [48] has been shown to outperform tradition methods in unpaired image generation field. In 2017, Zhu et al. [43] first proposed CycleGAN (Cycle-Consistent Adversarial Networks) to translate between domains with unpaired datasets. Using a shared-latent space assumption to obtain high quality images, the unsupervised image-to-image translation framework (UNIT) [44] is presented in 2017. The DRIT (2018) (Disentangled Representation Image Translation) [45] model apply disentangled representation to generate diverse synthetic images with unpaired data. Kim et al. [46] introduced U-GAT-IT model (2020) (Unsupervised Generative Attentional Networks). This model use AdaLIN (Adaptive Layer-Instance Normalization) function to fine-tune the image change on shapes and textures. The NICE-GAN (2020) (No-Independent Component-for-Encoding GAN) [47] introduced a decoupled training strategy and achieved better performance over other methods. In 2021, Ye et al. [48] proposed a new image generation

[3]http://www.changedetection.net/

[4]https://www.vis.uni-stuttgart.de/en/research/visual_analytics/visual-analyticsvideosteams/stuttgart_artificial_background_subtraction_dataset/index.html

[5]http://backgroundmodelschallenge.eu/

[6]http://synthia-dataset.net/

[7]https://europe.naverlabs.com/research/computer-vision/proxy-virtual-worlds-vkitti-1/

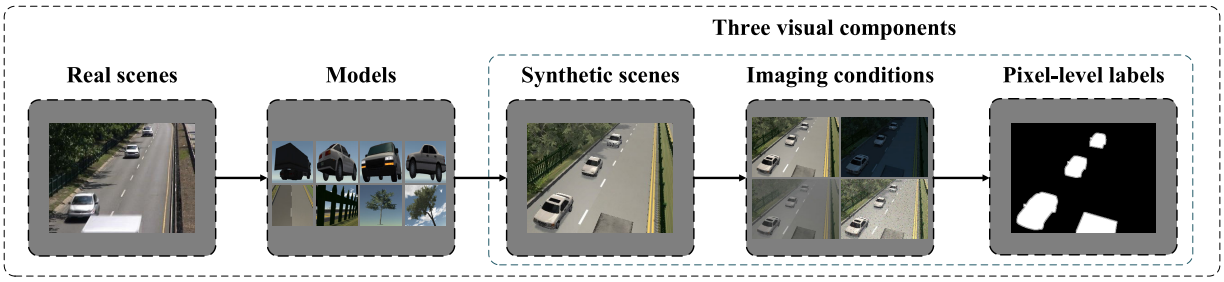

Fig. 3. The overall processes of Synthetic-HS (CG) dataset generation method. This framework uses a cloning method to construct synthetic scenes from real scenes. The synthetic scene is decomposed into three visual components (background models, foreground objects, and imaging conditions). In this scenes, accurate ground-truth labels can be automatically acquired, and imaging conditions can be flexibly changed. The role of each visual component is detailed in Section III.

architecture that removes the encoder for each network. This allows each network to focus on its own goals.

III. Detailed Implementation of the Synthetic-HS (CG) Dataset

Existing FD datasets face two major limitations that hinder computer vision research. First, the evaluation of algorithms with respect to the problems in real video surveillance are missing, inaccurately labeled, or low diversity. Second, the synthetic scene construction is often requires creating animation movies from scratch, which is expensive and labor-intensive. Due to these limitations, only a few previous efforts have fully utilized the potential of datasets to help algorithm design and evaluation. Consider these problems, we propose a synthetic traffic scene to solve two main challenges: (i) use cloning methods to generate realistic and diverse video sequences. (ii) edit scripts for modern platforms to automatically obtain ground truth. The overall processes of Synthetic-HS (CG) dataset generation method is illustrated in Fig. 3. The framework consists of the following steps: To begin with, the generation of synthetic scenes (Section III-A). Then, the creation of FD challenges for highway scenes (Section III-B). Furthermore, the automatic generation of pixel-level annotation (Section III-C).

A. Introducing Synthetic Models and Creating Synthetic Scenes

The FD datasets require a large number of scenarios to cover detection challenges. For instance: the CDnet2014 is a benchmark dataset, which contains 11 video challenges with 31 videos sequences. However, the dataset collected more than 10 videos sequences in surveillance field only to address limited challenges (bad weather, dynamic background). In fact, we do not need a reasonable coverage of all possible scenarios of interest. Instead, we use an existing real-world video sequence to initialize our synthetic world in highway surveillance scenarios. In this section, the original CDnet2014 benchmark (Baseline, highway category) is selected to initialize our synthetic worlds. We primary collect two types of data: camera, physical properties of important objects in real-world scenes. The real scene uses a color camera to capture 1,700 images containing 40 vehicles (cars, vans and trucks). In synthetic highway scenes, we use publicly assets and self-designed synthetic models to replace objects in the real scenes, as illustrated in Fig. 4. In addition, the main parameters of the virtual camera include height (4 meters) and field of view (FOV, 90 degrees). Here is a trick, we decompose synthetic scenes into three visual components (background models, foreground objects, and imaging conditions), which can help computer graphics engines (e.g., Unity3D) to reconstruct the synthetic scenes. This approach helps with the next two steps.

Fig. 4. Examples of synthetic models. (Top) foreground vehicles examples; (Bottom) background models.

B. Changing Imaging Conditions in Synthetic Scenes

On highway scenes, the vehicles often travel at 120km/h, and accidents can be disastrous for drivers. Extreme imaging conditions can easily cause serious accidents, which requires FD algorithms must cope with the challenges in specific scenes. As stated in [36], the typical problems for FD are defined. However, there are more extreme challenges in highway scenarios. Next, we propose the following challenges for evaluation, as shown in Fig. 5. Note that the imaging condition component is used to simulate the highway challenges in synthetic scenarios.

Basic: The basic highway scenario contains potential challenges for general performance training and evaluation.

Dynamic Background: Moving tree branches are the dynamic background. It is desirable that FD model adapts to uninteresting background movements.

Illumination Variation: The light intensity varies over time, resulting in gradual changes of the appearance and color of objects.

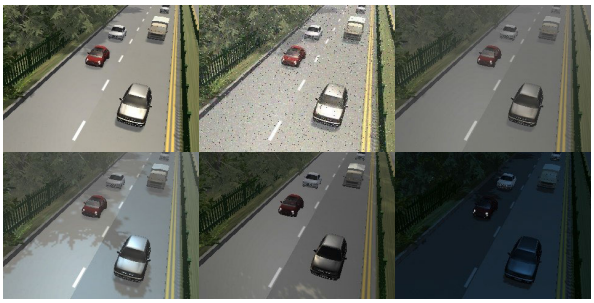

Fig. 5. Examples of our proposed Synthetic-HS (CG) dataset. Top: basic (left), noise (middle), and foggy (right). Bottom: thunder storm (left), illumination variation (middle), and night (right).

Fig. 6. Unity3D automatically generates the sample frames of the Synthetic-HS (CG) images and corresponding 4-class ground-truth labels (**Non-ROI** class is not present in this image). Note that **Unknown** border is represented by 16-neighborhood pixels. This protects evaluation metrics from being corrupted by motion blur. Best viewed with zooming.

