GAN-Based Synthetic Data Augmentation for Infrared Small Target Detection

Jun-Hyung Kim, *Member, IEEE*, and Youngbae Hwang, *Member, IEEE*

***Abstract*— Recently, convolutional neural networks (CNNs) have achieved state-of-the-art performance in infrared small target detection. However, the limited number of public training data restricts the performance improvement of CNN-based methods. To handle the scarcity of training data, we propose a method that can generate synthetic training data for infrared small target detection. We adopt the generative adversarial network framework where synthetic background images and infrared small targets are generated in two independent processes. In the first stage, we synthesize infrared images by transforming visible images into infrared ones. In the second stage, target masks are implanted on the transformed images. Then, the proposed intensity modulation network synthesizes realistic target objects that can be diversely generated from further image processing. Experimental results on the recent public dataset show that, when we train various detection networks using the dataset composed of both real and synthetic images, detection networks yield better performance than using real data only.**

***Index Terms*— Convolutional neural network (CNN), generative adversarial network (GAN), image-to-image translation, infrared small target, synthetic data augmentation.**

I. Introduction

DETECTING small targets in infrared images is an important task in infrared search and track (IRST) systems. It has many applications in the military field, such as infrared homing guidance, early warning, antiaircraft, and antimissile [1]. The main challenge of infrared small target detection lies in the lack of sufficient structural features because the target is usually represented by a tiny number of pixels in the image [2]. In addition, background clutters may cause severe false alarms to make target discrimination more difficult.

Traditional filtering approaches tackle the problem by manually designed filters that produce high responses in the target region because the appearance of targets generally resembles small bright spots in infrared images. Top-hat [3] and Laplacian of Gaussian [4] filters are applied to leverage this spatial characteristic of the target. On the other hand, a 2-D least-squares filter [5] estimates the background signal of the current region by using surrounding pixels. Then, the difference between the input and the predicted backgrounds is used to detect the targets. However, such simple models for the target [3], [4] and background [5] are not only difficult to adapt to various scenarios but can also generate many false alarms. Inspired by a human visual system, a local contrast measure (LCM) [6] is proposed to enhance the target signal and suppress the background one. The LCM is extended to improve the computational efficiency and performance [7], [8]. To exploit structural information of the target and backgrounds efficiently, a detection method using multiple directional filters is also proposed [9]. Recently, a novel morphological representation [10], [11], such as trees of different shapes, has been introduced to compensate for the lack of detailed information in infrared images by supplementing spatial information [12].

The infrared patch image (IPI) model [13] and its variants [14], [15] provide another alternative scheme to the traditional filtering methods. The IPI model formulates the infrared small target detection as a low-rank (background) and sparse signal (target) decomposition problem (for a thorough review of filtering and decomposition-based methods; see [12]). Although these approaches show promising results in some images, they still have difficulty in discriminating the small and dim target from the complex and nonrepetitive background region due to their inability to capture high-level semantics. Rapid progress in deep convolutional neural networks (CNNs) motivated recent works [1], [16]–[19] to formulate the problem of learning an end-to-end model that converts an input infrared image into the detection map. These learning-based methods are shown to produce better detection results than the traditional filtering or low-rank approximation ones.

In general, the performance of CNN-based approaches is enhanced as the number of training samples increases. Unfortunately, the size of existing public datasets [17], [18] for infrared small target detection is very limited compared to that for general object detection and classification. Collecting and manually annotating large datasets for infrared small target detection require considerable time, effort, and cost due to the following reasons.

1) The IRST system is mounted on platforms, such as the aircraft [20] or the ship [21], and then, infrared images are acquired during the operation. To construct datasets from the stored images, first, images containing targets of interest should be selected from the large set of images. Second, the raw data should be preprocessed to

Manuscript received April 26, 2022; revised May 24, 2022; accepted May 30, 2022. Date of publication June 2, 2022; date of current version June 14, 2022. This work was supported in part by the Grand Information Technology Research Support Program under Grant IITP-2022-2020-0-01462, in part by the National Research Foundation of Korea (NRF) under Grant 2020R1F1A1077110, and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under Grant 2021-0-02068 and Grant 2020-0-01077. *(Corresponding author: Youngbae Hwang.)*

The authors are with the Department of Intelligent Systems and Robotics, Chungbuk National University, Cheongju 28644, Republic of Korea (e-mail: ybhwang@cbnu.ac.kr).

Digital Object Identifier 10.1109/TGRS.2022.3179891

serve as an input to the CNN model (e.g., removing bad pixels as it can be confused with the real target). Finally, experts who can discriminate the target from background clutters should generate the ground-truth label. These processes take lots of time and effort.

2) A diverse training data should be prepared to improve the generalizability of the trained model. For that purpose, the platform equipped with the IRST system should collect a large number of images in various environments, which leads to increased costs. For example, the appearance of obtained images can be severely different even for the same scene since the at-aperture radiance depends not only on the background surface but also on the atmospheric condition.

Training with synthetic data is an efficient way because the burden of constructing the real dataset can be significantly mitigated. A synthetic image generator simulates the physical process of image formation and creates artificial images using the modeled imaging pipeline. By adjusting parameters governing the generation process, one can obtain a large number of training images with numerous variations. Moreover, the synthetic image generator can automatically produce annotations during the process. Synthetic datasets have been widely used to train CNN models in various applications, such as automatic target recognition in synthetic aperture radar data [22], target detection in hyperspectral images [23], and airplane detection in aerial images [24].

Another approach to synthesizing training data is to utilize the generative adversarial network (GAN) [25]. According to the source of generated samples, GAN-based data synthesis can be classified into two categories: sample synthesis and sample migration [26]. In sample synthesis, the generator tries to output fake but plausible data, while the discriminator is trained to discriminate between the real and fake data. After the training, the generator produces fake data that fit the distribution of real data as much as possible. The methods in the second category, also known as the image-to-image translation for image data [27], learn to convert the data from the source domain to the target by not only preserving the intrinsic source content but also transferring the extrinsic target style. The advantage of GAN-based methods over simulation-based data generation is that they can obtain synthetic data without modeling either the imaging pipeline or the physical world.

In this work, we introduce a new method to create synthetic training data for infrared small target detection. Here, we approach this problem in the context of image-to-image translation. Unlike previous methods that simultaneously handle the background and object instances in the single translation process, we take a two-stage approach. This is motivated by the following reasons.

1) Usually, targets of interest are located at a distant range in infrared small target detection. In such a case, it is highly likely that observable targets in the infrared image are not discernible in the visible image due to the relatively low atmospheric transmission in the visible spectrum [28]. Therefore, the paired target data in two spectral domains are hard to obtain.

2) Contrary to visible objects [29], the appearance of small targets in the infrared domain has few prominent characteristics. In addition, the spatial distribution of targets is not well discriminated with backgrounds [1]. Consequently, when we address the problems of translating the background and targets from one domain to another separately, the translation process can be more tractable.

In the first stage of the proposed method, we transfer visible images to an infrared domain by applying the image-to-image translation method [30] to produce synthetic background images. Then, we generate target masks for those images. To synthesize realistic targets, we propose an intensity modulation network that is trained within a GAN framework. The intensity modulation network adjusts the intensity of the masks to be fit with the surroundings, such that it can fool the discriminator. On the other hand, the discriminator is trained to distinguish real targets and intensity-modulated target masks. It is worth noting that Uddin *et al.* [31], similar to our work, propose a framework that converts visible images to infrared ones. However, their conversion network is trained and tested using an image dataset acquired at the same scene. In addition, the dataset used in their experiments deals with extended targets. The main contributions of this work are given as follows.

1) We decompose the problem of generating synthetic training data into two subproblems.
   a) We leverage the image-to-image translation techniques to produce synthetic background images from real visible spectrum images.
   b) Based on the proposed intensity modulation network, a GAN framework is adopted to render realistic targets on the synthetic background images.
2) We compare five deep networks including the baseline and four state-of-the-art models to show the effectiveness of the proposed data augmentation method.
3) In our ablation study, we show that each stage of the proposed method can gradually contribute to the improvement of the detection performance.

