

# To See in the Dark: N2DGAN for Background Modeling in Nighttime Scene

Zhenfeng Zhu<sup>1</sup>, Yingying Meng, Deqiang Kong, Xingxing Zhang<sup>1</sup>,  
Yandong Guo, and Yao Zhao<sup>1</sup>, *Senior Member, IEEE*

**Abstract**—Due to the deteriorated conditions of illumination lack and uneven lighting, the performance of traditional background modeling methods is greatly limited for the surveillance of nighttime video. To make background modeling under nighttime scene performs as well as in daytime condition, we put forward a promising generation-based background modeling framework for foreground surveillance. With a pre-specified daytime reference image as background frame, the GAN based generation model, called N2DGAN, is trained to transfer each frame of nighttime video to a virtual daytime image with the same scene to the reference image except for the foreground part. Specifically, to balance the preservation of background scene and the foreground object(s) in generating the virtual daytime image, we presented a two-pathway generation model, in which the global and local sub-networks were well combined with spatial and temporal consistency constraints. For the sequence of generated virtual daytime images, a multi-scale Bayes model was further proposed to characterize pertinently the temporal variation of background. We manually labeled ground truth on the collected nighttime video datasets for performance evaluation. The impressive results illustrated in both the main paper and supplementary show the effectiveness of our proposed approach.

**Index Terms**—GAN, background model, foreground detection, Bayes theory.

## I. INTRODUCTION

**B**ACKGROUND modeling originates in numerous applications, especially in visual surveillance [1]–[7]. In the last decades, state-of-the-art approaches for background modeling have been proposed for visual surveillance under daytime scenes. On a whole, they are popularly dominated by a family

Manuscript received November 9, 2019; revised March 5, 2020; accepted March 26, 2020. Date of publication April 15, 2020; date of current version February 4, 2021. This work was supported in part by the Science and Technology Innovation 2030—New Generation Artificial Intelligence Major Project under Grant 2018AAA0102101, in part by the National Natural Science Foundation of China under Grant 61976018 and Grant 61532005, and in part by the Fundamental Research Funds for the Central Universities under Grant 2018JBZ001. This article was recommended by Associate Editor W.-H. Peng. (*Corresponding author: Zhenfeng Zhu.*)

Zhenfeng Zhu, Yingying Meng, Xingxing Zhang, and Yao Zhao are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: zhfhzhu@bjtu.edu.cn; mengyingying@bjtu.edu.cn; zhangxing@bjtu.edu.cn; yzhao@bjtu.edu.cn).

Deqiang Kong is with Microsoft Multimedia, Beijing 100080, China (e-mail: kodeqian@microsoft.com).

Yandong Guo is with the OPPO Research Institute, Beijing 100026, China (e-mail: guoyandong@oppo.com).

Color versions of one or more of the figures in this article are available online at <https://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCSVT.2020.2987874

1051-8215 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.  
See <https://www.ieee.org/publications/rights/index.html> for more information.

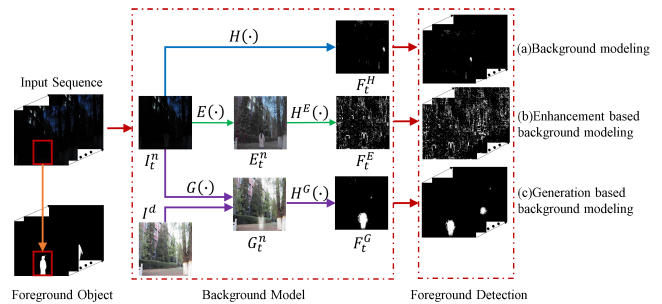


Fig. 1. Flowchart of three kinds of background modeling methods for foreground object detection in nighttime surveillance video. (a) Conventional. (b) Enhancement-based. (c) Generation-based.

of statistical based methods, like GMM [4] and KDE [2]. Besides, Codebook [3] and ViBe [1] are also two representative methods that achieve good performance for modeling the background. Most recently, several deep learning based works [5]–[7] for background modeling were also proposed.

Despite the achievements by these approaches, all of them face quite a challenge in the case of illumination lack and uneven lighting at night, especially in the presence of dynamic background, change of light, and some extreme weather conditions such as rain, snow and fog. As we can see from Fig.1 (a), the conventional background modeling methods, like GMM, etc., fail to distinguish the foreground object from background due to the deteriorated condition of illumination lack. To deal with such a case, an intuitive way as shown in Fig.1 (b) is to perform image enhancement through  $E(\cdot)$  first, and then build background model  $H^E(\cdot)$  on the bases of the enhanced frames  $E_t^n$ 's, just like background modeling under daytime scene. However, since these enhancement methods [8]–[10] are not task-driven, they usually lose sight of pixel-wise consistency of inter-frame, and yet it is of great significance for background modeling.

Although the captured images can be brightened by equipping a light with the camera, there are still deficiencies in the use of these devices. First, it will inevitably increase the cost of surveillance. Second, compared with the imaging under the natural light in the daytime, the visual quality by this way is still unsatisfactory. For example, some problems including serious noise, blurring, and unbalanced illumination will further bring difficulties for some vision processing tasks including object detection. In addition, it also won't work well for monitoring distant object.

To address this issue, we make a novel contribution in integrating generative model into background modeling. Fig.1 (c) shows the proposed generation-based background modeling framework. With a pre-specified daytime reference image  $I^d$  as ground-truth background frame, the generation model  $G(\cdot)$  is trained for transferring each frame  $I_t^n$  of nighttime video to a virtual daytime image  $G_t^n$  with the same scene to the reference image except for the foreground region. Furthermore, the background model  $H_G(\cdot)$  can be built to obtain  $F_t^G$  with the detected foreground object. In fact, the unique reference image plays a significant role for enforcing the pixel-wise temporal consistency of inter-frames in the generation of virtual daytime images.

To the best of our knowledge, this paper is one of the first attempts to introduce GANs based deep learning network for background modeling. In summary, the following points highlight several contributions of the paper:

- This paper proposes a reasonable and innovative solution, i.e., N2DGAN, to the longstanding problem of foreground object detection under nighttime scene. As a promising generation-based framework, it makes background modeling work as well as in daytime condition.
- To simultaneously preserve background scene and the foreground object(s) in generating the virtual daytime image, we present a two-pathway generation model, in which the global and local sub-networks are seamlessly combined with spatial and temporal consistency constraints.
- For the sequence of generated virtual daytime images, a multi-scale Bayes model is proposed to characterize pertinently the temporal variation of background. Thus, while suppressing effectively noise coming from virtual daytime image generation, we can ensure the favorable detection of foreground objects.
- We collect a benchmark dataset including indoor and outdoor scenes with manually labeled ground truth, which can serve as a good benchmark for the research community.

## II. RELATED WORK

### A. Nighttime Image Enhancement

Here we simply divide image enhancement methods into two categories: reference based and non-reference based methods.

Non-reference based methods mainly focus on how to improve low contrast images. As a naive method, Histogram Equalization (HE) [9] spreads out the most frequent intensity values, thus gaining a higher contrast for the areas of lower contrast. The purpose of Retinex based image enhancement [MSR] [8] is to estimate the illumination from original image, thereby decomposing reflectance image and eliminating the influence of uneven illumination. In recent years, some deep learning based low light image enhancement approaches were also proposed, such as LIME [44], LLNet [45], and Struct [46]. Although they generally can achieve better perceptual quality than HE and MSR, they highly depend on the amount of training data.

Reference based methods [11]–[13] usually combine images of a scene at different time intervals by image fusion. These methods usually produce unnatural effects in the enhanced images. Besides, it would increase signal to noise ratio, which is adverse for further video analysis and applications such as foreground detection.

### B. Generative Adversarial Nets

As a novel way to train generative models, GANs [14] proposed by Goodfellow et al. has received extensive applications in various of visual tasks [15]–[20]. In [15], GANs was applied for image completion with globally and locally consistent adversarial training. Reference [16] used back-propagation on a pretrained image generative network for image inpainting. To transfer the original image into a cartoon style, domain transfer network (DTN) [17] was proposed. In [18], [19], GANs has been employed for image super-resolution and image deblurring. Recently, a general-purpose solution to image-to-image translation based on conditional adversarial networks, also known as pixel2pixel network, was proposed in [20], and shows good performances on a variety of tasks like photo generation and semantic segmentation. In our previous work [21], a generative adversarial networks (GANs) based framework for nighttime image enhancement was proposed.