Light Switch: In some extreme cases, sudden light changes may occur, which are not covered by background models. Once-off changes are simulated at night (frame 500) and switch it at daytime (frame 1,000).

Foggy: Fog reduces visibility on highway, rendering the driver very hazy at about 400 meters. Many lives are lost each year worldwide from accidents (multiple-vehicle collisions) involving fog conditions on the highways.

Night: Basic sequence at night, low light intensity reduces foreground/background contrast, resulting in additional noise and camouflage.

Thunder Storm: Lightning causes sudden light changes (1-2 frames), so that the current object state can not covered by the background model.

Noise: Noise is a random dot pixel pattern (flicker of "dots" or "snow") displayed when no transmission signal is obtained by display devices. Basic sequence adds some artificial noise (Gaussian, salt and pepper) to cause more camouflage. Noise models are given by:

$$P_G(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \tag{1}$$

$$P_I(z) = \begin{cases} P_a & \text{for } z = a \\ P_b & \text{for } z = b \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where z, μ and σ represents the grey level, the mean value and the standard deviation. P_a and P_b are probabilities of pepper and salt, respectively. If $b > a$, grey-level b appears as a light dot (salt) in the image. Conversely, a will appear as dark dot (pepper).

C. Automatic Generating Pixel-Level Annotations

It is well known that annotation information is important to computer vision algorithms. In fact, different visual tasks have their corresponding labeling features. For example, object detection, object tracking, and foreground detection, semantic segmentation corresponds to 2D bounding boxes and pixel-level labels, respectively. In the CDnet2014 dataset, ground truth for training and evaluation were manually labeled. This common practice is however costly, and over 40 minutes is needed to label each image at pixel-level. Besides, manual annotation information is subjectivity, inconsistency and not-scalable. For instance, moving objects boundaries state is subjective and the underlying criterion may differ from case to case. It is worth noting that the annotation state of each pixel directly affects the performance of models.

In light of that, we propose a feasible approach to replace the controversial manual method. As stated above, the real scene is decomposed into three visual components. Our approach utilizes foreground objects and background models, which can automatically generate accurate and consistent ground truth annotations. To be specific, foreground objects and background models are divided into two categories by unlit shaders on the materials to efficiently and directly generate pixel-level labels. It is noted that the original image and labels are output independently, as shown in Fig. 6. The annotation of pixel-level images follows these rules:

(1) There are four label classes: **Static**, **Moving**, **Unknown**, and **Non-ROI** (ROI stands for "Region of Interest"), which are assigned grayscale value of 0, 255, 170, and 85, respectively.

(2) The **Static** and **Moving** classes are associated with pixels for background and foreground objects. The **Unknown** class is associated with pixels corrupted by motion blur. The **Non-ROI** class represents the metrics that are evaluated within the area of interest.

(3) The first 100 frames of each video sequence are labeled as **Non-ROI**. This helps algorithms to complete initialization process.

IV. Synthetic Image Generate From Generative Adversarial Network

In previous section, we use computer graphics to generate challenging synthetic datasets (Synthetic-HS (CG)). However, this approach has two disadvantages. First, editing synthetic scenarios still require a lot of manual work. Second, the Synthetic-HS (CG) datasets lack the details and textures of real scenes. In this part, we design an image generation network to transform the manual synthetic scene modeling into the inference process of machine learning models. The traditional image translation methods guide model to learn mapping functions between two domains. However, the mapping functions usually produce blurred details in the translated images. Our proposed method uses the Synthetic-HS (CG) images to generate images with fine textures. Here, we present the details of implementing the pipeline. First, the proposed model mainly consists of the U-Net and GAN components. The U-Net model ($\{E_1, DE_1\}$) and encoder (E_2) obtain the

TABLE I

THE IMAGE-TRANSLATION NETWORKS CONFIGURATIONS. THE PARAMETERS S AND FM ARE DENOTED AS STRIDE AND FEATURE MAPS

Network	Encoders	Generators	Discriminators
Image-Translation	Conv7*7,S=1,FM=64 Conv3*3,S=2,FM=128 Conv3*3,S=2,FM=256 Resblk3*3,S=1,FM=256 Resblk3*3,S=1,FM=256 Resblk3*3,S=1,FM=256 Resblk3*3,S=1,FM=256	Global Average & Max Pooling AdaResBlock3*3,S=1,FM=256 AdaResBlock3*3,S=1,FM=256 AdaResBlock3*3,S=1,FM=256 AdaResBlock3*3,S=1,FM=256 Deconv3*3,S=1,FM=128 Deconv3*3,S=1,FM=64 Conv7*7,S=1,FM=3	Conv4*4,S=1,FM=64 Conv4*4,S=2,FM=64 Conv4*4,S=2,FM=128 Conv4*4,S=2,FM=256 Conv4*4,S=2,FM=512 Conv4*4,S=2,FM=1,024 Conv4*4,S=2,FM=2,048 Global Average & Max Pooling Conv1*1,S=1,FM=512 Conv4*4,S=1,FM=1

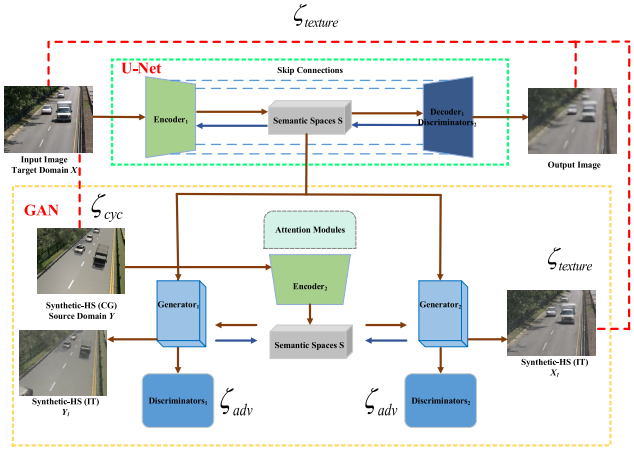

Fig. 7. The overall architecture of the image generation network. The real image is inputted into the U-Net model, which contains the skip connections. For generators and discriminators, the attention modules classify the image and localize class-specific image regions in local features. Therefore, this model can generate high-quality Synthetic-HS (IT) dataset with fine textures. Note that the orange dashed boxes are generative adversarial network (GAN) with attention module, the green dashed boxes indicate the U-Net structure, and the blue dashed lines mean skip connections between encoder and decoder. The brown and blue arrows represent forward propagation and back propagation, respectively. Best viewed with zooming.

semantic spaces S of the image and compares image textures in real and synthetic domains. For GAN, the generators (G_1 or G_2) map semantic spaces to generate different domain images (i.e. $G_1(S) \rightarrow X_1$ or $G_2(S) \rightarrow Y_1$). The discriminators (D_1 or D_2) are trained to output true for real images sampled from real or Synthetic-HS (CG) domain and false for images generated from generators (or decoders). Second, both U-Net and GAN can use the semantic spaces S, texture loss and cycle-consistency constraint to generate photo-realistic synthetic images. Fig. 7 illustrates the architecture of the image generation network.