The organization of this article is given as follows. In Section II, we briefly review the related work. In Section III, we introduce the proposed method for generating synthetic training data. In Section IV, we present our experimental results. The conclusion is given in Section V.

II. Related Work

A. Deep Learning-Based Infrared Small Target Detection

Contrary to the traditional filtering approaches, data-driven approaches, i.e., deep learning-based methods, do not assume any prior knowledge of the target and background, and learn it from the data. Wang *et al.* [17] adopt a conditional adversarial learning strategy [25] to train the target segmentation networks. In their work, there are two generators that produce segmentation results. One takes a role in minimizing the false alarms, and the other takes a role in reducing miss detection. Both generators are built with dilated convolutions [32] to gather context clues. A discriminator differentiates the segmentation result from the ground truth. Final results are obtained by averaging two segmentation outputs. Motivated by the success of the conditional adversarial networks for the image translation problem [27], Zhao *et al.* [1] propose a method to train a detection network that converts an infrared image into a target map. A generator built upon U-Net [33] and a discriminator are trained via adversarial learning from the paired training data. Since it is hard to acquire the real paired training data in practice, they propose a process to synthesize paired training data. Recently, based on the feature pyramid network (FPN) [34], Dai *et al.* [16] propose attentional local contrast (ALC) networks to segment a target region from an infrared image. In the ALC network, the multiscale LCM is performed at the feature level using the dilated convolution and the cyclic shift of the feature map. They also incorporate a new attention module to merge features at different layers. In another work of Dai *et al.* [18], an asymmetric contextual modulation (ACM) module is introduced and integrated into the U-Net and FPN. Recently, Li *et al.* [19] propose a dense nested detection network (DNA-Net) based on UNet++ [35] and the convolutional block attention module [36].

B. Image-to-Image Translation

An image-to-image translation is a task of converting one possible representation of a scene into another, given sufficient training data [27]. Since its introduction as a general-purpose solution, it has achieved great success and applied to many visual tasks [37]. In Pix2pix [27], the translation network is trained within the framework of conditional GAN [38] in which an input image is a conditioning variable. The network is guided by the L1 loss that forces the translated image close to the corresponding image in the target domain. Despite its great success, the training method of Pix2pix brings several problems. The first one is the mode collapse where the translation network fails to generate diverse outputs [27]. Another problem is the requirement of paired source–target images as it is rarely the case that one can obtain a sufficient number of paired training samples. To enable the translation network to generate multimodal outputs, BicycleGAN [30] introduces a bijective consistency between the latent code (i.e., an input noise vector to the translation network) and the output. In CycleGAN [39], the constraint of the paired training dataset is relaxed by a cycle consistency that translates an image from the source domain to the target domain and then back again to the source should be the identity function.

C. Modality Transfer

Existing data samples in one modality can be transferred into another one to supplement the insufficient training dataset. As there exist some large datasets available for visual tasks in the visible domain, we can utilize those datasets to remedy the scarcity of data in the infrared domain. In [40], labeled video sequences in the visible spectrum are translated into infrared video sequences. Together with the labels in the original domain, the translated videos are used to train a thermal infrared tracker. In that work, Pix2pix [27] and CycleGAN [39] are used for the translation. In the same manner as above, Pix2pixHD [41] is adopted to transform visible images annotated with semantic labels into thermal images for thermal image semantic segmentation [42].

Depending on the time and place of the data acquisition, the characteristic of remote sensing data can vary significantly. Therefore, the models trained with datasets collected during specific periods or for a particular region may show poor performance at the test time. Modality transfer can also be used to reduce the domain gap between the training and the testing datasets. In [43], a simple generator that consists of one scale and one shift matrices is trained to transform the spectral distribution of training images into that of test images. Then, the transformed training images are used to fine-tune the semantic segmentation network. In [44], a two-stage approach is used to increase the generalization capability of the segmentation network. First, a data augmentation network learns mappings among multiple domains using a style transfer technique [45]. Then, the segmentation network is trained from the diversified samples generated by the augmentation network.

When translating complex images that contain many disparate object instances, objects in the source image can be mapped incorrectly in the target domain (i.e., the data distribution of translated objects is inconsistent with that of objects in the target domain) [46], [47]. To solve the problem of image-level translation, several instance-level translation methods have been proposed [46]–[49]. These methods train their translation network so that the network applies adaptive mappings for the background and each object instance. In this work, we propose a new synthetic image generation method that unifies the translation of the image and objects of interest in a sequential manner, thus allowing us to leverage realistic synthetic images for infrared small target detection. The proposed data augmentation scheme can mitigate the limitation of an insufficient dataset in this area.

III. Synthetic Image Generation Process

Fig. 1 depicts the overall procedure of the proposed method. Visible images collected from the web are translated into the infrared domain. BicycleGAN [30] capable of generating multiple target images from one source image is adopted for the background translation in this work. To render targets in the synthetic background images, various target masks are built from the ground truth of real data. First, the position of the target mask is randomly shifted, and then, the shape of the target mask is modified arbitrarily by applying one of the operations among dilation, erosion, and identity. The processed mask is implanted in the synthetic background image. The proposed intensity modulation network takes the synthetic image together with the mask as input and modulates intensity values of the target area to transform the implanted mask into a realistic target. The discriminator is designed to distinguish between real and synthetic targets by focusing only on the target area. In Fig. 1, for simplicity, we omit the target mask used as an additional input to the discriminator. The adversarial loss forces the intensity modulation network to learn to map from the implanted target mask represented by a fixed pixel value to the synthetic target whose spatial