### C. Background Modeling Algorithms

Broadly speaking, background modeling methods can be divided into two categories: pixel-based methods and block-based methods.

One of the most popular pixel-based methods is Gaussian mixture models(GMM) [4], [22], [23]. It models the distribution at each pixel observed over time using a summation of weighted Gaussian distribution. Such methods generally perform well with the multi-modal nature of many practical situations. However, if high or low frequency changes appear in the background, the model can't be adaptively tuned in time and even may miss some information about fast moving objects. Consequently, Elgammal *et al.* [24] have developed a non-parametric background model, which estimates the probability of observing pixel intensity values based on a sample of intensity values for each pixel.

Different from GMM and KDE, some deep learning based works [5]–[7] for background modeling have also been proposed in recent years. But these models are all supervised and require many manually labeled data for model training. Thus, they obviously lack of scalability to a scene unknown beforehand, which also means their performances greatly depend on the collected training datasets.

Block-based methods [25], [26] divide each frame into multiple overlapped or non-overlapped small blocks, and then model the background using the features of each block. Compared with pixel-based methods, the image blocks can capture more spatial distribution information, which makes block-based methods insensitive to the local shift in the background. However, the detection performance will largely depend on the block-dividing technique, especially for small moving targets.

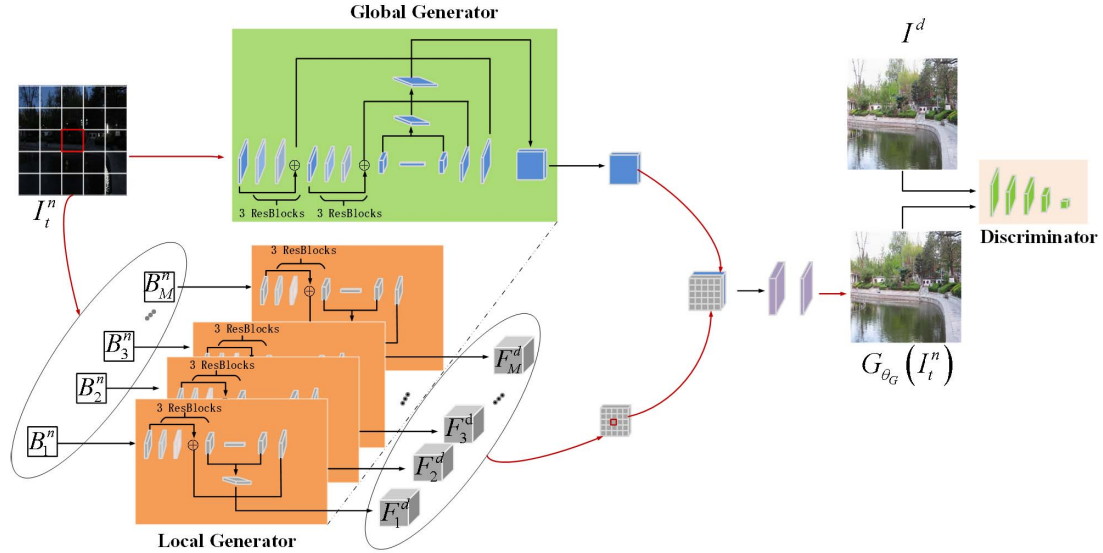


Fig. 2. The network architecture of N2DGAN.  $I_t^n$  is the input nighttime image, and we divide it into  $M$  blocks ( $B_i^n, i = 1, 2, \dots, M$ ) as the input of each local generator. Then, the output of global generator and local generators are concatenated together. Finally, after two convolutional layers, the output is the daytime image  $G_{\theta_G}(I_t^n)$ , where  $I^d$  is the reference daytime image. More details about our model architecture are provided in the appendix.

### III. NIGHTTIME TO DAYTIME GENERATIVE ADVERSARIAL NETWORKS (N2DGAN)

To maintain spatial and temporal consistency in the generation process, our goal is to train a generation model  $G(\cdot)$  as in Fig.1 to transfer each frame  $I_t^n$  of nighttime video to a virtual daytime image with the same scene to the unique reference image  $I^d$  except for the foreground region. Specifically, this generation problem can be formulated as:

$$\hat{\theta}_G = \underset{\theta_G}{\operatorname{argmin}} \frac{1}{N} \sum_{t=1}^N L(G_{\theta_G}(I_t^n), I^d) \quad (1)$$

where  $N$  is the number of training pairs,  $L(\cdot, \cdot)$  denotes a weighted combination of several loss components.

For the intent of learning the generation function  $G(\cdot)$ , the GANs is applied [14] due to its powerful generating ability. In particular, we propose a two pathway network N2DGAN with a generator network  $G_{\theta_G}$  and a discriminator network  $D_{\theta_D}$  parameterized by  $\theta_G$  and  $\theta_D$ , respectively. For the discriminator network  $D_{\theta_D}$ , we will have the following maximization problem given  $G_{\theta_G}$ :

$$\hat{\theta}_D = \underset{\theta_D \in \mathcal{D}}{\operatorname{argmax}} E_{I^d \sim P_d}[D_{\theta_D}(I^d)] - E_{I_t^n \sim P_n}[D_{\theta_D}(G_{\theta_G}(I_t^n))] \quad (2)$$

where  $P_d$  is the real daytime image distribution, and  $P_n$  is the nighttime image distribution. Here we adopt the same formulation for Eq.(2) as in WGAN [27],  $\mathcal{D}$  is the set of 1-Lipschitz function, and weight clipping is utilized to enforce the Lipschitz constraint. In essence, Eq.(1) and Eq.(2) are the alternative optimizing of a min-max optimization problem jointly parameterized by  $\theta_G$  and  $\theta_D$ .

#### A. Architecture

The overview on the network architecture of the proposed N2DGAN is shown in Fig.2. To leverage the preserving

of background scene and foreground object(s) in generating virtual daytime image, a two-pathway generator is proposed with global sub-network for maintaining background scene and  $M$  local sub-networks attending to capture local foreground information. As illustrated in Fig.2, both the global and local sub-networks are designed in an Encoder-Decoder manner as in most cases with modules of norm Convolution-BatchNorm-Relu, and each layer is followed by three residual blocks [28]. Following the architecture of ‘‘U-Net’’ adopted in [20], a fusion subnet is also designed to connect both the ‘‘Encoder’’ and the ‘‘Decoder’’ since symmetric layers can share some common information. This will be helpful for facilitating the information flow of foreground object between the input and output in the network chain. The details of the model architecture are provided in the Appendix.

#### B. Loss Function

The loss function  $L(\cdot, \cdot)$  in Eq.(1) plays a significant role in training GANs model. For an input nighttime image  $I_t^n$ , several kinds of loss functions are exploited to make the generated virtual daytime image  $G_{\theta_G}(I_t^n)$  retain most of the image information such as structure, objects, and texture as in the pre-specified reference image  $I^d$ .

1) *Adversarial Loss*: To encourage the generated images move towards the real daytime image manifold and generate images with more details, the adversarial loss is first considered for distinguishing the generated image  $G_{\theta_G}(I_t^n)$  from the daytime image  $I^d$ .

$$L_{adv} = \frac{1}{N} \sum_{t=1}^N -D_{\theta_D}(G_{\theta_G}(I_t^n)) \quad (3)$$

where  $N$  is the number of nighttime images in the training dataset.