A. Network Architectures

(1) **U-Net**, the U-Net is a typical Encoder-Decoder $\{E_1, DE_1\}$ structure. The real image is inputted into the U-Net model. The encoder E_1 uses max-pooling step to perform downsampling, while the decoder DE_1 consists of upsampling the feature map. Therefore, the real input image $\left(\{x_i\}_{i=1}^{N} \in X\right)$ is converted into a deep semantic space S, and then mapped into the desired result. However, the max-pooling step compress the detail information of image, leading to the loss of image texture. Here, the skip connections between encoder and decoder are used to reconstruct the semantic space with more features. Use this semantic space, the decoder can output more realistic images. Generative adversarial networks take detailed semantic features to generate synthetic images.

(2) **GAN**, Generative Adversarial Network: The U-Net model progressively downsample the input and only restore the input image. Next, our model uses two generative adversarial networks: $GAN_j = \left\{D_j, G_j\right\}_{j=1}^{2}$, the encoder E_2 and semantic space S to generate target domain (real) from source domain ($\left(\{y_i\}_{i=1}^{M} \in Y\right)$). The generators ($G_1$ and G_2) generate synthetic images in both domain (i.e. $G_1(S) \rightarrow Y_1$ or $G_2(S) \rightarrow X_1$), and the inputs of G_1 and G_2 from X (real) and Y (synthetic) domains. The attention modules [49] are used in both generators and discriminators, which allow the classification-trained CNN to both classify the image and localize class-specific image regions in local features or single forward-pass. The generator equips the residual blocks with AdaILN and two up-sampling layers. Thus, this method contains more patterns of the texture. It also flexibly controls the amount of change in shape and texture. The discriminator D_1 or D_2 are multi-scale models, which are trained to classifies images sampled from real or synthetic (G_1 or G_2) images. For the generators and the discriminators, the details architecture is given in Table I.

B. Objective

We jointly train the U-Net$_1$ $= \{E_1, DE_1\}$, Encoder (E_2) and GAN$_{1,2}$ $= \left\{G_{1,2}, D_{1,2}\right\}$ for image feature extraction, adversarial loss, and cycle-reconstruction. Therefore, the final objective of our model:

$$\begin{aligned}\zeta\left(E_1, DE_1, E_2, G_1, G_2, D_1, D_2\right) &= \lambda_1 \zeta_{tex}\left(E_1, DE_1, G_2\right) \\ &\quad + \lambda_2\left(\zeta_{adv}\left(E_2, G_2, D_2\right) + \zeta_{adv}\left(E_1, G_1, D_1\right)\right) \\ &\quad + \lambda_3 \zeta_{cyc}\left(E_1, E_2, G_1, G_2\right) \end{aligned} \tag{3}$$

where $\lambda_1, \lambda_2, \lambda_3$ are weights that control the importance of texture extraction, adversarial training and cycle-reconstruction terms. The ζ_{tex}, ζ_{adv} and ζ_{cyc} are texture loss, adversarial loss and cycle-consistency loss, respectively.

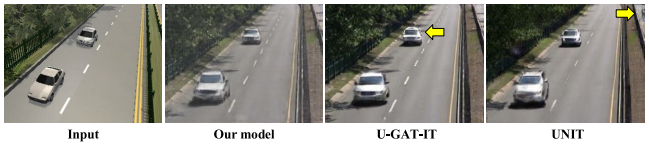

Fig. 8. Examples of results on synthetic images using different models. Column 1: input image. Column 2: the output of our proposed model. Column 3: the output of U-GAT-IT model. Column 4: the output of UNIT model. The yellow symbols indicate the a mismatched structure is generated. Best viewed with zooming.

1) Texture Loss: In this paper, the texture loss function is also called L_1 loss and has strong robustness. Even if the gradient has fluctuations, we can provide a higher pixel quality and more accurate texture capability. The texture loss is:

$$\zeta_{tex}\left(E_1, DE_1, G_2\right) = \mathbb{E}_{x\sim p(X)}\left[x - \|DE_1\left(E_1\left(x\right)\right)\|_1\right] + \mathbb{E}_{x\sim p(X)}\left[x - \|G_2\left(E_1\left(x\right)\right)\|_1\right] \quad (4)$$

2) Adversarial Loss: To better obtain realistic synthetic images, we employ adversarial loss to match the distribution of source images to target images. The GAN objective functions are given by:

$$\zeta_{adv}\left(E_2, G_2, D_2\right) = \mathbb{E}_{y\sim p(Y)}\left[\log D_2(y)\right] + \mathbb{E}_{y\sim p(Y)} \times \left[\log\left(1 - D_2\left(G_2\left(E_2(y)\right)\right)\right)\right] \quad (5)$$

$$\zeta_{adv}\left(E_1, G_1, D_1\right) = \mathbb{E}_{x\sim p(X)}\left[\log D_1(x)\right] + \mathbb{E}_{x\sim p(X)} \times \left[\log\left(1 - D_1\left(G_1\left(E_1(x)\right)\right)\right)\right] \quad (6)$$

First, these functions are used to reduce the differences between output image and target image. Second, semantic space can obtain more classification information, which is helpful to distinguish foreground and background.

3) Cycle-Consistency Constraint: Target domain images X are downsample to semantic spaces $E_1(x)$, and generative adversarial networks use semantic spaces and source images Y to generate synthetic images. Here, cycle-consistency constraint ensures the stability of the transition from source to target images, vice versa. The cycle-consistency constraint is given by:

$$\zeta_{cyc}\left(E_1, E_2, G_1, G_2\right) = \mathbb{E}_{x\sim p(X)}\left[G_1\left(E_1(x)\right) - x\right] + \mathbb{E}_{y\sim p(Y)}\left[G_2\left(E_2(y)\right) - y\right] \quad (7)$$

In this way, the cycle consistency constraint further ensures that the image generation process synthesizes the target domain images. The above functions can ensure the effectiveness, controllability and robustness of the whole image generation network. The Fig. 8 shows that the foreground textures and background contents translated by IT are closer to real images.

C. Optimization and Implementation Details

In our experiments, the architecture has several hyper-parameters ($\lambda_1 = 0.2, \lambda_2 = 0.3, \lambda_3 = 5$). The default values for these hyper-parameters are determined by experiments (the quality of synthetic images). To optimize our network, we apply ADAM for training, with a learning rata of 0.0001 and momentum parameters are set to 0.5 and 0.999.

TABLE II

TABLE OF INCEPTION SCORES FOR SAMPLES GENERATED BY VARIOUS MODELS. SCORE HIGHLY CORRELATES WITH HUMAN VISUAL JUDGMENT, AND THE HIGHER THE BETTER

Model	Synthetic-HS (IT)	UGATIT	NICE-GAN	UNIT
Score	**1.308**	1.294	1.235	1.185

V. EXPERIMENTAL EVALUATION

A. Experimental Procedure

The traditional datasets contain only one challenge and use inaccurate annotations for a given scenario. This leads to limited training and unreliable evaluation of FD models. On the contrary, we generate the auto-labeled synthetic datasets "Synthetic-HS (CG)" from the real scenes. The Synthetic-HS (CG) consists of eight video challenges (each sequence has 1,230 frames, which can be added and configured flexibly), one of which is predominant and also influenced by others. In addition, we also use our model and other image translation models to obtain realistic synthetic images. The Fig. 8 illustrates synthetic examples of translated output using different image translation models. Then, we conduct experiments to demonstrate that our proposed method improves the inception score [50], presented in Table II. Samples from the Synthetic-HS (IT) dataset achieve the highest value. Currently, the Synthetic-HS dataset consists of 20,400 frames (Synthetic-HS (CG) with 13,600 frames, and Synthetic-HS (IT) with 6,800 frames) taken from the highway surveillance scenarios. To verify the reliability of our proposed database, we select a wide range of FD methods including recent learning-based methods and traditional approaches. The content of the experimental procedure is briefly described as follows: First and foremost, the comparative experiments are conducted on real and synthetic datasets. Then, the eight visual challenges involving synthetic highway scenes are used to explore the impact of extreme conditions on FD algorithms. Moreover, some supervised models are trained and tested on real and synthetic database. Finally, we summarize the experimental results briefly.

