Pattern Recognition 153 (2024) 110501

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Generative data augmentation by conditional inpainting for multi-class object detection in infrared images

Peng Wang [a,b,1], Zhe Ma [a,b,1], Bo Dong [a,b,d], Xiuhua Liu [a,b], Jishiyu Ding [a,b], Kewu Sun [a,b], Ying Chen [c,*]

[a] Intelligent Science and Technology Academy Limited of CASIC, 100043, Beijing, China
[b] Key Lab of Aerospace Defense Intelligent System and Technology, 100043, Beijing, China
[c] Institute for Stroke and Dementia Research, University Hospital, Ludwig-Maximilians University Munich, 81377, Munich, Germany
[d] Department of Control Science and Engineering, Harbin Institute of Technology, 150006, Harbin, China

ARTICLE INFO

Dataset link: https://github.com/RonaldoPeng/MCTIFD

Keywords:
Data augmentation
Image inpainting
GAN
Infrared image
Multi-class object detection

ABSTRACT

Multi-class object detection in infrared images is important in military and civilian use. Deep learning methods can obtain high accuracy but require a large-scale dataset. We propose a generative data augmentation framework DOCI-GAN, for infrared multi-class object detection with limited data. Contributions of this paper are four-folds. Firstly, DOCI-GAN is designed as a conditional image inpainting framework, yielding paired infrared multi-class object image and annotation. Secondly, a text-to-image converter is formulated to transform text-format object annotations to bounding box mask images, leading the augmentation to be mask-image-to-raw-image translation. Thirdly, a multiscale morphological erosion-based loss is created to alleviate the intensity inconsistency between inpainted local backgrounds and global background. Finally, for generating diverse images, artificial multi-class object annotations are integrated with real ones during augmentation. Experimental results demonstrated that DOCI-GAN augments dataset with high-quality infrared multi-class object images, consequently improving the accuracy of object detection baselines.

1. Introduction

Infrared (IR) cameras are resistant to illumination variations, and thus have robust all-day performance. IR cameras steadily gain popularity in many domains, especially in military surveillance and remote sensing applications. When IR cameras are integrated into the intelligent systems designed for these applications, automatic object detection in IR images becomes a fundamental task.

In recent years, flourishing deep learning technology has brought vigor and vitality to object detection in natural image domain, e.g. RCNN family models [1,2], YOLO-like models [3,4]. Profiting from the excellent representation learning capability, deep learning-based object detection methods break the bottleneck of traditional methods and increase the accuracy by a large margin. The success of deep learning methods in natural image domain, spurs their application in IR image domain for detecting vehicles, pedestrian, and small targets [5].

However, the impressive performance of deep learning heavily relies on large-scale annotated datasets, which proves challenging to obtain in the case of massive IR images. More crucially, manually annotating bounding boxes for objects is time-consuming. As a consequence, deep learning for IR object detection is confronted with the data scarcity problem in real-world scenarios. Data augmentation (DA) is a promising technique to alleviate this problem by artificially enlarging datasets through the generation of new samples from existing samples, thus providing deep learning methods with sufficient training data.

In this work, our focus lies on IR multi-class object detection task. We study DA technique to tackle this task in scenarios where training data is scarce. We formulate a DA pipeline rooted in an image inpainting framework [6], to generate IR object images based on provided bounding box annotations. A novel inpainting framework is proposed to reconstruct an object with given position and category as conditions. The contributions of this paper are four-folds. Firstly, object **d**etection **o**riented **c**onditional **i**npainting GAN (DOCI-GAN), is designed to automatically produce both IR multi-class images and object annotations without requiring additional manual labeling. Secondly, a text-to-image converter is constructed to transform text-format object annotations

* Corresponding author.
E-mail addresses: paulwp@buaa.edu.cn (P. Wang), zhemazhe@yeah.net (Z. Ma), bob_dongbo@yeah.net (B. Dong), xiuhualiu@pku.edu.cn (X. Liu), dingjishiyu@126.com (J. Ding), sun_kewu@126.com (K. Sun), Ying.Chen@campus.lmu.de (Y. Chen).
[1] Authors contribute equally.

https://doi.org/10.1016/j.patcog.2024.110501
Received 21 December 2023; Received in revised form 8 April 2024; Accepted 12 April 2024
Available online 16 April 2024

into image-format bounding box masks. Accordingly, DA is reformulated as image-to-image translation, thereby boosting augmentation efficiency. Thirdly, a noticeable distinction exists between the local background intensity distributions of inpainted objects and those of the surrounding global background. To address this issue, we create a novel multiscale morphological erosion-based Mean Squared Error (MSE) loss, aimed at mitigating intensity inconsistencies around the borders of inpainted regions. Finally, to enrich the diversity of augmented IR images, various artificial object annotations are introduced alongside real ones.