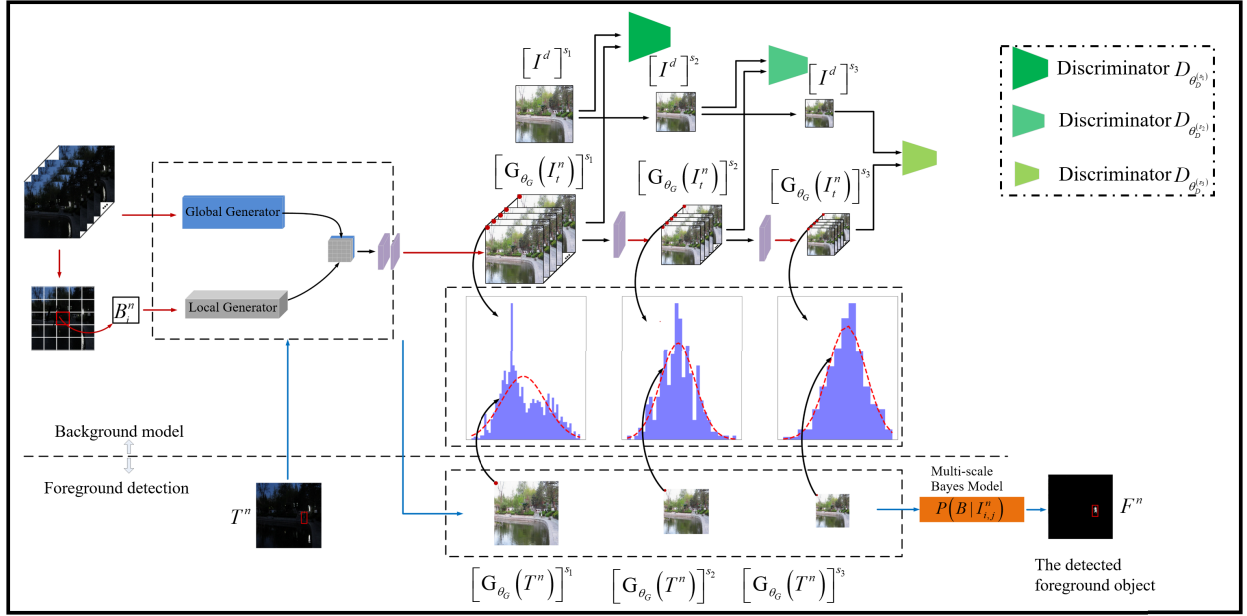


Fig. 3. Multi-scale Bayes inference framework for foreground detection. In the training phase, for input frame  $I_t^n$ , multi-scale daytime images  $[G_{\theta_G}(I_t^n)]^{s_i}$ ,  $i = 1, 2, \dots, Q$ , are generated, then we model the temporal distribution of each pixel of the generated image at each scale.  $D_{\theta_D^{(s_i)}}$  is discriminator network for scale  $s_i$ . In the testing phase, that is, foreground detection phase, for the input frame  $T_t^n$ , the pre-trained N2DGAN network outputs its corresponding multi-scale daytime images  $[G_{\theta_G}(T_t^n)]^{s_i}$ ,  $i = 1, 2, \dots, Q$ . Then the multiple probabilities of each pixel belonging to the background at each scale are integrated elegantly based on Bayes inference for detecting the foreground object.

2) *Perceptual Loss*: In order to minimize the high-level perceptual and semantic differences between  $G_{\theta_G}(I_t^n)$  and  $I^d$  while preventing unexpected overfitting coming from  $I^d$ , we follow the idea that minimizes the difference in convolutional layer of a pre-trained network [29] between two images. The motivation behind it lies in that the neural network pre-trained by image classification task has already learnt effective representation, which can be transferred into other tasks such as our enhancement processing. Specifically, we define  $\phi_i$  as the activation of the  $i^{\text{th}}$  convolutional layer of the pre-trained network, and the perceptual loss is defined as:

$$L_p = \frac{1}{C} \sum_{i=1}^C \frac{1}{W_i} \frac{1}{H_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} \left( [\phi_i(G_{\theta_G}(I_t^n))]_{x,y} - [\phi_i(I^d)]_{x,y} \right)^2 \quad (4)$$

where  $C$  is the number of convolutional layers,  $W_i$  and  $H_i$  describe the dimensions of the respective feature maps within the VGG network.

3) *Pixel-Wise Loss*: To facilitate further background modeling task with spatial consistency of intra-frame and pixel-wise temporal consistency of inter-frame, the most widely used pixel-wise MSE loss (Eq.(5)) and total variation loss (Eq.(6)) are also adopted.

$$L_{mse} = \frac{1}{W} \frac{1}{H} \sum_{i=1}^W \sum_{j=1}^H \left( I_{i,j}^d - [G_{\theta_G}(I_t^n)]_{i,j} \right)^2 \quad (5)$$

$$L_{tv} = \sum_{i=1, j=1}^{W, H} \sqrt{\left( (\hat{I}_t^d)_{i+1,j} - (\hat{I}_t^d)_{i,j} \right)^2 + \left( (\hat{I}_t^d)_{i,j+1} - (\hat{I}_t^d)_{i,j} \right)^2} \quad (6)$$

where  $\hat{I}_t^d$  denotes the generated virtual daytime image  $G_{\theta_G}(I_t^n)$ .

Since each of the loss functions mentioned above is provided with a unique view on characterizing the visual quality of the generated virtual image, an intuitive way is to make a combination of them. Thus, we have the final overall loss function as:

$$L = \lambda_{adv} L_{adv} + \lambda_{tv} L_{tv} + \lambda_{mse} L_{mse} + \lambda_p L_p \quad (7)$$

where  $\lambda_{adv}$ ,  $\lambda_{tv}$ ,  $\lambda_{mse}$ , and  $\lambda_p$  are weights of the corresponding terms, respectively.

#### IV. MULTI-SCALE BAYES INFERENCE FOR FOREGROUND DETECTION

N2DGAN ensures that there is a detectable difference between foreground object and background. All these characters match the major premise of GMM, that the background is more frequently visible than the foreground and that its variance is significantly slight. However, as we know, the neural network has the properties of both randomness and uncertainty. Thus, there exists inevitably pixel-wise difference between  $I_{i,j}^d$  and  $[G_{\theta_G}(I_t^n)]_{i,j}$  in generating  $G_{\theta_G}(I_t^n)$ , and it essentially can be regarded as some kind of random noise arising from both spatial and temporal domains. In other words, given the total error value, this difference at pixel  $(i, j)$  may also occur at any other pixels with equal probability. This case will be doomed to bring some unexpected negative influence on pixel-level background modeling.

To mitigate this issue, inspired by some works on multi-scale multiplication for edge detection [37], [38] that tend to yield significant localized detection, we extend N2DGAN to a multi-scale generative model as shown in Fig.3 to facilitate the background modeling to be noise-free.

### A. Multi-Scale Generation

As illustrated in Fig.3, we reformulate the generation problem for background modeling as follows: to train a generator network  $G_{\theta_G}$  parametrized by  $\theta_G$ , which learns a mapping function from the source domain  $\Omega_n$  of nighttime to the target domain  $\Omega_d$  of daytime. For every input nighttime frame  $I_t^n \in \Omega_n$ , to generate a multi-scale set of images  $[G_{\theta_G}(I_t^n)]^{s_i}$ ,  $i = 1, \dots, Q$ , in daytime domain  $\Omega_d$  will be equivalent to:

$$\hat{\theta}_G = \frac{1}{N} \underset{\theta_G}{\operatorname{argmin}} \sum_{t=1}^N \sum_{i=1}^Q L([G_{\theta_G}(I_t^n)]^{s_i}, [I^d]^{s_i}) \quad (8)$$

where  $N$  is the number of training pairs as before, and  $L(\cdot, \cdot)$  denotes the loss function as mentioned above. In addition, we introduce  $Q$  adversarial discriminators to distinguish the reference daytime image  $[I^d]^{s_i}$  from the generated virtual daytime image  $[G_{\theta_G}(I_t^n)]^{s_i}$  under scale  $s_i$ . Particularly, similar to Eq.(2), in each generation task of different scales, the discriminator network will be reformulated as:

$$\hat{\theta}_D^{(s_i)} = \underset{\theta_D^{(s_i)} \in \mathcal{D}}{\operatorname{argmax}} E_{[I^d]^{s_i} \sim P_d} [D_{\theta_D^{(s_i)}}([I^d]^{s_i})] - E_{I^n \sim P_n} [D_{\theta_D^{(s_i)}}([G_{\theta_G}(I^n)]^{s_i})] \quad (9)$$

It should be noted that both the generative network architecture and the discriminator architecture in Fig.3 are same as those in Fig.2. But different from N2DGAN, more convolutional layers are employed to generate multi-scale daytime images.