B. Metrics of Performance Measure

It is well known that FD training and evaluation rely on pixel-level information, which is accurate and objective. Several metrics for FD assessment are provided by the CDnet2014 organizers. In our experiments, the recall, precision and F-Measure are selected to identify the correctness of binary classification. These metrics are generally considered to be good indicators for overall evaluation, and expressed as:

$$Recall = \frac{T_P}{T_P + F_N} \quad (8)$$

$$Precision = \frac{T_P}{T_P + F_P} \quad (9)$$

$$F_{measure} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

where TP = correctly classified foreground pixels, $TP + FN$ = foreground pixels in ground truth, $TP + FP$ = pixels

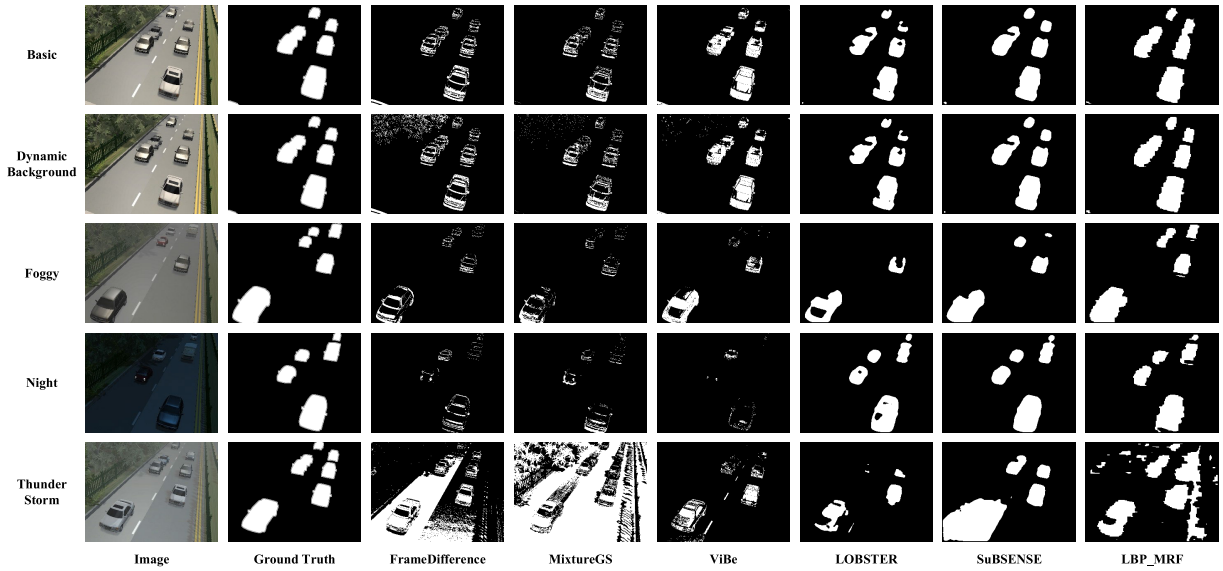

Fig. 9. Visualization results on the synthetic multi-challenge datasets. The column 1 shows input images; the column 2 contains ground truth masks. The column 3-8 correspond to foreground detection results predicted by the FrameDifference, MixtureGS, ViBe, LOBSTER, SuBSENSE and LBP_MRF models, respectively. Note that **Unknown** border is added around the foreground objects in the ground truth image. Best viewed with zooming.

classified as foreground. Most of the experimental results give precision, recall and F-measure tables, since this presentation is favored over the commonly used in the case of FD.

C. Assessing the Usefulness of Synthetic Dataset

Compared to real data, the generation and annotation of Synthetic-HS (CG) datasets are key innovations in our approach. In real highway sequence, only 1,230 frames in the sequence are full-labeled (see Fig. 1). The remaining frames are labeled as **Non-ROI** to prevents the corruption during initialization. In this work, the synthetic data has annotation information for each frame. We select 1,700 full-labeled frames from Synthetic-HS (CG, basic)) as the evaluation data. The six traditional FD methods (FrameDifference [51], MixtureGS [52], ViBe [53], LOBSTER [54], SuBSENSE [55], LBP_MRF [56]) are used in experiments on evaluation data and real data. For more experiments, the readers can employ algorithms available at the BGSLibrary[8] and at the LRSLibrary.[9] Table III shows the results as compared with above methods. The performance of considered algorithms has no significant difference between real data and synthetic data, which indicates that Synthetic-HS (CG) dataset, as a supplement to real data, can effectively distinguish various approaches.

TABLE III

COMPARISON RESULTS OF THE SIX TRADITIONAL FOREGROUND DETECTION ON THE REAL HIGHWAY SEQUENCE AND CORRESPONDING S-HS (CG, BASIC)) SEQUENCE. NOTATION: 'S-HS (CG)' MEANS SYNTHETIC-HS (CG) DATASET GENERATED BY COMPUTER GRAPHICS

Method	Testing	Recall	Precision	F-Measure
FrameDifference [51]	Real	51.2%	76.6%	61.4%
MixtureGS [52]	Real	45.5%	83.4%	58.8%
ViBe [53]	Real	67.8%	90.2%	77.3%
LOBSTER [54]	Real	84.8%	91.9%	88.2%
SuBSENSE [55]	Real	95.0%	**94.1%**	**94.6%**
LBP_MRF [56]	Real	**95.8%**	80.2%	87.3%
FrameDifference	S-HS (CG)	35.0%	91.4%	50.6%
MixtureGS	S-HS (CG)	31.5%	87.0%	46.3%
ViBe	S-HS (CG)	71.1%	93.0%	80.5%
LOBSTER	S-HS (CG)	77.9%	95.2%	85.7%
SuBSENSE	S-HS (CG)	85.9%	**95.9%**	**90.6%**
LBP_MRF	S-HS (CG)	**87.7%**	90.6%	89.1%

D. Evaluation on Synthetic Multi-Challenge Datasets

On highway, most accidents occur under extreme conditions. For example: the Tesla collided with a huge and white truck, because the algorithm detects that it was a cloud. As described above, our proposed Synthetic-HS (CG) dataset has great potential for model evaluation. Compared with real scenes, the proposed synthetic scenes are controllable, considerable, and repeatable. Therefore, we provide 8 imaging conditions for highway surveillance setting to evaluate the performance of detectors, including Basic, Dynamic Background, Illumination Variation, Light Switch, Foggy, Night, Thunder Storm and Noise. The quantitative results are shown in Table IV. We then analyze the impact of each imaging condition on algorithmic performance.