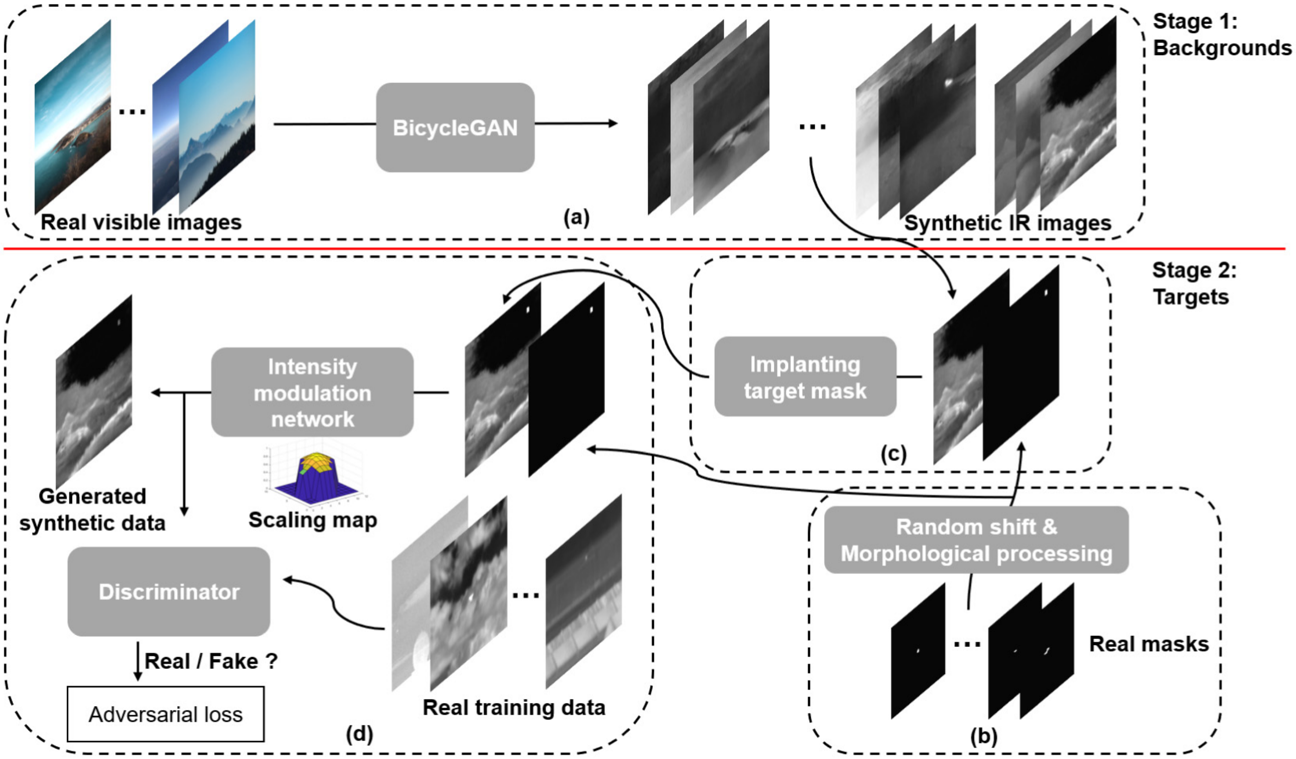

Fig. 1. Overview of the proposed synthetic image generation process. (a) BicycleGAN transfers visible images collected from the web into the infrared domain. (b) First, binary masks that represent the target area with one and the background area with zero are randomly shifted, and then, a morphological operation is performed on the shifted mask. (c) Using the binary mask, a target is implanted in the translated image with the highest pixel value. (d) Proposed intensity modulation network manipulates the pixel values of the target area with the aim to fool the discriminator. The discriminator differentiates between the real and synthetic targets.

distribution resembles that of the real target. The details of each stage will be introduced in Sections III-A and III-B.

A. Synthetic Background Images

As mentioned in Section I, the size of publicly available datasets for infrared small target detection is limited. To create various background images, we transform visible images into infrared images using the image-to-image translation method. In [40], two image-to-image translation methods, Pix2pix [27] and CycleGAN [39], are considered and compared to generate labeled synthetic infrared videos from labeled real color videos for the thermal infrared tracking problem. By inspecting the quality of translated images subjectively and objectively, they concluded that Pix2pix that uses the paired training dataset is superior to CycleGAN that uses the unpaired training dataset. Based on this result, we utilize BicycleGAN [30] that is an extension of Pix2pix with the capability of producing multimodal outputs. We require the translation network to have an additional property, multimodality, due to the following reasons.

1) Multimodal outputs mimic the diversity of infrared images. For the same scene, the intensity distribution of captured infrared images can vary depending on the time of the day and environmental conditions. Trained on the multimodal synthetic background images, the detection network for infrared small target detection becomes more robust to such a change.

TABLE I
COMPARISON OF SEVERAL IMAGE TRANSLATION METHODS

Method	Paired/Unpaired	Multi-modality
Pix2pix	Paired	No
CycleGAN	Unpaired	No
BicycleGAN	Paired	Yes

2) Given the same number of visible images, we can increase the number of synthetic background images generated by the image-to-image translation module.

In summary, we choose BicycleGAN for the background image translation as it is multimodal and produces better results than the method utilizing an unpaired training dataset. The three image translation methods introduced are compared in Table I.

From the web, we collect various images of the sky, mountain, cloud, seashore, and the horizon. Randomly generated latent vectors and the collected images are used as inputs to the trained BicycleGAN. Examples of translated infrared images are shown in Figs. 2 and 3. Unfortunately, not all translated images are consistent with the input visible image, as seen in Fig. 3. We commonly observe the structural mismatch between the visible and translated infrared images and distortions, especially in smooth regions. As can be seen in Fig. 4, the contours of the two images should be similar even though modalities differ. This problem may be mitigated

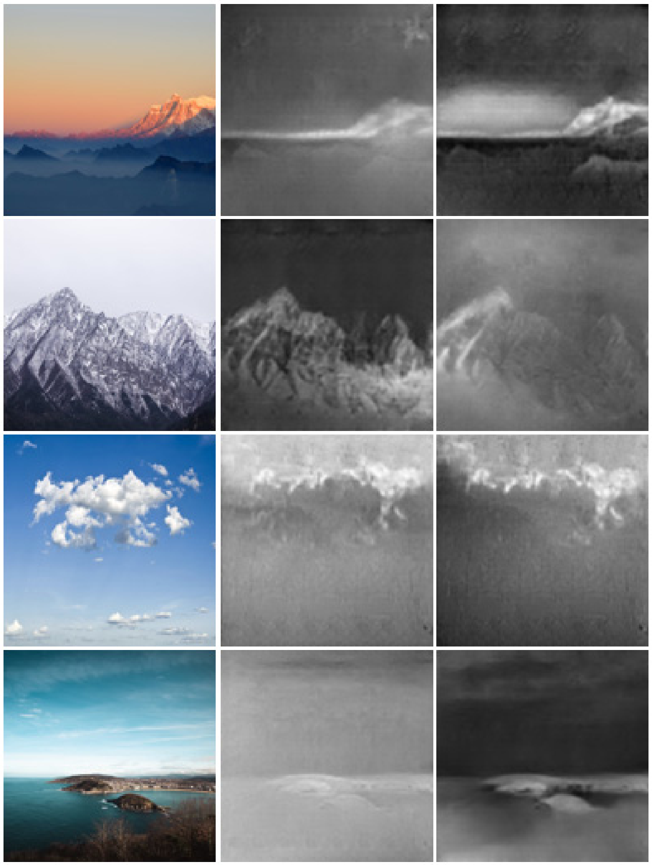

Fig. 2. Successful cases. Given visible images (leftmost), BicyclGAN produces multimodal outputs that have different visual appearances.

by the recently proposed multimodal translation methods [50], [51]. In this work, we manually select appropriate translated images to be used in the second stage when the silhouette of a generated infrared image significantly differs from that of a visible image.

B. Intensity Modulation Network

Let M be a binary target mask, the output of the procedure in Fig. 1(b), with value 0 for the background region and 1 otherwise. Given the translated background image I, an image with the target mask implanted on I_t is given by

$$I_t = 255 \times M + I \odot (1 - M) \tag{1}$$

where $\odot$ denotes the elementwise multiplication. Here, we assume that we deal with an 8-bit image. For the proposed intensity modulation network G, we use I_t as the input. G predicts the scaling map S for I_t

$$S = G(I_t). \tag{2}$$

By applying S only to the target region, we can obtain the final synthetic image I_f as follows:

$$I_f = S \odot I_t \odot M + I_t \odot (1 - M). \tag{3}$$

To encourage G to produce a scaling map that changes the implanted mask into the realistic target, we use a discriminator

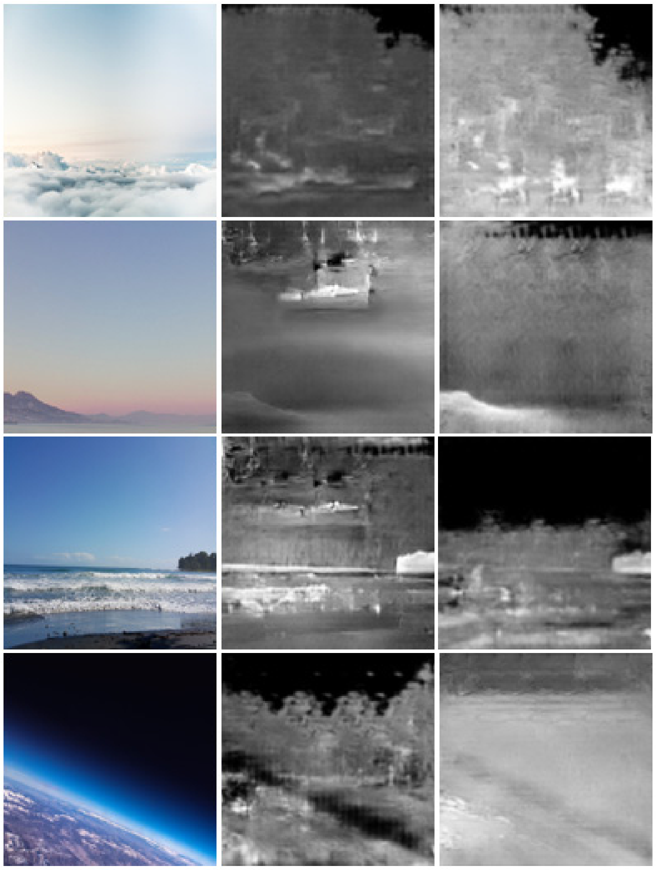

Fig. 3. Failure cases. There are large structural differences between the visible (leftmost) and the synthetic infrared images.