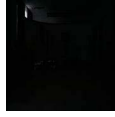

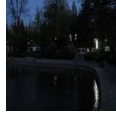



### B. Multi-Scale Bayes Inference Based on Scale Multiplication

N2DGAN enforces the sequence of generated virtual daytime images  $G_{\theta_G}(I_t^n)$ ,  $t = 1, \dots, N$ , to be as approximated closely as possible to the pre-specified unique reference frame  $I^d$ . The inescapable fact, however, is that there is certain difference between them, one is noise accompanied by the neural network, the other part is the foreground region. To suppress effectively noise coming from virtual daytime generation while strengthening the discriminant of foreground object, a multi-scale Bayes model is proposed to characterize pertinently the temporal variation of background.

For each pixel  $I_{i,j}^n$ , we use  $P(B | I_{i,j}^n)$  to serve as the background model, denoting the probability of pixel  $I_{i,j}^n$  to be background. Given the multi-scale representations  $[G_{\theta_G}(I^n)]_{i,j}^{s_k}$ ,  $k = 1, \dots, Q$ , for pixel  $I_{i,j}^n$ , the background model  $P(B | I_{i,j}^n)$  can be given with Bayes criterion by Eq.(10). Here,  $\lfloor \cdot \rfloor$  represents rounding down to the nearest

TABLE I

THE DETAILS OF OUR FOUR DATASETS

| Datasets        | Lab  | Tree  | Lake1   | Lake2 |
|-----------------|--|---|---|-------|
| #Frames         | 693  | 722   | 579   | 1482  |
| Nighttime Scene |  |  |  |       |
| Reference Image |  |  |  |       |

whole number. On the assumption that the generation of virtual daytime images with different scales is independent of each other, thus the background model given by Eq.(10), as shown at the bottom of this page, will further reduce to the following Eq.(11), as shown at the bottom of this page. As we can see from Eq.(11), the background model  $P(B | I_{i,j}^n)$  is equivalent to the multi-scale multiplication of multiple background models at different scales. In addition, for each background model  $P(B | [G_{\theta_G}(I^n)]_{\lfloor i/2^{k-1} \rfloor, \lfloor j/2^{k-1} \rfloor}^{s_k})$ ,  $k = 1, \dots, Q$ , a single gaussian model instead of GMM can be simply applied as shown in Fig.3.

## V. EXPERIMENTAL RESULTS

We evaluate the proposed background modeling approach visually and quantitatively, by comparing with state-of-the-arts and providing extensive ablation studies

### A. Datasets and Experiment Settings

1) *Datasets and Evaluation Metrics*: Our work in this paper mainly focuses on background modeling under nighttime scene with low illumination. However, to the best of our knowledge, there are no public open datasets to evaluate such a task. For this reason, we collect several benchmark datasets by a Canon IXY 210F video camera including indoor and outdoor scenes with manually labeled ground truth. The details about our four datasets, including **Lab**, **Tree**, **Lake1**, and **Lake2**, are shown in Tab.I.<sup>1</sup> For each dataset, the corresponding pre-specified daytime images that serve as ground truth background frames are also provided. It is worth of noting that both the ‘Lake’ and ‘Tree’ datasets were captured outdoor on windy

<sup>1</sup>The datasets and code of our method will be released at <https://github.com/anqier0468/N2DGAN>.

$$P(B | I_{i,j}^n) = P(B | [G_{\theta_G}(I^n)]_{i,j}^{s_1}, \dots, [G_{\theta_G}(I^n)]_{\lfloor i/2^{Q-1} \rfloor, \lfloor j/2^{Q-1} \rfloor}^{s_Q}) P([G_{\theta_G}(I^n)]_{i,j}^{s_1}, \dots, [G_{\theta_G}(I^n)]_{\lfloor i/2^{Q-1} \rfloor, \lfloor j/2^{Q-1} \rfloor}^{s_Q} | I_{i,j}^n) \quad (10)$$

$$\begin{aligned} P(B | I_{i,j}^n) &= \prod_{k=1}^Q P(B | [G_{\theta_G}(I^n)]_{\lfloor i/2^{k-1} \rfloor, \lfloor j/2^{k-1} \rfloor}^{s_k}) P([G_{\theta_G}(I^n)]_{\lfloor i/2^{k-1} \rfloor, \lfloor j/2^{k-1} \rfloor}^{s_k} | I_{i,j}^n) \\ &\propto \prod_{k=1}^Q P(B | [G_{\theta_G}(I^n)]_{\lfloor i/2^{k-1} \rfloor, \lfloor j/2^{k-1} \rfloor}^{s_k}) \end{aligned} \quad (11)$$

TABLE II  
QUANTITATIVE COMPARISON OF FOREGROUND DETECTION ACCURACY BY DIFFERENT METHODS

| Methods                   |                           | Accuracy(%)                          |                                      |                                      |
|---------------------------|---------------------------|--------------------------------------|--------------------------------------|--------------------------------------|
|                           |                           | Tree                                 | Lake1                                | Lake2                                |
| Typical methods           | GMG [35]                  | 13.07% $\pm$ 2.97%                   | 21.45% $\pm$ 2.03%                   | 10.70% $\pm$ 1.30%                   |
|                           | GMM [4]                   | 5.85% $\pm$ 0.47%                    | 6.90% $\pm$ 0.16%                    | 7.19% $\pm$ 0.25%                    |
|                           | ASOM [32]                 | 10.01% $\pm$ 1.93%                   | 16.41% $\pm$ 1.49%                   | 18.40% $\pm$ 0.80%                   |
|                           | FASOM [31]                | 4.42% $\pm$ 0.53%                    | 13.55% $\pm$ 1.61%                   | 10.16% $\pm$ 0.27%                   |
|                           | IMBGS [34]                | 12.13% $\pm$ 2.69%                   | 10.98% $\pm$ 3.74%                   | 10.46% $\pm$ 2.16%                   |
|                           | MRF-UV [36]               | 16.43% $\pm$ 4.71%                   | 17.36% $\pm$ 2.32%                   | 8.82% $\pm$ 0.60%                    |
|                           | LOBSTER [30]              | 26.89% $\pm$ 2.60%                   | 37.87% $\pm$ 3.89%                   | 22.23% $\pm$ 3.64%                   |
|                           | SUBSENSE [33]             | 28.43% $\pm$ 7.93%                   | 25.89% $\pm$ 7.60%                   | 26.73% $\pm$ 3.42%                   |
| Enhancement-based methods | HE [9]+GMM [4]            | 13.08% $\pm$ 1.45%                   | 2.02% $\pm$ 0.02%                    | 2.93% $\pm$ 0.12%                    |
|                           | HE [9]+SUBSENSE [33]      | 33.95% $\pm$ 8.09%                   | 22.89% $\pm$ 2.66%                   | 22.33% $\pm$ 2.88%                   |
|                           | MSR [8]+GMM [4]           | 6.53% $\pm$ 0.33%                    | 2.21% $\pm$ 0.03%                    | 2.73% $\pm$ 0.07%                    |
|                           | MSR [8]+SUBSENSE [33]     | 29.05% $\pm$ 7.25%                   | 19.99% $\pm$ 4.29%                   | 17.24% $\pm$ 2.02%                   |
|                           | LIME [44]+GMM [4]         | 8.91% $\pm$ 1.45%                    | 6.18% $\pm$ 0.17%                    | 5.69% $\pm$ 0.35%                    |
|                           | LIME [44]+SUBSENSE [33]   | 17.73% $\pm$ 6.59%                   | 38.46% $\pm$ 7.37%                   | 30.74% $\pm$ 4.60%                   |
|                           | LLNet [45]+GMM [4]        | 6.11% $\pm$ 0.72%                    | 12.74% $\pm$ 0.37%                   | 8.87% $\pm$ 0.39%                    |
|                           | LLNet [45]+SUBSENSE [33]  | 19.03% $\pm$ 7.60%                   | 37.17% $\pm$ 3.06%                   | 25.51% $\pm$ 2.90%                   |
|                           | Struct [46]+GMM [4]       | 2.85% $\pm$ 0.20%                    | 6.44% $\pm$ 0.14%                    | 4.81% $\pm$ 0.14%                    |
|                           | Struct [46]+SUBSENSE [33] | 16.00% $\pm$ 6.33%                   | 39.88% $\pm$ 2.60%                   | 27.14% $\pm$ 3.37%                   |
| Ours                      | N2DGAN(global)            | 20.61% $\pm$ 2.72%                   | 35.74% $\pm$ 1.82%                   | 22.91% $\pm$ 1.52%                   |
|                           | N2DGAN(local)             | 36.46% $\pm$ 1.82%                   | 37.61% $\pm$ 0.93%                   | 27.68% $\pm$ 1.03%                   |
|                           | N2DGAN                    | <b>48.45% <math>\pm</math> 1.93%</b> | <b>40.32% <math>\pm</math> 1.69%</b> | <b>39.29% <math>\pm</math> 1.91%</b> |

nights, and the 'Lab' dataset is taken indoors where we control the intensity of lighting by pulling curtains and switching incandescent lights on purpose. Actually, these datasets are much challenging for background modeling task since they feature the undulation of lake, reflection of lights in the water, leaves shaking, and illumination variation. In order to make a quantitative evaluation, the foreground object(s) in the datasets are also manually labeled. Following the previous works on foreground detection, **IoU** is employed as our evaluation metric [39].