Video Basic (V_B): This sequence provides a first impression and challenge of the highway scenario. This sequence shows different vehicles driving on the road, and the algorithm needs

[8]https://github.com/andrewssobral/bgslibrary
[9]https://github.com/andrewssobral/lrslibrary

TABLE IV
COMPARISON RESULTS OF THE SIX TRADITIONAL FOREGROUND DETECTION MODELS ON THE SYNTHETIC MULTI-CHALLENGE DATASETS, INCLUDING: BASIC, DYNAMIC BACKGROUND, ILLUMINATION VARIATION, LIGHT SWITCH, FOGGY, NIGHT, THUNDER STORM, AND NOISE

Method	Testing	Recall	Precision	F-Measure
FrameDifference [51]	**V_B**	35.4%	91.3%	51.0%
MixtureGS [52]	**V_B**	31.2%	86.6%	45.9%
ViBe [53]	**V_B**	69.6%	92.4%	79.4%
LOBSTER [54]	**V_B**	77.8%	95.0%	85.6%
SuBSENSE [55]	**V_B**	85.1%	**95.7%**	**90.1%**
LBP_MRF [56]	**V_B**	**88.5%**	90.3%	89.4%
FrameDifference	**V_DB**	36.1%	63.3%	46.0%
MixtureGS	**V_DB**	31.7%	80.9%	45.6%
ViBe	**V_DB**	69.6%	90.3%	78.6%
LOBSTER	**V_DB**	77.6%	94.7%	85.3%
SuBSENSE	**V_DB**	85.1%	**95.6%**	**90.1%**
LBP_MRF	**V_DB**	**89.0%**	89.0%	89.0%
FrameDifference	**V_IV**	34.7%	**89.1%**	50.0%
MixtureGS	**V_IV**	30.0%	82.6%	44.0%
ViBe	**V_IV**	70.7%	72.5%	71.5%
LOBSTER	**V_IV**	77.9%	83.5%	**80.6%**
SuBSENSE	**V_IV**	84.8%	73.3%	78.6%
LBP_MRF	**V_IV**	**89.1%**	62.4%	73.4%
FrameDifference	**V_LS**	25.1%	85.2%	38.8%
MixtureGS	**V_LS**	20.8%	70.7%	32.2%
ViBe	**V_LS**	80.9%	16.1%	26.8%
LOBSTER	**V_LS**	**87.9%**	13.3%	23.1%
SuBSENSE	**V_LS**	81.6%	74.5%	77.9%
LBP_MRF	**V_LS**	84.0%	**89.6%**	**86.7%**
FrameDifference	**V_F**	23.4%	98.0%	37.8%
MixtureGS	**V_F**	16.8%	85.3%	28.1%
ViBe	**V_F**	38.2%	99.9%	55.3%
LOBSTER	**V_F**	38.6%	99.9%	55.7%
SuBSENSE	**V_F**	62.8%	**99.9%**	77.1%
LBP_MRF	**V_F**	**79.1%**	95.7%	**86.6%**
FrameDifference	**V_N**	16.3%	97.5%	27.9%
MixtureGS	**V_N**	11.2%	79.5%	19.6%
ViBe	**V_N**	8.0%	**99.8%**	14.8%
LOBSTER	**V_N**	79.9%	94.9%	86.8%
SuBSENSE	**V_N**	**87.9%**	95.1%	**91.3%**
LBP_MRF	**V_N**	79.0%	90.2%	84.2%
FrameDifference	**V_TS**	23.7%	77.6%	36.3%
MixtureGS	**V_TS**	16.9%	55.4%	26.0%
ViBe	**V_TS**	38.0%	97.6%	54.6%
LOBSTER	**V_TS**	38.1%	**97.8%**	54.8%
SuBSENSE	**V_TS**	62.5%	95.6%	75.6%
LBP_MRF	**V_TS**	**79.1%**	94.4%	**86.1%**
FrameDifference	**V_No**	34.7%	89.1%	50.0%
MixtureGS	**V_No**	35.4%	31.3%	33.2%
ViBe	**V_No**	62.0%	69.5%	65.5%
LOBSTER	**V_No**	64.6%	97.6%	77.7%
SuBSENSE	**V_No**	70.0%	**98.3%**	**81.8%**
LBP_MRF	**V_No**	**82.4%**	74.3%	78.2%

to determine the state of each pixel. According to Table IV, the FrameDifference and MixtureGS methods show high precision (over 85%) at low recall, resulting in unsatisfactory F-measure metric. This seems that multimodal techniques have trouble capturing fast-moving vehicles (holes left in the foreground, called the "blobs" problems). Segmentation masks in Fig. 9 (row 1) exhibit that the rest four methods can well detect the moving vehicles.

Video Dynamic Background (V_DB): Some uninteresting movement that has to be deemed as background. In this sequence, moving tree branches is a difficult interference for some algorithms. Compared with the basic sequence, the precision and F-measure metrics of most models are decreased. Also remarkable is that moving tree branches even directly affects the segmentation masks of the FrameDifference, MixtureGS and ViBe algorithms in Fig. 9 (row 2). Approaches using LBSP features (Local Binary Similarity Pattern), such as the LOBSTER and SuBSENSE models can properly handle this region.

Video Illumination Variation (V_IV): The illumination variation challenge causes gradual changes for the foreground and background in this scene, such as vehicle appearance and branch shadows. Compared with the dynamic background sequence, the experimental results prove that the changes of two elements bring greater challenges to all models. For instance, it is worth noting that F-measure metric of the LBP-MRF and SuBSENSE models decreased by 15.6% and 13%, respectively. Although the SuBSENSE model has a feedback method to automatically adjust threshold, it still can not handle the pixel classification well in this challenge.

Video Light Switch (V_LS): In this sequence, 500-1,000 frames are at night and the rest are daytime. The Table IV shows that only the SuBSENSE and LBP-MRF maintain good performance metrics. Other FD methods do not satisfactorily handle sudden once-off changes in illumination. In particular, the background models of the ViBe and LOBSTER are unable or slow to update, which lead to major performance loss in this experiment.

Video Foggy (V_F): Many lives are lost each year worldwide from accidents involving fog conditions on the highways. The experimental results show that many tested FD methods are affected by this challenge. Segmentation masks in Fig. 9 (row 3) exhibit that the ViBe and LOBSTER algorithms are most affected, such as error detection, foreground blobs. This phenomenon is caused by poor visibility at a distance, resulting in camouflage of foreground objects.

Video Night (V_N): Driving at night is a dreadful and dangerous. Road fatalities triple during the night, and human eyes do not help much either. The results are depicted in Table IV, note that for the FrameDifference, MixtureGS and ViBe algorithms do not properly handle the more camouflage generated in this experiment. Therefore, the results of these three methods show quite low performance. In contrast, the three most recent FD algorithms (LOBSTER, SuBSENSE, and LBP-MRF) are well suited to increasing gain level and low background/foreground contrast.

Video Thunder Storm (V_TS): In thunder storm sequence, lightning frequently occurs 1 frame out of every 100 frames, causing the foreground and background change intermittently. The light switch sequence is similar to this challenge.