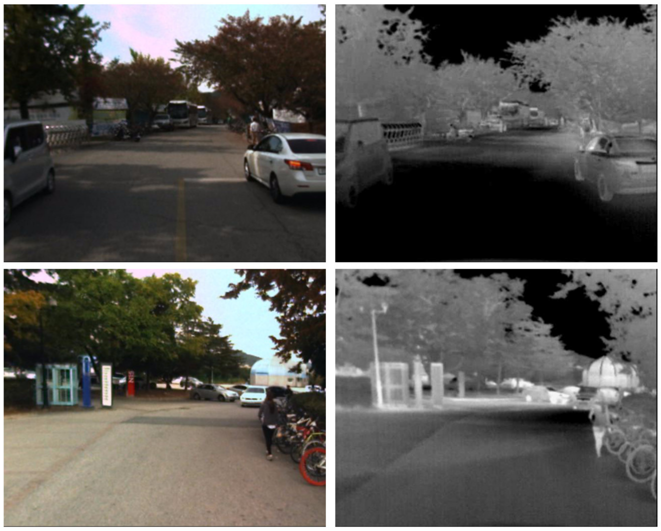

Fig. 4. Samples of training image pairs. (Left) Visible images. (Right) IR images.

D to perform adversarial learning. Even though there is no available supervision for S, S should adapt to pixel values of the background region that surrounds the target mask. Adversarial training can solve these problems without manual intervention. To train D and G, we use objective functions of

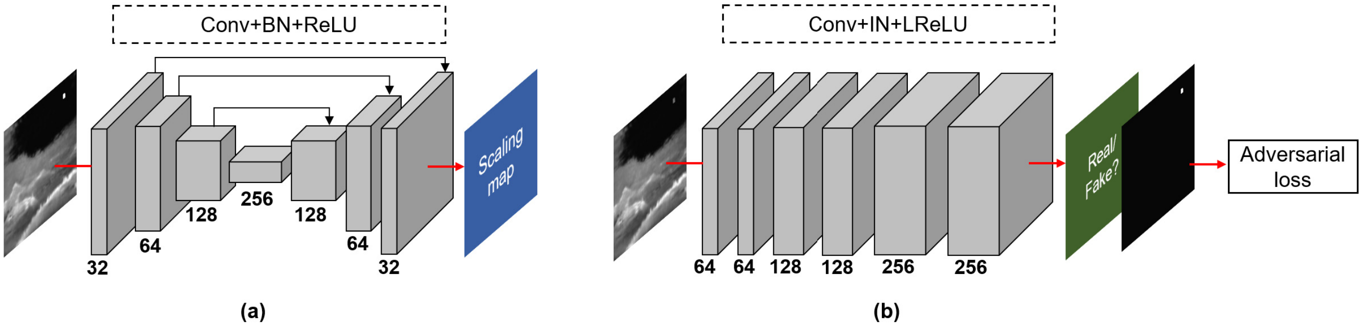

Fig. 5. Architecture of (a) U-Net-based intensity modulation network G and (b) PatchGAN discriminator D. The numbers of channels are given below each activation. The adversarial loss is evaluated only with those patches whose center is within the target region.

least-squares GAN (LSGAN) [52], which are given by

$$\min_D \mathcal{L}_{\text{adv}}(D) = \frac{1}{2}\mathbb{E}_{I_r}\left[(D(I_r)-1)^2\right] + \frac{1}{2}\mathbb{E}_{I_f}\left[(D(I_f))^2\right] \tag{4}$$

$$\min_G \mathcal{L}_{\text{adv}}(G) = \frac{1}{2}\mathbb{E}_{I_f}\left[(D(I_f)-1)^2\right] \tag{5}$$

where I_r is the real image. In the original GAN framework [25], the discriminator produces the output based on the features extracted from the whole image. However, as the proposed intensity modulation network is designed to adjust pixel values of the masked region only, the output of the discriminator should be derived only from the features of the target region. Contrary to the conventional global GAN, a discriminator in PatchGAN [27] slides a window over the input image and generates a score map that indicates whether each patch is real or fake. Therefore, we train the intensity modulation network by using the adversarial losses of local patches whose center is located within the target region.

We include one additional loss to train the intensity modulation network. Generally, an infrared small target appears as a light spot in the image [16], [21], which indicates that intensity values of the target should be much higher than those of the surrounding region. To penalize the reversal of the contrast between the target and its surrounding regions, we introduce contrast loss that is given by

$$\mathcal{L}_{\text{cont}}(G) = -\sum_{x\in R_{\text{tg}}} \ln\left(\frac{I_f(x)}{M_{\text{sr}}}\right) \tag{6}$$

where M_{sr} is the mean value of the surrounding region and the summation is calculated over the target region R_{tg}. Then, our overall loss for the intensity modulation network is

$$\mathcal{L}_{\text{total}}(G) = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} \tag{7}$$

where λ_{adv} and λ_{cont} are hyperparameters that adjust the contribution of different losses.

Fig. 5 shows the architecture of the proposed intensity modulation network and the PatchGAN discriminator. We adopt U-Net [33] architecture to build the proposed intensity modulation network, as shown in Fig. 5(a). It is composed of an encoder and a decoder with symmetric skip connections. Modules of the form 3 × 3 convolution (Conv)-batch normalization (BN)-rectified linear unit (ReLU) are used to construct the encoder and the decoder. The last 1 × 1 convolutional layer followed by the sigmoid function predicts the scaling map. As depicted in Fig. 5(b), the PatchGAN discriminator is composed of 6 convolutional layers, including instance normalization (IN) and leaky ReLU (LReLU) with a slope of 0.1. The size of the convolutional kernel starts from 7 × 7 and decreases by 2 for each dimension whenever the number of channels doubles. The extracted feature map is converted to the discriminator output by 1 × 1 convolution. In other studies [30], [44] that utilize PatchGAN discriminator, the spatial dimension of a feature map is reduced with multiple strided convolutional layers. As the reduction of width and height of a feature map makes it difficult or impossible to identify the patches corresponding to the target region, however, no strided convolution is adopted in our PatchGAN discriminator.