2) *Implementation Details*: The training of the proposed N2DGAN model is implemented on 2 NVIDIA TITAN Xp GPUs. The first 300 frames of each nighttime video in Tab.I paired with the corresponding daytime reference image are used to train the model, and the remaining are for testing. All of the images are downsampled to resolution of  $256 \times 256$ . Specially, we split each image into multiple image blocks with size  $32 \times 32$ , and then each block is used as the input of each local generator subnet. Considering the computational efficiency, only two scales are adopted, i.e.,  $Q = 2$ , to eliminate the influence of noise and spatial shift of background pixels. Based on RMSProp, the mini-batch gradient descent method is used with a batch size of 4 and a learning rate of  $10^{-4}$ . Since WGAN [27] is used as the backbone of our generation model, the weight need to be clamped to a fixed box  $(-0.01, 0.01)$  after each gradient update to avoid gradient vanishing and mode collapse problems during the learning process. In all our experiments, we uniformly set the total training epoch number to 30, and empirically set  $\lambda_{adv} = 10^{-1}$ ,  $\lambda_p = 10^{-5}$ ,  $\lambda_{mse} = 10^{-1}$ , and  $\lambda_{tv} = 10^{-3}$  in Eq.(7) to maintain the same order of magnitude. Besides, the first 3 convolutional layers of VGG network is used to calculate perceptual loss.

## B. Performance Evaluation

1) *Comparisons With State-of-the-Arts*: We compare the proposed method with state-of-the-arts on foreground

detection, including eight typical background modeling methods and four enhancement-based methods. Implementations of all these methods are based on the BGSLibrary [40] with default parameters. The quantitative comparison results on 3 sequences are shown in Tab.II. Obviously, our method can always achieve the best performance. Specifically, for sequences 'Tree' and 'Lake2', our method even outperforms the state-of-the-art method SUBSENSE [33] by 75% and 44%, respectively. Fig.4 presents the qualitative comparison results, which illustratively show that the proposed N2DGAN performs better. This mainly lies in two facts: 1) compared with the directly background modeling methods, generating daytime images makes the flatten pixel distribution sharper and easier to detect foreground; 2) compared with the enhancement-based methods, the unique daytime reference frame in generative process ensures the inter-frame consistency of the generated daytime images.

2) *Stability Comparison*: Consistency stability of successive frames is of great importance for further background modeling. For the  $t$ -th frame  $I_t \in R^{W \times H}$ , we use the following metric to measure the stability between  $I_t$  and its adjacent frame  $I_{t+1}$ .

$$s_t = Sim(I_t, I_{t+1}) \quad (12)$$

where  $Sim(\cdot, \cdot)$  denotes the distance between  $I_t$  and  $I_{t+1}$ . Here, we adopt Kullback-Leibler Divergence, which represents stronger stability if its value is close to 0. Fig.5 shows the stability comparison of several representative methods on a randomly selected sequence consisting of 200 consecutive frames from the test set of three datasets. As we can see, the result of our method (yellow line) is quite close to real nighttime images (blue line), while both HE (red line) and MSR (green line) show distinct difference from real sequence. Particularly, the large fluctuation by MSR also indicates that the pixel values between two adjacent frames differ greatly, which is unfavorable for background modeling. To sum up, with the joint constrain of spatial and temporal

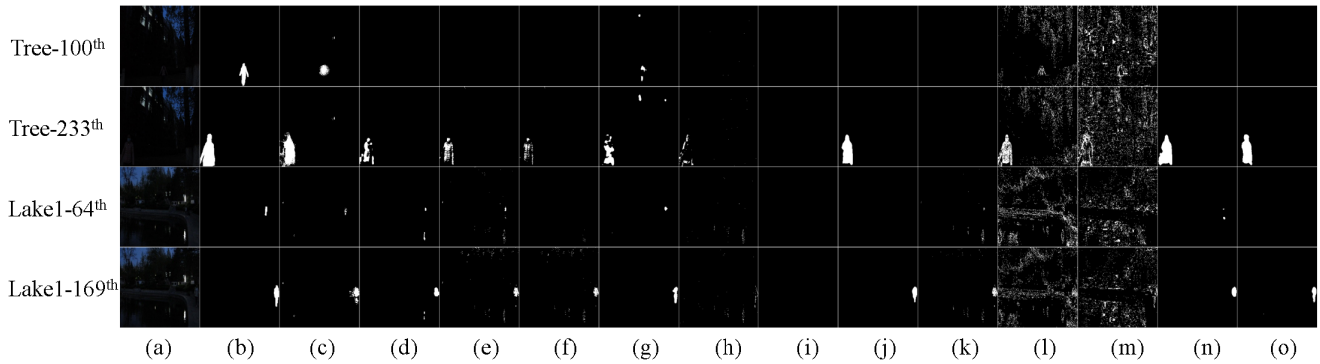


Fig. 4. Qualitative comparison of different foreground detection methods. (a) nighttime image, (b) groundtruth, (c) our method N2G-GAN, (d) GMG [35], (e) ASOM [32], (f) FASOM [31], (g) LOBSTER [30], (h) GMM [4], (i) MCueBGS [41], (j) SUBSENSE [33], (k) MRF-UV [36], (l) HE [9]+GMM [4], (m) MSR [8]+GMM [4], (n) HE [9]+SUBSENSE [33], (o) MSR [8]+SUBSENSE [33].

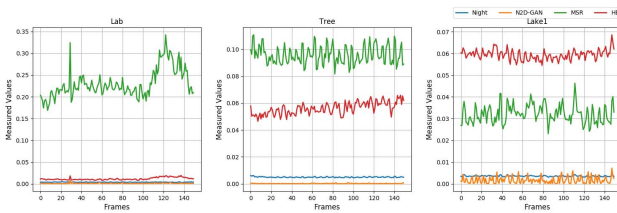


Fig. 5. Sequence stability comparison with HE, MSR and N2D-GAN on Lab (left), Tree (middle), and Lake1 (right) datasets.

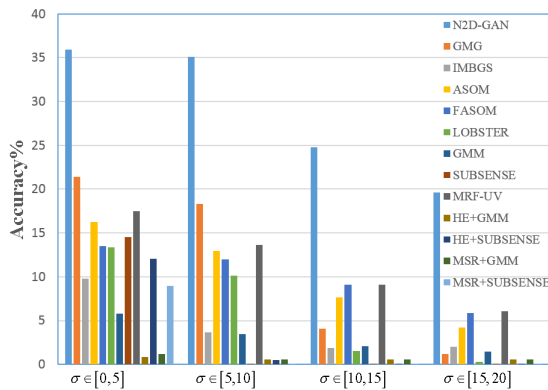


Fig. 6. Foreground detection accuracy on Lake1 dataset with different levels of noises.

consistency, our generation based method performs better than enhancement-based method.

### C. Robustness to Noise and Illumination Variation

Some extreme weather conditions such as rain, snow, and fog usually bring great challenges to background modeling. To demonstrate N2D-GAN's scalability under such environment, additive Gaussian noise is added to nighttime video sequence to simulate extreme weather. We randomize the noise standard deviation  $\sigma \in \{[0, 5], [5, 10], [10, 15], [15, 20]\}$  separately for each testing example. As illustrated in Fig.6, the behaviors are quantitatively different in all three datasets. This demonstrates that our method is the only technique that manages to perform well with different levels of noises. On two

randomly selected frames Lake1-13th and Lake1-158th, Fig. 7 and Fig. 8 present visually the comparison results with conventional background modeling methods and enhancement based methods, which clearly shows that the proposed N2D-GAN achieves the best performance.