TABLE V

ABLATION STUDY. RESULTS OF METHODS ARE ALL INTHE SAME SETTING. THEREARE SOME NOTATIONS; IT: IMAGE TRANSLATION, AT: ATTENTION MODULE, TE: TEXTURE, CY: CYCLE-CONSISTENCY CONSTRAINT

Model	Recall	Precision	F-Measure
IT	**69.5%**	74.6%	**71.9%**
IT w/o AT	58.1%	**82.5%**	68.2%
IT w/o Te loss	62.0%	80.7%	70.1%
IT w/o Cy loss	46.9%	58.7%	52.1%

The results show that the ViBe and LOBSTER have little performance improvement under thunderstorm conditions. The evaluation performance of other algorithms remains stable. Note that the above results are compared with the light switch sequence.

Video Noise (V_No): Video noise is an undesirable by-product of image capture that obscures the desired information. In this sequence, artificial noise (Gaussian noise, salt and pepper) is used to evaluate algorithm performance. Accordingly, the results of this sequence show quite low performance for all evaluated approaches. For instance, the MixtureGS algorithm does not reach F-measure value above 34%, due to multimodal techniques can not fit the scene distributions with random noise. In addition, the precision of ViBe method decreased by 22.9%. Because this background model updates slowly and is sensitive to noise changes.

E. Ablation Study

We performed an ablation study (on Synthetic-HS (CG) dataset and our image translation framework) measuring impact of the different kinds of loss functions, attention module to the translation performance and showed the recall, precision and F-Measure results in Table V. Ideally, we expect all metrics to perform better. However, there is a trade-off indicating that more recall is associated with less precision. We analyze three key components including attention module, texture loss, and cycle-consistency constraint loss (Note that the discriminator requires adversarial loss to distinguish the authenticity of images). Table V shows that the cycle-consistency constraint loss can greatly improve the performance of recall and F-Measure metrics, because this method prevents mode collapse. The F-measure metric is maximized when texture loss is not used (compared with other loss functions), which indicates that the texture has little influence on detection performance. Besides, the application of attention module can effectively introduce classification information and slightly improve the detection performance. Overall, by combing all components, our proposed image generation network remarkably outperforms all other variants.

F. Experiments for Supervised Foreground Detection Model Trained on Real or Synthetic Datasets

The above experiments verify the effectiveness of synthetic dataset, which can be used for the comprehensive evaluation of FD algorithms. However, if the proposed dataset can not be used in deep learning models, it will contribute little to the field. Therefore, our method uses real images, Synthetic-HS (CG) data and image generation networks to generate synthetic datasets. Then, we compare the performance of our method with other image generation methods including UNIT [44], NICE-GAN [46] and U-GAT-IT [47]. We randomly select the 1,133 images from real highway sequence and corresponding Synthetic-HS (CG) images as image pairs to train above image generation networks. Then, the Synthetic-HS (IT), Synthetic-HS (UNIT), Synthetic-HS (NICE-GAN), Synthetic-HS (U-GAT-IT) images can be generated. We select 50 images from real BMC (CG), V-KITTI (CG) datasets and above different domains as training sets. The supervised FD algorithm (MU-Net1 [57]) is trained on these domain datasets (Real, Synthetic-HS (CG), BMC (CG), V-KITTI (CG) and Synthetic-HS (IT)). The trained models are tested on 1,180 real images. Fig. 10 displays some results by using MU-Net1 models trained on real images and Synthetic-HS images. The numerical results are shown in Table VI. The results demonstrate that the performance of our dataset (Synthetic (CG)) is much better than traditional synthetic datasets (BMC (CG) and V-KITTI (CG)), which benefits from the overall similarity between synthetic datasets and real scenes. Moreover, we also find that the image generation method (Synthetic-HS (IT)) reduces the gap between datasets in different domains, which allows the Synthetic-HS (IT) images to be used in real scene. Compared with other models, it is obvious that the metrics (recall and F-Measure) of our model achieves the highest scores while maintaining a high precision value. In addition, the MU-Net1 model trained on the Synthetic-HS (IT) dataset, which can obtain more accurate foreground detection performance. Because the proposed image generation model not only translate high-quality foreground images, but also better retain the background object of the image. This allows the detector to achieve better visualization performance.

TABLE VI

PERFORMANCE OF MU-NET1 MODEL EVALUATED ON REAL OR SYNTHETIC DATASETS. NOTATION: '(REAL)' MEANS USING THE REAL DATASET AS THE TRAINING DATA, '(SYNTHETIC-HS (UNIT))' MEANS USING THE SYNTHETIC DATASET GENERATED BY THE UNIT MODEL AS TRAINING DATA

Training Dataset	Testing	Recall	Precision	F-Measure
Synthetic-HS (CG)	Real	**29.6%**	**98.7%**	**45.5%**
BMC (CG)	Real	18.8%	97.2%	31.5%
V-KITTI (CG)	Real	28.7%	68.0%	40.4%
Real (Baseline)	Real	69.2%	97.8%	81.1%
Synthetic-HS (UNIT)	Real	28.7%	18.6%	22.6%
Synthetic-HS (NICE-GAN)	Real	59.6%	**81.8%**	68.9%
Synthetic-HS (U-GAT-IT)	Real	66.9%	75.0%	70.7%
Synthetic-HS (IT)	Real	**69.5%**	74.6%	**71.9%**

G. Experiments for Supervised Foreground Detection Model Trained on Real and Mixed Datasets

Previous experimental results show that the image generation network can effectively narrow the gap between source

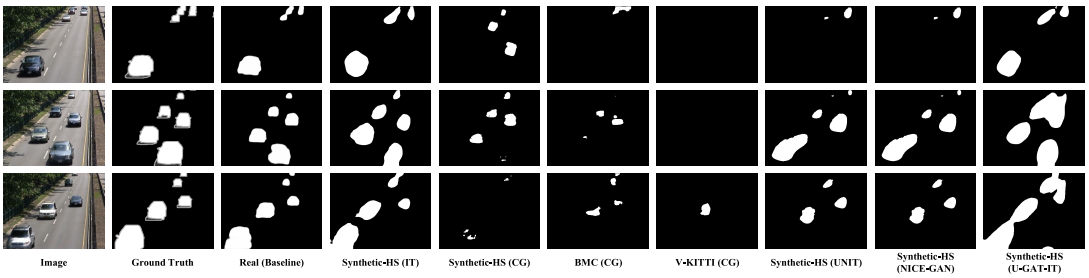

Fig. 10. Examples of foreground detection results on CDnet2014 (highway sequence) using the MU-Net1 model. Column 1: testing images; Column 2: ground truth masks; Column 3-10: the MU-Net1 model trained on the real images and Synthetic-HS images. Best viewed with zooming.