C. Baseline Network for Target Detection

In order to show the benefit of the proposed synthetic data augmentation method, we perform experiments on the baseline detection network and the state-of-the-art deep learning-based detection algorithms. The baseline detection network is based on the original U-Net [33]. First, U-Net was designed for automatic biomedical image segmentation. Due to its excellent performance on many vision-related problems, U-Net has become the preferred starting point for designing new network architectures. Fig. 6 shows the structure of the baseline detection network. U-Net consists of two subparts: the convolutional encoder and decoder with skip connections (Copy). The basic 3 × 3 convolutional operations (3 × 3 Conv) are followed by a BN layer and an ReLU activation. The encoder gradually decreases the spatial dimension of feature maps through 2 × 2 max-pooling operations with a pooling size of 2 × 2 and stride of 2 (MaxPool 2 × 2). This process is repeated three times in our network. After each downsampling step, the number of feature channels is doubled. The reverse of the operations performed at the encoder occurs at the decoder. The decoder upsamples feature maps using a 2 × 2 transposed convolution operation (Up-conv 2 × 2) and then concatenates the upsampled feature maps with the corresponding feature maps from the encoder. In the meantime, the number of

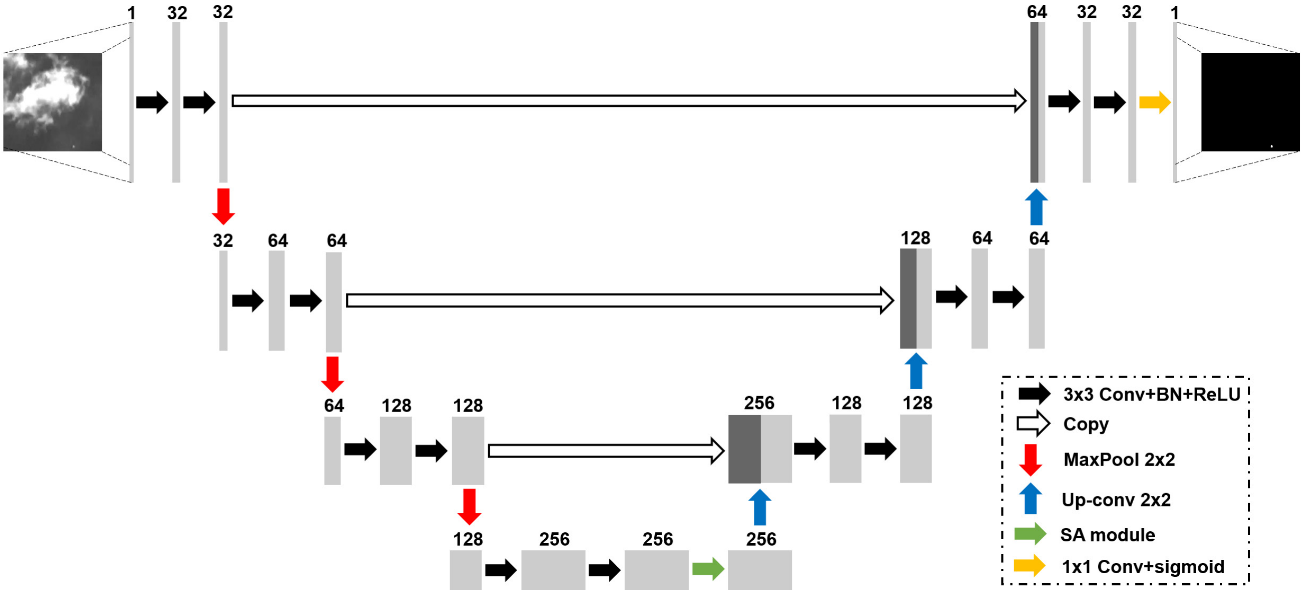

Fig. 6. Structure of the baseline detection network. The number of feature channel is given at the top of each feature map.

concatenated feature channels is halved. At the last layer, the 1 × 1 convolution (1 × 1 Conv) and the sigmoid activation function are used to produce the final result.

Between the encoder and the decoder, we insert the spatial attention (SA) module that automatically learns where to focus in the feature map [36]. Initially, the SA module was proposed as one of the components of the convolutional block attention module, and its effectiveness was shown in generic image classification and object detection [36]. Using the spatial relationship of features, the SA module outputs a 2-D weight map. To compute the weight map, max- and average-pooling operations are applied to the input feature map along the channel axis. Then, two pooling results are concatenated and further processed by a 7 × 7 convolution operator. Finally, the processed 2-D map is converted to the weight map using a sigmoid activation function. The output feature map is obtained by the elementwise multiplication of the input feature map with the calculated weight map.

IV. Experimental Results

A. Dataset and Evaluation Metrics

As noted before, our final goal is to improve the performance of CNN-based detection network trained with limited samples. We first describe the infrared small target detection dataset on which we evaluate the proposed synthetic training data generation process. In [16] and [18], the Single-frame InfraRed Small Target (SIRST) dataset is released to the public. From hundreds of infrared sequences, 427 representative frames are selected and manually annotated. Approximately 90% of image frames contain a single target. The SIRST dataset is split into 50% for train, 20% for validation, and 30% for the test.

To train BicycleGAN, we utilize two paired visible-infrared datasets, the KAIST multispectral pedestrian dataset [53] and the CVC-14 dataset [54]. Images of both datasets were captured by one infrared and one visible camera mounted on a moving vehicle. The KAIST multispectral pedestrian dataset contains approximately 95.3k images, and the CVC-14 dataset comprises roughly 8.4k images. Among them, we exclude nighttime sequences and sample training images only from daytime sequences. A total of 6212 visible-infrared image pairs are used to train BicycleGAN.

Traditional filtering methods are evaluated by metrics such as background suppression factor and signal-to-clutter gain. These metrics are designed to assess how successful the filtering approach can be at mitigating the background signal and improving (or at least preserving) the target signal. However, the deep networks output a binary mask, where the values of these metrics would be infinity in most cases [18]. To evaluate the detection performance for deep learning-based methods, we use four different evaluation metrics.

The first evaluation metric is normalized intersection over union (nIoU), which is adopted in [16] and [18]. nIoU is defined as

$$\mathrm{nIoU} = \frac{1}{N}\sum_{i}^{N}\frac{\mathrm{TP}[i]}{\mathrm{T}[i]+\mathrm{P}[i]-\mathrm{TP}[i]} \tag{8}$$

where TP, T, and P are the true positive, true, and positive, respectively. N denotes the total sample number. In order to achieve higher nIoU scores, the detection results should have the maximum overlap with the ground truth and the minimum false alarms. Therefore, nIoU quantifies the precision of target localization.

We also adopt the probability of detection (P_d) and the false alarm rate (F_a) to evaluate the effectiveness of the proposed data augmentation process. In the practical infrared small target detection system, a cluster of pixels is reported as a single alarm [21]. We calculate P_d and F_a by counting the number of clusters of pixels in the detection map. If the distance between the centroid of ground truth and that of

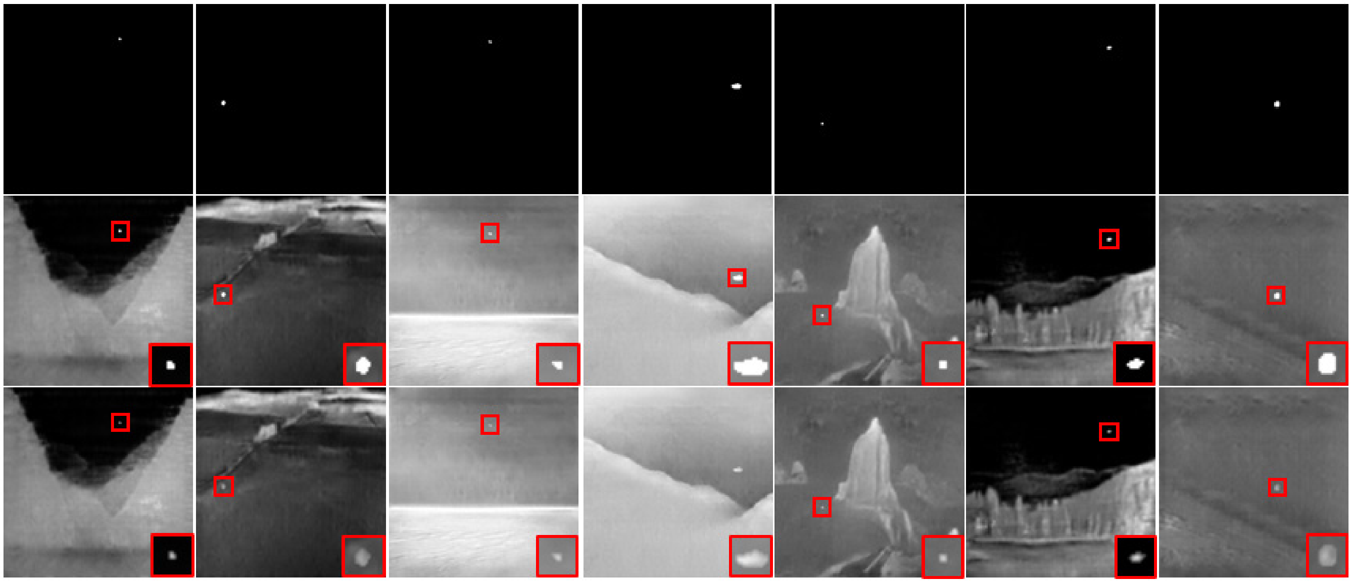

Fig. 7. Examples of synthetic images. (Top) Target masks. (Middle) Synthetic background images with a target mask implanted on them. (Bottom) Final results obtained by multiplying the pixel values of the target region with the scaling factor produced by the proposed intensity modulation network.