For the experiments on evaluating the robustness to illumination variation, Fig. 9 illustrates the comparison results on the indoor nighttime dataset Lab with illumination variation. As we can see from Fig.9, the N2D-GAN model is much more insensitive to the instantaneous changes in light compared with other state-of-the-art background modeling methods. Here, the enhancement result based on HE (Fig. 9(b)) is only utilized to clarify the foreground object since it is not easy to find the ground truth in dark.

Two factors must be credited for our high resilience to noise and illumination change. The first originates from our model design, which allows noisy pixel to be outfitting to the reference background image. The second lies in our background model on successive multi-scale images, which is more robust to noise by hierarchical Bayes modeling.

### D. Ablation Study

1) *Global and Local Consistency Evaluation*: To verify the effectiveness of combining both local and global consistency together in our model, we first perform foreground detection when using global subnetwork alone, called  $N2D-GAN(global)$ . As observed from Fig.10, small targets in nighttime images are lost in this case. Meanwhile, when using local sub-net alone, called  $N2D-GAN(local)$ , the detected foreground objects are incomplete on the edge of patches, since there exists blocking-artifact problem caused by patch enhancement. Additionally, quantitative comparison results shown in Tab.II (bottom) demonstrate that our baseline improves detection accuracy by more than 10% and 5% compared with  $N2D-GAN(global)$  and  $N2D-GAN(local)$ , respectively.

*Evaluation on Multi-scale Bayes modeling* To further demonstrate the effectiveness of our background modeling on successive multi-scale images of daytime domain, we attempt to perform on a single scale generated images  $[G_{\theta_G}(I_n)]^{s_1}$ . As illustrated in Fig.11, due to the fact that multi-scale bayes model can suppress the noise caused by network, then

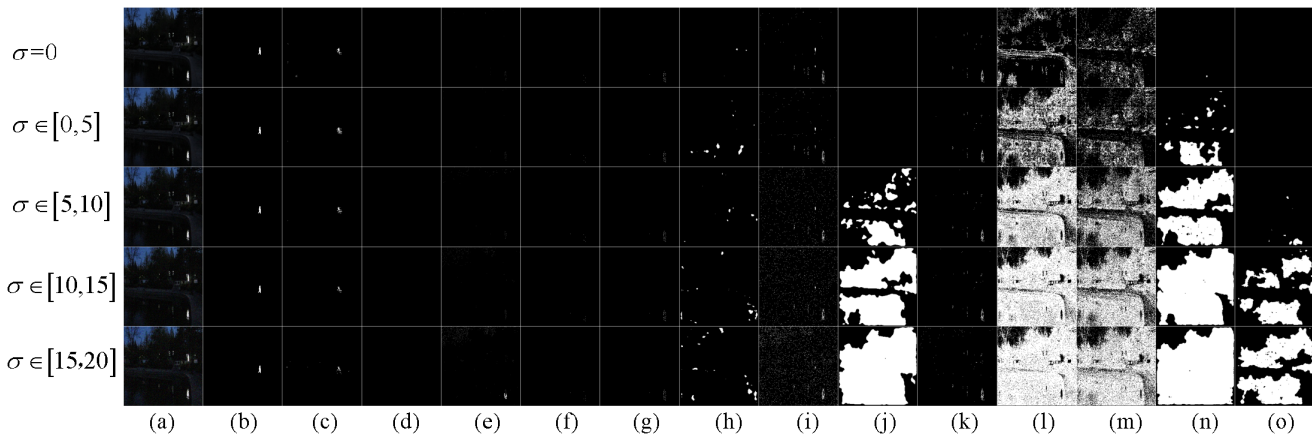


Fig. 7. Performance evaluation of robustness to noise on Lake1-13th. (a) The input nighttime image. (b) Groundtruth. (c) N2DGAN. (d) GMG [35], (e) IMBGS [34]. (f)ASOM [32], (g)FASOM [31], (h)LOBSTER [30], (i)GMM [4], (j)SUBSENSE [33], (k)MRF-UV [36], (l) HE [9]+GMM [4], (m) MSR [8]+GMM [4], (n) HE [9]+SUBSENSE [33], (o) MSR [8]+SUBSENSE [33].

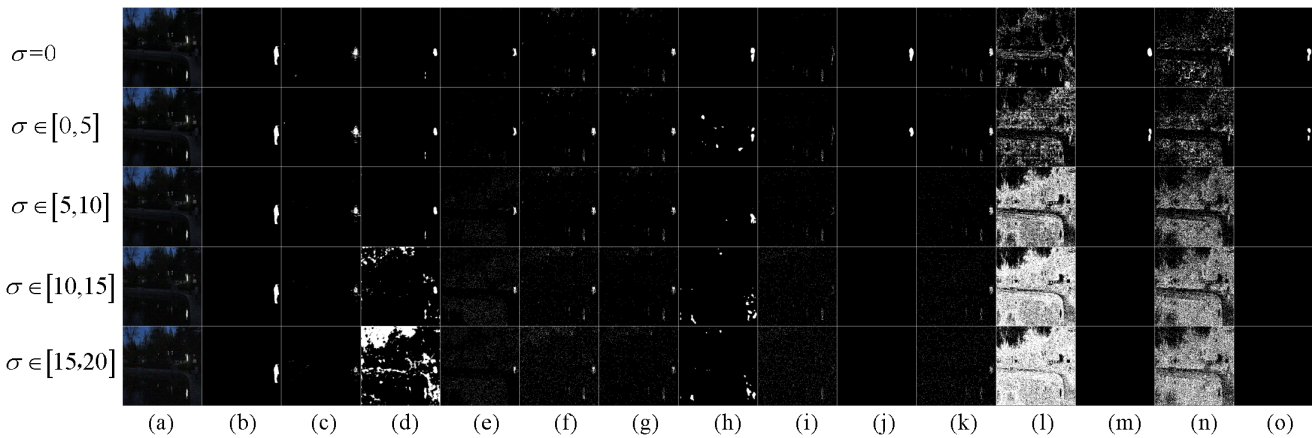


Fig. 8. Performance evaluation of robustness to noise on Lake1-158th. (a) The input nighttime image. (b) Groundtruth. (c) N2DGAN. (d) GMG [35], (e) IMBGS [34]. (f)ASOM [32], (g)FASOM [31], (h)LOBSTER [30], (i)GMM [4], (j)SUBSENSE [33], (k)MRF-UV [36], (l) HE [9]+GMM [4], (m) MSR [8]+GMM [4], (n) HE [9]+SUBSENSE [33], (o) MSR [8]+SUBSENSE [33].

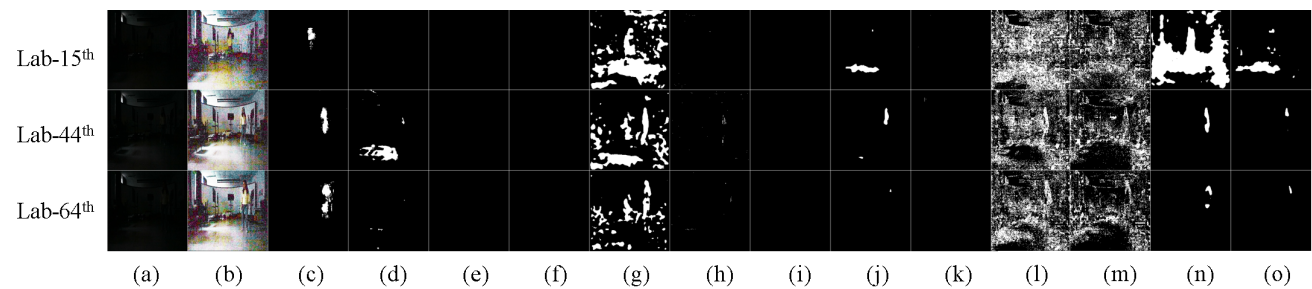


Fig. 9. Performance evaluation of robustness to illumination variation. (a) nighttime image, (b) HE, (c) N2DGAN, (d) GMG [35], (e) ASOM [32], (f) FASOM [31], (g) LOBSTER [30], (h) GMM [4], (i) MCueBGS [41], (j) SUBSENSE [33], (k) MRF-UV [36], (l) HE [9]+GMM [4], (m) MSR [8]+GMM [4], (n) HE [9]+SUBSENSE [33], (o) MSR [8]+SUBSENSE [33].

our baseline N2DGAN makes the foreground region more remarkable.