TABLE VII

PERFORMANCE OF MU-NET1 MODEL TRAINED ON REAL AND MIXED DATASETS. NOTATION: 'SYNTHETIC-HS (CG)' MEANS USING THE SYNTHETIC DATASET GENERATED BY COMPUTER GRAPHICS AS THE TRAINING DATA, 'SYNTHETIC-HS (IT)' MEANS USING THE SYNTHETIC DATASET GENERATED BY OUR GENERATIVE ADVERSARIAL NETWORK AS THE TRAINING DATA

Training Dataset	Testing	Recall	Precision	F-Measure
Real (baseline)	Real	69.2%	97.8%	81.1%
Real + Synthetic-HS (CG)	Real	75.8%	96.7%	84.9%
Real + Synthetic-HS (IT)	Real	**76.6%**	**99.6%**	**86.5%**

and target domain. More importantly, researchers expect to use synthetic data to train models applied in real-world settings. In this scheme, we randomly select 50 images from real and synthetic images (Synthetic-HS (CG) and Synthetic-HS (IT)) as training sets. The real (baseline) and mixed datasets (Synthetic-HS (CG) and Synthetic-HS (IT)) are used to train the MU-Net1 model, and the trained models are tested on 1,180 real images. The performance of experimental results is shown in Table VII. Compared with baseline, the recall of models trained on "real + Synthetic-HS (IT)" is significantly increased by 7.4%, and the precision is improved slightly by 1.8%. The F-measure represents the balance of these two metrics and also boosted by nearly 5.4%. In addition, the training data mixed with Synthetic-HS (IT) has higher metrics than Synthetic-HS (CG). This suggests that the synthetic data generated by our well-designed model can help supervised models to capture more detailed features. Therefore, this approach improves visual perception performance of highway surveillance in the real world.

VI. CONCLUSION

In our experiments, we have trained and evaluated different FD models for the challenges of highway surveillance. The experimental results prove that computer graphics can automatically generate more accurate binary labels and be used for detector training and testing. In addition, we create a variety of Synthetic-HS (CG) sequences so that FD models can be quantitatively analyzed under different imaging conditions. Although most of the FD approaches exhibit flaws in some experiments, we find the SuBSENSE and LBP_MRF to be the most promising ones. But it is also worth noting that the results of some video challenges (such as foggy, light switch and thunder storm) show quite lower performance for all evaluated approaches. This offers a direction for future research in the area. Besides, our method uses image generation networks to generate realistic synthetic dataset. The ablation study measures the effect of different components on image translation performance. More importantly, the supervised model trained by our datasets (Synthetic-HS (IT)) can greatly promote the performance in practical applications compared to other models.

In this paper, we present challenging synthetic highway scenarios for promoting video visual intelligence. In the synthetic scenarios, accurate and consistent binary masks can be automatically generated by computer graphics. What is more, the synthetic scenarios are controllable, considerable, and repeatable. This allows us to design multi-challenge video datasets, including: basic, dynamic background, illumination variation. These characteristics make synthetic datasets particularly suitable for algorithm's training and result evaluation. Experimental results demonstrate that the proposed Synthetic-HS(CG) dataset can quantitatively analyze the potential influence of each element on the FD algorithms. Furthermore, we propose an image translation network, which consists of U-Net with skip connections and GAN with attention module. This method narrows the gap between source and target images and synthesize photo-realistic images. Thus, the results show that the proposed Synthetic-HS (IT) datasets can improve the performance of the supervised FD model in real scenarios. The future research can be conducted from the following aspects: (1) since adversarial training can benefit from loss functions, we will present or use novel loss functions (such as dual contrastive loss [58]) which encourage the distinguishability power of the discriminator representations for their classification task. (2) some results [59] show that the improved performance come from structure itself, and we will explore the creative modules to promote image generation quality.

REFERENCES

[1] B. Tian et al., "Hierarchical and networked vehicle surveillance in its: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 25–48, Jan. 2017.

[2] S. Wan, X. Xu, T. Wang, and Z. Gu, "An intelligent video analysis method for abnormal event detection in intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4487–4495, Jul. 2021.

[3] H. Duan, Y. Sun, and Y. Shi, "Bionic visual control for probe-and-drogue autonomous aerial refueling," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 2, pp. 848–865, Apr. 2021.

[4] X. Li, H. Duan, J. Li, Y. Deng, and F.-Y. Wang, "Biological eagle eye-based method for change detection in water scenes," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108203.

[5] J.-J. Qiao, X. Wu, J.-Y. He, W. Li, and Q. Peng, "SWNet: A deep learning based approach for splashed water detection on road," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3012–3025, Apr. 2022.

[6] W. J. Kim, S. Hwang, J. Lee, S. Woo, and S. Lee, "AIBM: Accurate and instant background modeling for moving object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9021–9036, Jul. 2022.

[7] M. Shah, J. Deng, and B. Woodford, "Illumination invariant background model using mixture of Gaussians and SURF features," in *Proc. Asian Conf. Comput. Vis.*, Daejeon, South Korea, 2012, pp. 308–314.

[8] B. N. Subudhi, M. K. Panda, T. Veerakumar, V. Jakhetiya, and S. Esakkirajan, "Kernel-induced possibilistic fuzzy associate background subtraction for video scene," *IEEE Trans. Computat. Social Syst.*, early access, Jan. 13, 2022, doi: 10.1109/TCSS.2021.3137306.

[9] L. Maddalena and A. Petrosino, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 179–186, 2010.

[10] H. Sajid and S.-C. S. Cheung, "Universal multimode background subtraction," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3249–3260, Jul. 2017.

[11] A. Babaryka, I. Katerynchuk, and O. Komarnytska, "Technologies for building intelligent video surveillance systems and methods for background subtraction in video sequences," *Future Intent-Based Netw.*, pp. 468–480, 2022.

[12] S. Javed, P. Narayanamurthy, T. Bouwmans, and N. Vaswani, "Robust PCA and robust subspace tracking: A comparative evaluation," in *Proc. IEEE Stat. Signal Process. Workshop (SSP)*, Jun. 2018, pp. 836–840.

[13] 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.*, vol. 23, no. 7, pp. 6101–6122, Jul. 2022.

[14] J. H. Giraldo, S. Javed, N. Werghi, and T. Bouwmans, "Graph CNN for moving object detection in complex environments from unseen videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 225–233.

[15] T. Minematsu, A. Shimada, and R.-I. Taniguchi, "Rethinking background and foreground in deep neural network-based background subtraction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3229–3233.

[16] W. Zheng, K. Wang, and F.-Y. Wang, "A novel background subtraction algorithm based on parallel vision and Bayesian GANs," *Neurocomputing*, vol. 394, pp. 178–200, Jun. 2020.

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

[18] J. H. Giraldo, S. Javed, and T. Bouwmans, "Graph moving object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2485–2503, 2020.

[19] J. H. Giraldo et al., "The emerging field of graph signal processing for moving object segmentation," in *Proc. Int. Workshop Frot. Comput. Vis.*, Daegu, South Korea, 2021, pp. 31–45.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[21] X. Li, K. Wang, Y. Tian, L. Yan, F. Deng, and F.-Y. Wang, "The ParallelEye dataset: A large collection of virtual images for traffic vision research," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2072–2084, Jun. 2019.

[22] T. Q. Phan, P. Shivakumara, S. Bhowmick, S. Li, C. L. Tan, and U. Pal, "Semiautomatic ground truth generation for text detection and recognition in video images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1277–1287, Aug. 2014.