a detected position is within a threshold (three pixels), then the detection is considered as being a correct prediction and otherwise being a false alarm. To obtain the detection map, we fixed the threshold to 0.5.

Finally, the receive operating characteristic (ROC) curve is used to evaluate the detection performance. By varying the detection threshold from 0 to 1, it is possible to observe the trend of the probability of detection under various false alarm rates.

B. Implementation Details

To synthesize background infrared images from real visible images, we train the network of BicycleGAN from the scratch. The same network architecture in the original paper [30] is used. We train the translation network using the Adam optimizer with a batch size of 2 and a learning rate of 2×10^{-4}. Real visible images downloaded from the web are converted to synthetic infrared background images by the trained translation network. Prior to feeding the real images to the translation network, the collected real images are resized to a fixed resolution and cropped, as is done in the training. To construct the target mask library, we utilize the annotation data in the SIRST dataset. Specifically, each binary target map in the SIRST dataset is shifted so that the center of a target locates in the middle of the map. Then, each shifted annotation data are registered to the target mask library. The intensity modulation network is trained by the Adam optimizer with a batch size of 2 and a learning rate of 1×10^{-4}. We set the hyperparameters $\lambda_{\text{adv}} = 1$ and $\lambda_{\text{cont}} = 0.03$. Only the training set in the SIRST dataset is utilized for the adversarial training. All the networks used in our experiments are implemented with *PyTorch* and run on an NVIDIA Geforce RTX 3090.

C. Results on Real and Synthetic Data

First, we perform the visual comparison between the real and synthetic data. Fig. 7 depicts the target mask, the intermediate results after implanting the target mask into the synthetic infrared image, and the final results. Examples of some real images from the SIRST dataset are also given in Fig. 8 for comparison. As can be seen from Figs. 7 and 8, the appearance of the synthetic target resembles that of a real target. The border of a target is naturally smeared out, and the pixel values of the target are no longer the brightest in the image.

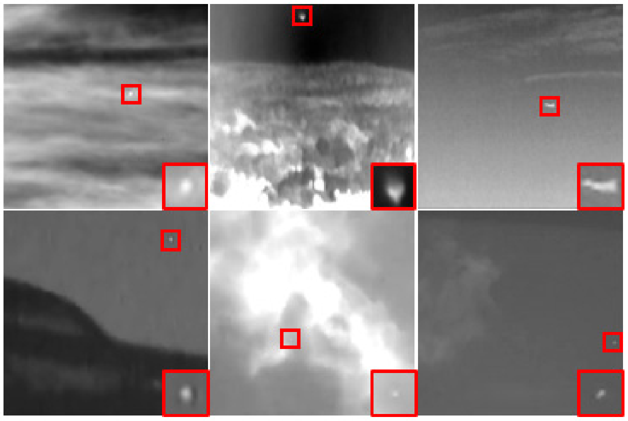

Fig. 8. Examples of real images from the SIRST dataset.

Note again that the final goal of this work is to improve the performance of the detection network with the help of synthetic training data, not to render realistic infrared images. Therefore, we analyze detection results by combining the real training data with different amounts of synthetic training data. Five detection networks, including four state-of-the-art detection models and the baseline, are selected to show the effectiveness of the proposed synthetic data augmentation. The selected models are MDvsFA [17], IRSTD-GAN [1], DNA-Net [19], and ACM U-Net [18]. Table II shows nIoU, P_d, and F_a values for the different amounts of synthetic

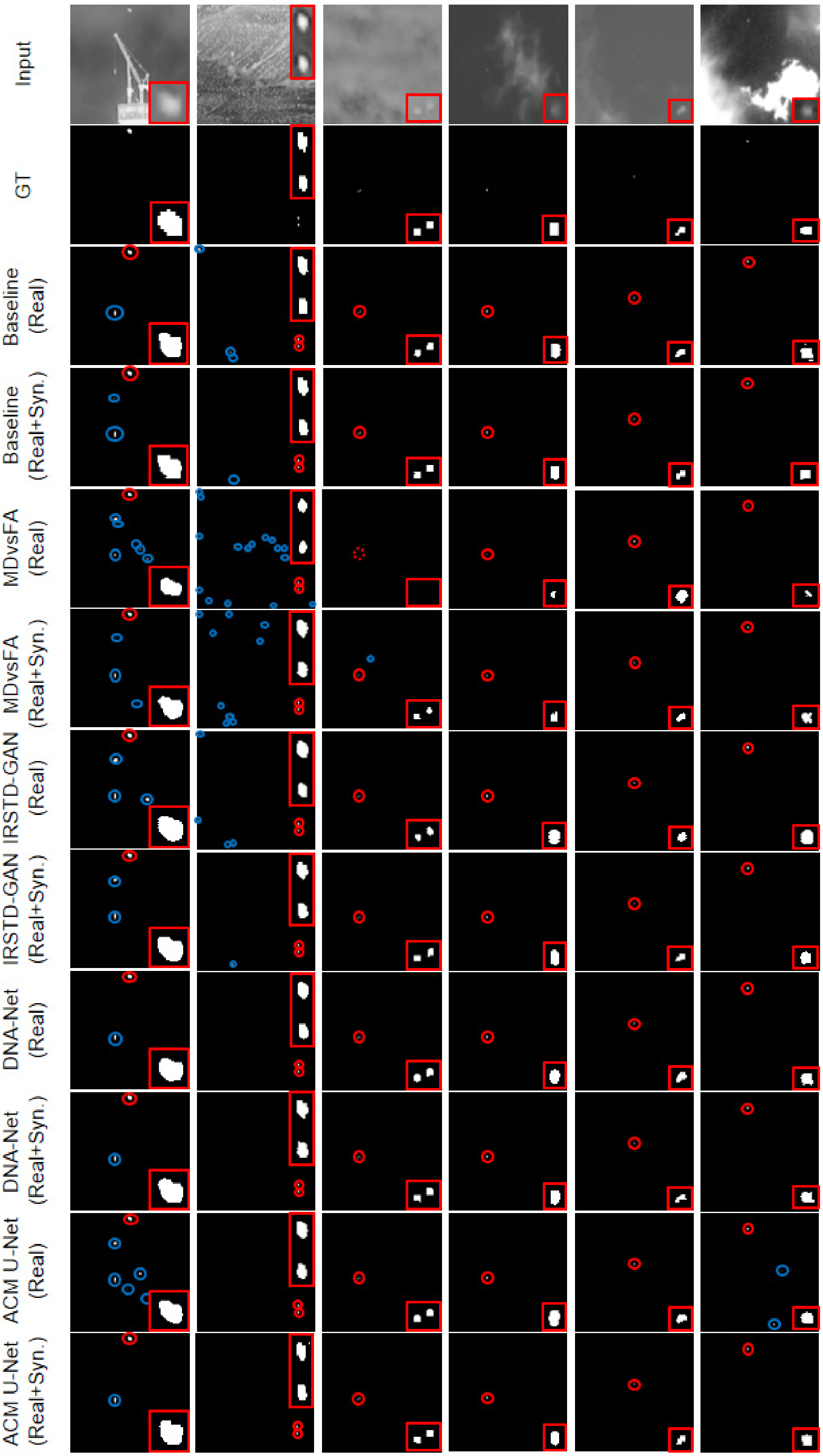

Fig. 9. Examples of the detection map. The target region is highlighted for better visualization. The correctly detected target, missing target, and the false alarm are encircled by red, red dots, and blue circles, respectively. The combined real and synthetic training data improve the detection results irrespective of detection models.