*E. Time Complexity Analysis*

For a background model, the computational complexity is one of the key issues worthy of attention. For our N2DGAN model, its computational complexity mainly consists of two

parts, i.e., virtual daytime image generation and foreground detection. In the foreground detection stage, since our model holds only a single gaussian model which can be off-line available and without need for online model updating, thus the time complexity in this stage is much lower than the traditional GMM and can be negligible compared with the one in generation stage.



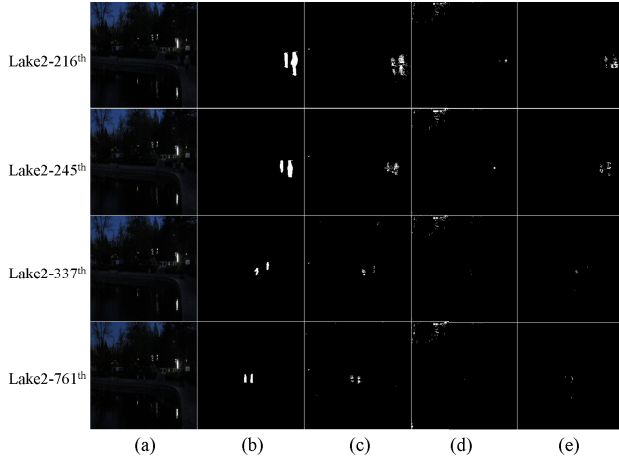


Fig. 10. Comparison of foreground detection results using global subnet and local subnet independently. (a) Input nighttime image. (b) Ground truth. (c) N2DGAN. (d) N2DGAN(global). (e) N2DGAN(local).

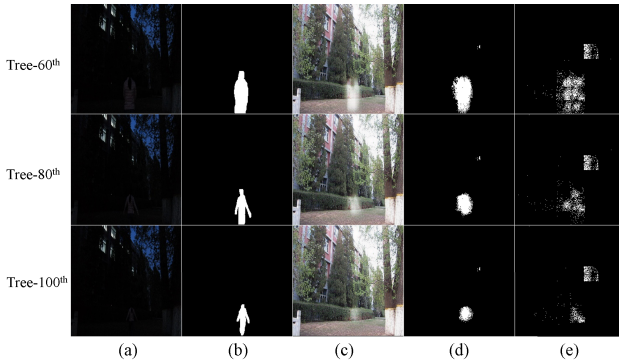


Fig. 11. Comparison of foreground detection results using multi scales and single scale. (a) nighttime image. (b) Groundtruth. (c) Generated image  $[G_{\theta_G}(I_n)]^{s_1}$ . (d) N2DGAN. (e) The detection result using single scale  $s_1$ .

TABLE III  
TIME COMPLEXITY ANALYSIS OF GENERATION PROCESS

| Time Complexity (fps) |                |               |
|-----------------------|----------------|---------------|
| N2DGAN                | N2DGAN(global) | N2DGAN(local) |
| 8                     | 56             | 10            |

By comparison, to generate a virtual daytime image will occupy most of the time with a frame rate of 8 fps without any code optimisation. As shown in Table III, the frame rate of **N2DGAN(local)** to generate  $8 \times 8$  blocks of local sub-images<sup>2</sup> is around 10 fps, which is much more slowly than **N2DGAN(global)** with a frame rate of 56 fps. However, considering that we can generate each block of local virtual sub-image in parallel, the frame rate of **N2DGAN(local)** will dramatically increased, approximating around 640 fps in an ideal situation. It means that the global generation process with 56 fps will dominate the overall computational complexity of

<sup>2</sup>For an input nighttime image, it is divided into  $8 \times 8$  blocks and then the corresponding local generation sub-network is trained on each block to generate a local sub-image. For details please refer to Section III.

TABLE IV  
ARCHITECTURE OF THE LOCAL GENERATOR SUB-NETWORK

| Layer  | Inputs                     | Kernel/Stride             | Outputs                         |
|--------|----------------------------|---------------------------|---------------------------------|
| cov0   | $B_1^n$                    | $3 \times 3 \times 32/1$  | $32 \times 32 \times 32$        |
| cov1   | cov0                       | $3 \times 3 \times 64/2$  | $16 \times 16 \times 64$        |
| cov2   | cov1                       | $3 \times 3 \times 128/2$ | $8 \times 8 \times 128$         |
| decov0 | cov2                       | $3 \times 3 \times 64/2$  | $16 \times 16 \times 64$        |
| decov1 | cov0                       | $3 \times 3 \times 32/2$  | $32 \times 32 \times 32$        |
| decov2 | cov2                       | $3 \times 3 \times 128/2$ | $16 \times 16 \times 128$       |
| decov3 | decov0,<br>cov1,<br>decov2 | $3 \times 3 \times 256/2$ | $32 \times 32 \times 256$       |
| decov4 | decov3,<br>cov0,<br>decov1 | $3 \times 3 \times 64/2$  | $F_1^d(32 \times 32 \times 64)$ |

TABLE V  
ARCHITECTURE OF THE GLOBAL GENERATOR SUB-NETWORK

| Layer  | Inputs                     | Kernel/Stride              | Outputs                    |
|--------|----------------------------|----------------------------|----------------------------|
| cov0   | $I_t^n$                    | $3 \times 3 \times 32/1$   | $256 \times 256 \times 32$ |
| cov1   | cov0                       | $3 \times 3 \times 64/2$   | $128 \times 128 \times 64$ |
| cov2   | cov1                       | $3 \times 3 \times 128/2$  | $64 \times 64 \times 128$  |
| cov3   | cov2                       | $3 \times 3 \times 256/2$  | $32 \times 32 \times 256$  |
| cov4   | cov3                       | $3 \times 3 \times 512/2$  | $16 \times 16 \times 512$  |
| cov5   | cov4                       | $3 \times 3 \times 1024/2$ | $8 \times 8 \times 1024$   |
| fc1    | cov5                       | -                          | 512                        |
| fc2    | fc1                        | -                          | $8 \times 8 \times 128$    |
| decov0 | fc2                        | $3 \times 3 \times 64/2$   | $16 \times 16 \times 64$   |
| decov1 | decov0                     | $3 \times 3 \times 32/2$   | $32 \times 32 \times 32$   |
| decov2 | decov1                     | $3 \times 3 \times 16/2$   | $64 \times 64 \times 16$   |
| decov3 | decov2                     | $3 \times 3 \times 8/2$    | $128 \times 128 \times 8$  |
| decov4 | decov3                     | $3 \times 3 \times 4/2$    | $256 \times 256 \times 4$  |
| decov5 | cov5                       | $3 \times 3 \times 512/2$  | $16 \times 16 \times 512$  |
| decov6 | decov5,<br>cov4,<br>decov0 | $3 \times 3 \times 256/2$  | $32 \times 32 \times 256$  |
| decov7 | decov6,<br>cov3,<br>decov1 | $3 \times 3 \times 128/2$  | $64 \times 64 \times 128$  |
| decov8 | decov7,<br>cov2,<br>decov2 | $3 \times 3 \times 64/2$   | $128 \times 128 \times 64$ |
| decov9 | decov8,<br>cov1,<br>decov3 | $3 \times 3 \times 32/2$   | $256 \times 256 \times 32$ |
| cov6   | decov9,<br>cov0,<br>decov4 | $3 \times 3 \times 32/1$   | $256 \times 256 \times 32$ |

the virtual daytime image generation. By this way, the need for online real-time foreground object detection can be met.

## VI. CONCLUSION

For the challenge of background modeling under daytime scene, an innovative N2DGAN model is proposed, which paves a new way completely different from the existing methods. To the best of our knowledge, this is the first time to introduce GANs based deep learning for this practical problem. As an unsupervised model, N2DGAN is provided with good scalability and practical significance.