[23] X. Li, P. Ye, J. Li, Z. Liu, L. Cao, and F.-Y. Wang, "From features engineering to scenarios engineering for trustworthy AI: I&I, C&C, and V&V," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 18–26, Jul./Aug. 2022.

[24] X. Li, Y. Tian, P. Ye, H. Duan, and F.-Y. Wang, "A novel scenarios engineering methodology for foundation models in metaverse," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Dec. 26, 2022, doi: 10.1109/TSMC.2022.3228594.

[25] X. Li, Y. Wang, L. Yan, K. Wang, F. Deng, and F.-Y. Wang, "ParallelEye-CS: A new dataset of synthetic images for testing the visual intelligence of intelligent vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 9619–9631, Oct. 2019.

[26] T. Shen, C. Gou, J. Wang, and F.-Y. Wang, "Simultaneous segmentation and classification of mass region from mammograms using a mixed-supervision guided deep model," *IEEE Signal Process. Lett.*, vol. 27, pp. 196–200, 2020.

[27] W. S. Bainbridge, "The scientific research potential of virtual worlds," *Science*, vol. 317, no. 5837, pp. 472–476, 2007.

[28] V. S. Subrahmanian and J. Dickerson, "What can virtual worlds and games do for national security?" *Science*, vol. 326, no. 5957, pp. 1201–1202, Nov. 2009.

[29] Y. Zhang, W. Qiu, Q. Chen, X. Hu, and A. Yuille, "UnrealStereo: Controlling hazardous factors to analyze stereo vision," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 228–237.

[30] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, May 2016, pp. 4340–4349.

[31] F. Lateef, M. Kas, and Y. Ruichek, "Saliency heat-map as visual attention for autonomous driving using generative adversarial network (GAN)," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5360–5373, Jun. 2022.

[32] K. Zhang, Y. Zhang, and H. D. Cheng, "CrackGAN: Pavement crack detection using partially accurate ground truths based on generative adversarial learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1306–1319, Feb. 2021.

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

[34] D. P. Young and J. M. Ferryman, "PETS metrics: On-line performance evaluation service," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2005, pp. 317–325.

[35] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.

[36] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "A novel video dataset for change detection benchmarking," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4663–4679, Nov. 2014.

[37] F. Tiburzi, M. Escudero, J. Bescos, and J. M. Martinez, "A ground truth for motion-based video-object segmentation," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 17–20.

[38] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proc. CVPR*, Jun. 2011, pp. 1937–1944.

[39] A. Vacavant, T. Chateau, A. Wilhelm, and L. Lequievre, "A benchmark dataset for outdoor foreground/background extraction," in *Proc. Asian Conf. Comput. Vis.*, Daejeon, South Korea, 2012, pp. 291–300.

[40] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.

[41] X. Yue, B. Wu, S. A. Seshia, K. Keutzer, and A. L. Sangiovanni-Vincentelli, "A LiDAR point cloud generator: From a virtual world to autonomous driving," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2018, pp. 458–464.

[42] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8198–8207.

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[44] M. Liu et al., "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 700–708.

[45] H. Lee et al., "Diverse image-to-image translation via disentangled representations," in *Proc. Euro. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 35–51.

[46] J. Kim et al., "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translationn," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–19.

[47] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8168–8177.

[48] K. Ye, Y. Ye, M. Yang, and B. Hu, "Independent encoder for deep hierarchical unsupervised image-to-image translation," 2021, *arXiv:2107.02494*.
[49] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 700–708.
[50] T. Salimans et al., "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 29–39.
[51] Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Comparative study of background subtraction algorithms," *J. Elec. Imag.*, vol. 19, pp. 1–12, Jul. 2010.
[52] 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.
[53] 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.
[54] P.-L. St-Charles and G.-A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 509–515.
[55] 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.
[56] C. Kertész, "Texture-based foreground detection," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 4, no. 4, pp. 51–62, 2011.
[57] G. Rahmon, F. Bunyak, G. Seetharaman, and K. Palaniappan, "Motion U-Net: Multi-cue encoder–decoder network for motion segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8125–8132.
[58] N. Yu et al., "Dual contrastive loss and attention for GANs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6731–6742.
[59] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

Xuan Li received the Ph.D. degree in control science and engineering from the Beijing Institute of Technology, Beijing, China, in 2020. He joined the Peng Cheng Laboratory and became an Assistant Professor with the Virtual Reality Laboratory. From October 2018 to October 2019, he was a Visiting Scholar with the Department of Computer Science, Stony Brook University, Stony Brook, NY, USA. His research interests include image synthesis, computer vision, bionic vision computing, intelligent transportation systems, and machine learning.

Haibin Duan (Senior Member, IEEE) is currently a Full Professor with the School of Automation Science and Electrical Engineering, Beihang University. He is also the Vice Director of the State Key Laboratory of Virtual Reality Technology and Systems and also the Head of the Bio-Inspired Autonomous Flight Systems (BAFS) Research Group. He received the National Science Fund for Distinguished Young Scholars of China. He is also enrolled in the Scientific and Technological Innovation Leading Talent of Ten Thousand Plan-National High Level Talents Special Support Plan. He has authored or coauthored more than 70 publications and three monographs. His current research interests are bio-inspired computing, biological computer vision, and multi-UAV autonomous formation control.

Bingzi Liu received the M.Sc. degree from the School of Software and Microelectronics, Peking University, Beijing, China, in 2015. She was an Assistant Engineer with the Institute of Automation, Chinese Academy of Sciences, from 2015 to 2018. She is currently a Engineer with the Peng Cheng Laboratory. Her research interests include software engineering and software crowdsourcing.

Xiao Wang (Member, IEEE) received the bachelor's degree in network engineering from the Dalian University of Technology, Dalian, China, in 2011, and the Ph.D. degree in social computing from the University of Chinese Academy of Sciences, Beijing, China, in 2016. She is currently an Associate Professor with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences. She has published more than a dozen SCI/EI articles and translated three technical books (English to Chinese). Her research interests include social transportation, cybermovement organizations, artificial intelligence, and social network analysis.

Fei-Yue Wang (Fellow, IEEE) received the Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. He joined the University of Arizona in 1990 and became a Professor and the Director of the Robotics and Automation Laboratory (RAL) and Program in Advanced Research for Complex Systems (PARCS). In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Oversea Chinese Talents Program from the State Planning Council and "100 Talent Program" from CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory for Management and Control of Complex Systems. His current research focuses on methods and applications for parallel intelligence, social computing, and knowledge automation. He was the Founding Editor-in-Chief (EiC) of the *International Journal of Intelligent Control and Systems* from 1995 to 2000, the *IEEE ITS Magazine* from 2006 to 2007, the IEEE/CAA JOURNAL OF AUTOMATICA SINICA from 2014 to 2017, and the *Journal of Command and Control* (China) from 2015 to 2020. He was the Editor-in-Chief (EiC) of the IEEE INTELLIGENT SYSTEMS from 2009 to 2012 and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS from 2009 to 2016. He is also the Editor-in-Chief (EiC) of the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS since 2017 and the Founding Editor-in-Chief of *Journal of Intelligent Science and Technology* (China) since 2019. Currently, he is the President of CAA's Supervision Council, IEEE Council on RFID, and the Vice President of IEEE Systems, Man, and Cybernetics Society.