TABLE II

PERFORMANCE COMPARISON USING NIOU, PROBABILITY OF DETECTION (p_d), AND FALSE ALARM (F_a). THE BEST SCORES FOR EACH METHOD ARE HIGHLIGHTED IN BOLD

Number of Training Images	Real Synthetic Total	160 0 160	160 68 228	160 160 320	160 262 422	160 376 536	0 376 376
Baseline	nIoU	0.743	0.748	0.765	**0.769**	0.747	0.459
	P_d	0.971	**0.990**	0.980	0.980	0.961	0.775
	$F_a(10^{-5})$	5.062	2.801	4.486	**1.176**	3.286	66.706
MDvsFA [17]	nIoU	0.340	0.483	0.501	0.534	**0.606**	0.372
	P_d	0.676	0.8333	0.804	0.892	**0.951**	0.804
	$F_a(10^{-5})$	5.225	14.679	6.024	**3.564**	7.669	20.238
IRSTD-GAN [1]	nIoU	0.66	0.688	0.688	0.694	**0.704**	0.331
	P_d	0.961	0.961	**0.980**	0.951	0.961	0.569
	$F_a(10^{-5})$	5.285	4.911	7.506	4.582	**3.521**	10.343
DNA-Net [19]	nIoU	0.722	0.719	0.73	**0.743**	0.736	0.441
	P_d	0.951	0.980	0.990	0.981	**1.000**	0.686
	$F_a(10^{-5})$	2.889	1.081	**0.946**	1.864	1.232	3.699
ACM U-Net [18]	nIoU	0.709	0.724	0.733	0.738	**0.748**	0.3
	P_d	0.961	0.961	**0.980**	0.971	0.951	0.716
	$F_a(10^{-5})$	3.203	3.509	4.967	2.861	**2.265**	255.366

training data in the total training set. The following can be observed. First, the detection network trained only by synthetic training data produces much less accurate localization (nIoU) and detection (P_d and F_a) results. This implies that there still exists a domain gap between the real and synthetic data. Second, we can improve the accuracy of target localization and the detection performance by combining real training data with synthetic training data. Specifically, the proposed data augmentation method always improves the accuracy of target localization. For the detection performance, the model trained with combined training data shows better performance than the model trained with real training data only in at least one of the two metrics. These two observations are consistent in all five models.

For the qualitative comparison of the detection results, some examples of the detection map are shown in Fig. 9. We depict the results of the two models: one trained by 160 real images and the other trained by a combined 160 real and 262 synthetic images. Compared with the model trained by only real images, the model trained by the combined images produces fewer false alarms and better localization accuracies.

Finally, ROC curves for the five models are given in Fig. 10. In most cases, the state-of-the-art network models show higher true positive rates with lower false positive rates when the combined data of real and synthetic images are used for training. However, there are some mismatches between ROC curves and the quantitative results in Table II, especially for the baseline and ACM U-Net. In the traditional filtering approach, the output signal of the target area is generally proportional to the intensity of the target in the input image, given that the background is successfully suppressed. In our experiments, a deep network produces a detection map that mostly consists of zeros and ones. Thus, evaluating detection results by gradually changing the threshold as in ROC curves may be insufficient to represent the performance of the method for deep learning-based infrared small target detection.

TABLE III

EFFECTS OF λ_{cont} ON THE DETECTION RESULTS. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

λ_{cont}	nIoU	P_d	$F_a(10^{-5})$
0	0.745	0.980	3.461
0.03	**0.769**	0.980	**1.176**
0.1	0.761	**0.990**	4.437
0.3	0.760	0.980	2.774

TABLE IV

ABLATION STUDY ON THE INTENSITY MODULATION NETWORK

Description	nIoU	P_d	$F_a(10^{-5})$
No processing	0.751	0.971	**1.113**
Random values	0.746	0.971	4.900
Intensity modulation	**0.769**	**0.980**	1.176

D. Ablation Study on the Intensity Modulation Network

Ablation studies in this section are conducted using the baseline detection network. First, we investigate the effect of the hyperparameter λ_{cont} on the detection performance. We generate the synthetic dataset using the given value of λ_{cont}. Then, we train the detection network using 160 real and 262 synthetic images. The comparative results on the detection performance are given in Table III. Introducing the contrast loss can enhance the accuracy of target localization, as shown in Table III. Although $\lambda_{\text{cont}} = 0.03$ provides the highest nIoU value, λ_{cont} larger than 0.03 achieves similar target localization performance. Regarding detection performance, all of the compared models give similar P_d, and only F_a varies. As shown in Table III, the highest P_d is reported when λ_{cont} is set to 0.1, and the lowest F_a is achieved when λ_{cont} is set to 0.03.

Next, we evaluate the effectiveness of the intensity modulation network. We compare the proposed intensity modulation network with two variants of assigning pixel values to the target region: one is directly using the implanted target, and the other is to set a random value to the target. When we randomly select the pixel value for the target, only the value that is larger than the mean value of the region surrounding the target is considered. Then, the Gaussian blur is applied to the target region. The detection results evaluated with nIoU, P_d, and F_a are given in Table IV. Even if the proposed intensity modulation network is not applied in the data generation

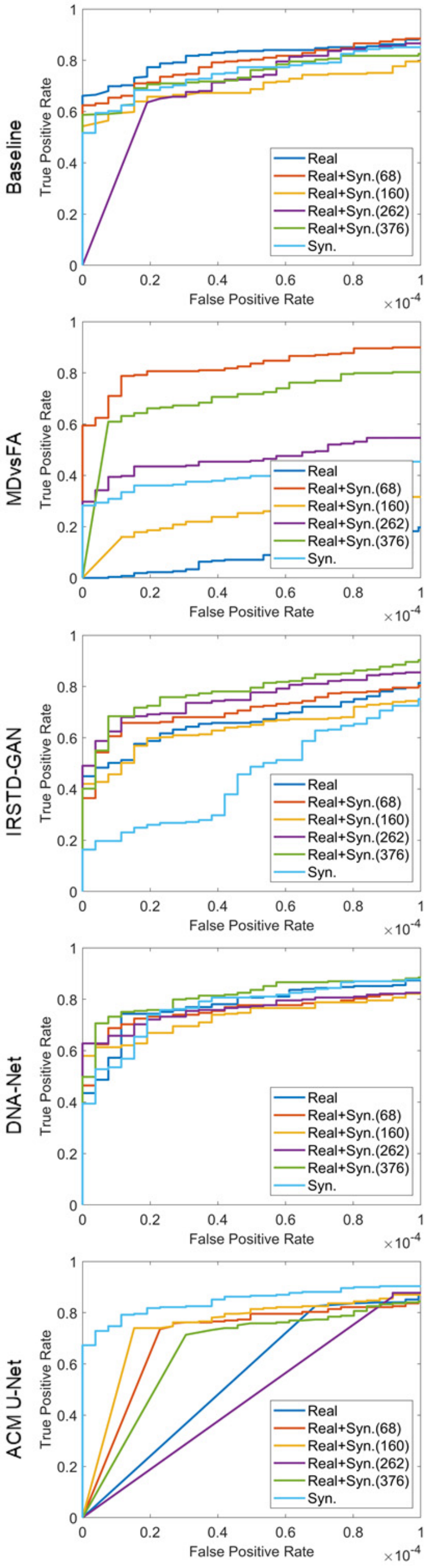

Fig. 10. ROC curves for the five detection models.

process, translated images with artificial targets can increase the nIoU score and decrease F_a compared to that of the model trained only with real images. By adopting the proposed intensity modulation network, we can further improve the nIoU score and P_d.