As for the time complexity of N2DGAN, it takes about 0.125 seconds (8 fps) for each frame. Considering each local generation model can be implemented in parallel, the proposed N2DGAN could be highly parameterizable. Besides, some model compression works [42], [43] can also be feasible

solutions for network acceleration. It is also worth noting that we have assumed that the images in the training set is free of foreground objects. To overcome this limitation, we will try to extend this method to deal with the situation that the images in the training set may contain some foreground objects.

## APPENDIX

The detailed structures of the global sub-network and local sub-network are provided in Table IV and Table V, respectively. Each convolution layer is followed by 3 residual block [28]. Then we simply concatenate the output from each local generator and the global generator to produce a fused feature tensor and then feed it to the successive convolution layers to generate the final output.

## REFERENCES

- [1] 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.
- [2] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.
- [3] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imag.*, vol. 11, no. 3, pp. 172–185, 2005.
- [4] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, 2004, pp. 28–31.
- [5] L. A. Lim and H. Yalim Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognit. Lett.*, vol. 112, pp. 256–262, Sep. 2018.
- [6] L. Yang, J. Li, Y. Luo, Y. Zhao, H. Cheng, and J. Li, "Deep background modeling using fully convolutional network," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 254–262, Jan. 2018.
- [7] D. Zeng and M. Zhu, "Background subtraction using multiscale fully convolutional network," *IEEE Access*, vol. 6, pp. 16010–16021, 2018.
- [8] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.
- [9] D. J. Ketcham, "Real-time image enhancement techniques," *OSA Image Process.*, vol. 74, no. 2, pp. 120–125, 1976.
- [10] Y. Gong, Y. Lee, and T. Q. Nguyen, "Nighttime image enhancement applying dark channel prior to raw data from camera," in *Proc. Int. Soc. Design Conf. (ISOCC)*, Oct. 2016, pp. 173–174.
- [11] Y. Cai, K. Huang, T. Tan, and Y. Wang, "Context enhancement of nighttime surveillance by image fusion," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 980–983.
- [12] Y. Rao, "Image-based fusion for video enhancement of night-time surveillance," *Opt. Eng.*, vol. 49, no. 12, Dec. 2010, Art. no. 120501.
- [13] R. Raskar, A. Ilie, and J. Yu, "Image fusion for context enhancement and video surrealism," in *Proc. Int. Symp. Non-Photorealistic Animation Rendering*, 2004, p. 4.
- [14] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [15] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.
- [16] R. Yeh *et al.*, "Semantic image inpainting with perceptual and contextual losses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1–10.
- [17] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–8.
- [18] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
- [19] S. Zheng, Z. Zhu, J. Cheng, Y. Guo, and Y. Zhao, "Edge heuristic GAN for non-uniform blind deblurring," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1546–1550, Oct. 2019.
- [20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
- [21] Y. Meng, D. Kong, Z. Zhu, and Y. Zhao, "From night to day: GANs based low quality image enhancement," *Neural Process. Lett.*, vol. 50, no. 1, pp. 799–814, Aug. 2019.
- [22] P. Kaewtrakulpong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*. Berlin, Germany: Springer, 2002, pp. 135–144.
- [23] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1999, pp. 246–252.
- [24] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 751–767.
- [25] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657–662, Apr. 2006.
- [26] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1301–1306.
- [27] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1–11.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 770–778.
- [29] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2016, pp. 2414–2423.
- [30] 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.
- [31] 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, Mar. 2010.
- [32] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, Jul. 2008.
- [33] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Flexible background subtraction with self-balanced local sensitivity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 408–413.
- [34] D. Bloisi and L. Iocchi, "Independent multimodal background subtraction," in *Proc. Comput. Modeling Objects Represented Images Fundam. Methods Appl.*, 2012, pp. 39–44.
- [35] A. B. Godbehere, A. Matsukawa, and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *Proc. Amer. Control Conf.*, 2012, pp. 4305–4312.
- [36] Z. Zhao, T. Bouwmans, X. Zhang, and Y. Fang, "A fuzzy background modeling approach for motion detection in dynamic backgrounds," in *Proc. Int. Conf. Multimedia Signal Process.*, 2012, pp. 177–185.
- [37] Z. Zhu, H. Lu, and Y. Zhao, "Scale multiplication in odd Gabor transform domain for edge detection," *J. Vis. Commun. Image Represent.*, vol. 18, no. 1, pp. 68–80, Feb. 2007.
- [38] L. Zhang and P. Bao, "Edge detection by scale multiplication in wavelet domain," *Pattern Recognit. Lett.*, vol. 23, no. 14, pp. 1771–1784, Dec. 2002.
- [39] Z. Liu, K. Huang, and T. Tan, "Foreground object detection using top-down information based on EM framework," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4204–4217, Sep. 2012.
- [40] A. Sobral, "BGSLibrary: An OpenCV C++ background subtraction Library," in *Proc. IX Workshop Vis. Comput. (WVC)*, Rio de Janeiro, Brazil, 2013, pp. 1–8.
- [41] S. J. Noh and M. Jeon, "A new framework for background subtraction using multiple cues," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 493–506.
- [42] K. Jia, D. Tao, S. Gao, and X. Xu, "Improving training of deep neural networks via singular value bounding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4344–4352.
- [43] X. Yu, T. Liu, X. Wang, and D. Tao, "On compressing deep models by low rank and sparse decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7370–7379.
- [44] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

- [45] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.
- [46] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.



**Zhenfeng Zhu** received the M.E. degree from the Harbin Institute of Technology, Harbin, China, in 2001, and the Ph.D. degree from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, CAS, Beijing, China, in 2005, respectively.

He was a Visiting Scholar with the Department of Computer Science and Engineering, Arizona State University, Tempe, AZ, USA, in 2010. He is currently a Professor with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing. He has authored or coauthored more than 100 articles in journals and conferences, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (T-KDE), the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (T-NNLS), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (T-CSVT), the IEEE TRANSACTIONS ON MULTIMEDIA (T-MM), the IEEE TRANSACTIONS ON CYBERNETICS (T-CYB), the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), AAAI, IJCAI, and ACM Multimedia. His group has won the Honorable Award (Rank First in stage two) in KDD Cup2016 competition and Rank3 in CIKM Cup2016 competition, respectively. His current research interests include image and video understanding, computer vision, machine learning, and recommendation systems.



**Yingying Meng** received the B.S. degree in computer science and technology from the Shandong University of Science and Technology, Qingdao, China, in 2016, and the M.E. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2019. She is currently a Software Engineer with the China CITIC Bank.



**Deqiang Kong** received the M.E. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2018. He is currently an Applied Scientist with Microsoft, Beijing. His research interests include video understanding, natural language processing, and recommendation systems.



**Xingxing Zhang** received the B.S. degree in communication engineering from Henan Normal University, Xinxiang, China, in 2015. She is currently pursuing the Ph.D. degree with the Institute of Information Science, Beijing Jiaotong University, Beijing, China.

Her research interests include data analysis, image and video understanding, computer vision, and machine learning.



**Yandong Guo** received the B.S. and M.S. degrees in ECE from the Beijing University of Posts and Telecommunications, China, in 2005 and 2008, respectively, and the Ph.D. degree in ECE from Purdue University at West Lafayette in 2013, under the supervision of Prof. Bouman and Prof. Allebach. He is currently the Chief Scientist of intelligent perception with the OPPO Research Institute. He also holds an Adjunct Professor position at the Beijing University of Posts and Telecommunications, and the University of Electronic Science and Technology of

China. Before he joined OPPO in 2020, he was the Chief Scientist with XPeng Motors, China, and a Researcher with the Microsoft Research, Redmond, WA, USA. His professional interests lie in the broad area of computer vision, imaging systems, human behavior understanding and biometric, and autonomous driving.



**Yao Zhao** (Senior Member, IEEE) received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996.

From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Delft University of Technology, Delft, The Netherlands. Since 2001, he has been a Professor with BJTU, where he is currently the Director of the Institute of Information Science. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding.

Dr. Zhao serves on the editorial boards of several international journals, including as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE SIGNAL PROCESSING LETTERS, and *Circuits, Systems, and Signal Processing*, and an Area Editor of *Signal Processing: Image Communication*. He was named as a Distinguished Young Scholar by the National Science Foundation of China in 2010 and was elected as a Chang Jiang Scholar of the Ministry of Education of China in 2013.