V. Conclusion

In this work, a new GAN-based synthetic data generation process is proposed to supplement the training dataset. Without a complex simulation model for the scene and imaging process, we can generate the synthetic training data using visible images obtained from the web and the recently released public dataset for infrared small target detection. Although there is a domain gap between the generated images and real ones, experimental results show that various detection networks trained with the mix-up of real and synthetic data outperform than using real images alone. The ablation studies show that the intensity modulation network contributes to the improvement of detection performance effectively. In the future work, we plan to explore domain adaptation techniques to fill the domain gap between the synthetic and real.

References

[1] B. Zhao, C. Wang, Q. Fu, and Z. Han, “A novel pattern for infrared small target detection with generative adversarial network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4481–4492, May 2021.

[2] D. Pang, T. Shan, W. Li, P. Ma, and R. Tao, “Infrared dim and small target detection based on greedy bilateral factorization in image sequences,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 99, pp. 3394–3408, Jun. 2020.

[3] V. T. Tom, T. Peli, M. Leung, and J. E. Bondaryk, “Morphology-based algorithm for point target detection in infrared backgrounds,” *Proc. SPIE*, vol. 1954, pp. 2–11, May 1993.

[4] S. Kim and J. Lee, “Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track,” *Pattern Recognit.*, vol. 45, no. 1, pp. 393–406, Jan. 2012.

[5] T. W. Bae, F. Zhang, and I. S. Kweon, “Edge directional 2D LMS filter for infrared small target detection,” *Infr. Phys. Technol.*, vol. 55, no. 1, pp. 137–145, Jan. 2012.

[6] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, “A local contrast method for small infrared target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014.

[7] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, “A robust infrared small target detection algorithm based on human visual system,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2168–2172, Dec. 2014.

[8] Y. Wei, X. You, and H. Li, “Multiscale patch-based contrast measure for small infrared target detection,” *Pattern Recognit.*, vol. 58, pp. 216–226, Oct. 2016.

[9] P. Yang, L. Dong, and W. Xu, “Detecting small infrared maritime targets overwhelmed in heavy waves by weighted multidirectional gradient measure,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[10] M. Zhao, L. Li, W. Li, R. Tao, L. Li, and W. Zhang, “Infrared small-target detection based on multiple morphological profiles,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6077–6091, Jul. 2021.

[11] M. Zhao, W. Li, L. Li, P. Ma, Z. Cai, and R. Tao, “Three-order tensor creation and tucker decomposition for infrared small-target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[12] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, “Single-frame infrared small-target detection: A survey,” *IEEE Geosci. Remote Sens. Mag.*, early access, Feb. 16, 2022, doi: 10.1109/MGRS.2022.3145502.

[13] C. Q. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, and A. G. Hauptmann, “Infrared patch-image model for small target detection in a single image,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4996–5009, Dec. 2013.

[14] Y. Dai, Y. Wu, Y. Song, and J. Guo, “Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values,” *Infr. Phys. Technol.*, vol. 81, pp. 182–194, Mar. 2017.

[15] Y. Dai and Y. Wu, “Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3752–3767, Aug. 2017.

[16] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021.

[17] H. Wang, L. Zhou, and L. Wang, "Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8509–8518.

[18] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 950–959.

[19] B. Li *et al.*, "Dense nested attention network for infrared small target detection," 2021, *arXiv:2106.00487*.

[20] R. Driggers *et al.*, "Detection of small targets in the infrared: An infrared search and track tutorial," *Appl. Opt.*, vol. 60, no. 16, pp. 4762–4777, 2021.

[21] K. Sungho and L. Joohyoung, "Small infrared target detection by region-adaptive clutter rejection for sea-based infrared search and track," *Sensors*, vol. 14, no. 7, pp. 13210–13242, 2014.

[22] N. Inkawhich *et al.*, "Bridging a gap in SAR-ATR: Training on fully synthetic and testing on measured data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2942–2955, 2021.

[23] J. H. Kim, J. Kim, and J. Joung, "Siamese hyperspectral target detection using synthetic training data," *Electron. Lett.*, vol. 56, no. 21, pp. 1116–1118, Oct. 2020.

[24] W. Liu, B. Luo, and J. Liu, "Synthetic data augmentation using multiscale attention CycleGAN for aircraft detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[25] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[26] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2387–2402, 2021.

[27] 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.

[28] J. R. Schott, *Remote Sensing: The Image Chain Approach*. Oxford, U.K.: Oxford Univ. Press, 2007.

[29] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2014, pp. 740–755.

[30] J.-Y. Zhu *et al.*, "Multimodal image-to-image translation by enforcing bi-cycle consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[31] M. S. Uddin, R. Hoque, K. A. Islam, C. Kwan, D. Gribben, and J. Li, "Converting optical videos to infrared videos using attention GAN and its impact on target detection and classification performance," *Remote Sens.*, vol. 13, no. 16, p. 3257, Aug. 2021.

[32] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–13.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[34] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[35] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[37] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-Image translation: Methods and applications," 2021, *arXiv:2101.08629*.

[38] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[39] 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.

[40] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, and F. S. Khan, "Synthetic data generation for end-to-end thermal infrared tracking," *IEEE Trans. Image process.*, vol. 28, no. 4, pp. 1837–1850, Apr. 2019.

[41] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[42] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3069–3082, Jul. 2020.

[43] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.

[44] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "DAugNet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1067–1081, Feb. 2020.

[45] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[46] S. Liu, M. Gao, V. John, Z. Liu, and E. Blasch, "Deep learning thermal image translation for night vision perception," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 1, pp. 1–18, Feb. 2021.

[47] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3683–3692.

[48] S. Mo, M. Cho, and J. Shin, "InstaGAN: Instance-aware image-to-image translation," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–26.

[49] D. Bhattacharjee, S. Kim, G. Vizier, and M. Salzmann, "DUNIT: Detection-based unsupervised image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4787–4796.

[50] D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee, "Diversity-sensitive conditional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–23.

[51] K. Li, S. Peng, T. Zhang, and J. Malik, "Multimodal image synthesis with conditional implicit maximum likelihood estimation," *Int. J. Comput. Vis.*, vol. 128, nos. 10–11, pp. 2607–2628, Nov. 2020.

[52] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[53] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[54] A. González *et al.*, "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, p. 820, Jun. 2016.

Jun-Hyung Kim (Member, IEEE) received the B.S. and Ph.D. degrees in electronics engineering from Korea University, Seoul, Republic of Korea, in 2006 and 2012, respectively.

He worked for the Agency for Defense Development, Daejeon, Republic of Korea, from 2012 to 2021, as a Senior Researcher, working on developing algorithms for electro-optical/infrared sensors. He then spent a half year at the Department of Intelligent Systems and Robotics, Chungbuk National University, Cheongju, Republic of Korea, as a Post-Doctoral Researcher. His research interests include image processing, deep learning, and their applications in target detection and image understanding.

Youngbae Hwang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2001, 2003, and 2009, respectively.

After six months as a Post-Doctoral Researcher at KAIST, he worked for the Robot Business Group, Samsung Techwin, Changwon, Republic of Korea, as a Senior Research Engineer, for one and a half years. He was a Senior Researcher with the Korea Electronics Technology Institute, Seongnam, Republic of Korea, for eight years. He is currently an Assistant Professor with Chungbuk National University, Cheongju, Republic of Korea. His research interests include image noise modeling, low-level image processing, computational photography, and medical image processing.