We conducted experiments to assess the efficacy of DOCI-GAN. Results indicate that DOCI-GAN generates IR multi-class object images with high plausibility and diversity. Thereby, DOCI-GAN provides deep learning detection models with sufficient training data backup, resulting in significant improvement in object detection accuracy.

2. Related works

Our study builds upon previous literature in image inpainting, data augmentation, IR object detection domains. This section provides an overview of relevant works.

2.1. Image impainting

Image inpainting aims to fill in missing regions of images with reasonable and fine-grained texture. Nowadays, deep learning neural networks show their advance in capturing local context and texture information, thereby enabling accurate inference of missing regions. Among these networks, generative adversarial networks (GANs) have become dominant in the past decade. For example, Chen et al. [7] integrated both a global discriminator and a local discriminator within the GAN framework to ensure the global coherence and authenticity of local details in the generated images. Besides utilizing two discriminators, Zhang et al. [8] further embedded domain knowledge through a hierarchical variational auto-encoder into the latent variable space to guide inpainting process.

Lately, diffusion models have emerged as a cutting-edge category of deep generative models. Founded on the non-equilibrium thermodynamics theory, J. Ho et al. [9] proposed the denoising diffusion probabilistic model (DDPM), showcasing remarkable ability in generating diverse image with high-level details. R. Rombach et al. [10] introduced cross-attention layers into diffusion models, to generate image from conditional inputs, such as text. For image inpainting task, Zhang et al. [11] incorporated a Bayesian framework into diffusion model, to jointly modify both revealed and unrevealed regions, leading to improved coherence in the inpainted images. Corneanu et al. [12] simplified the process of conditioning diffusion models by a new conditioning mechanism that works in latent spaces, reducing computational costs for inpainting diffusion models.

2.2. Data augmentation

Data augmentation (AD) presents as an explicit solution to data scarcity by exploiting deep generative models for generating samples with complex and rich variations. Recent works have explored the potential of multiple GANs for DA. For example, CycleGAN can be utilized in domain transfer, translating images from one domain with sufficient annotated data to enlarge the training dataset for a target domain [13]. Bosquet et al. [14] designed a Downsampling GAN tailored for small object augmentation, which generates smaller objects from larger ones then places them into an existing background. Bailo et al. [15] utilized conditional GAN to generate photorealistic blood cell images, utilizing segmentation masks as conditional information. In this work, we focus on adopting the conditional GAN framework for multi-class object DA. The main difference between Bailo's work and ours is that we aim to generate IR images for object detection task, where multi-class objects are inpainted with given conditional bounding box information.

2.3. Object detection in IR images

In the IR image domain, object detection tasks are usually centered around pedestrian, vehicles and small targets. Small target detection is currently a trending topic [16] where the low contrast between small targets and noisy background poses a challenge [17]. A typical strategy to improve the performance of small target detection models is to effectively fuse high-level and low-level features, ensuring simultaneous extraction of semantics and preservation of details [18,19]. Another frequently used approach is utilizing attention modules to enhance features of small targets [20,21].

Different from small targets, IR objects we study in this work are big enough to discern their object categories, thus we can carry out multi-class object detection, including vehicles and pedestrian. In the realm of IR multi-class object detection, Li. et al. [22] adapted YOLO model by devising a feature extraction module to fully exploit both shallow and detailed features, along with a multi-layer detection head for identifying weak and small objects within IR dataset. Dai. et al. [23] integrated a SSD with a residual branch, capable of being removed during inference, to construct a lightweight network with high detection efficiency for vehicle and pedestrian datasets. Similarly, Jiang et al. [24] leveraged YOLO models for object detection in UAV thermal IR images.

3. Methods

The flowchart of the proposed DOCI-GAN is illustrated in Fig. 1. In this section, we introduce the core idea and main architecture of DOCI-GAN. Then we present additional details of DOCI-GAN, including bounding box mask generation and training loss.

3.1. Detection oriented conditional inpainting GAN

Augmenting training data for supervised learning in IR object detection necessitates generating paired IR image and bounding box annotation. Bounding box annotations can be readily fabricated artificially. Therefore, we propose to address the DA problem of IR object images through conditional image generation. Here, multi-class IR objects are generated under the condition of bounding box annotation.

Center on the aforementioned idea, we build a model for IR image generation using conditional GAN. Conditional GAN shows its particular competence in image-to-image translation. The most famous work is Pix2Pix [25], which enables image reconstruction from label maps or edge maps. Accordingly, in our work, rather than employing word embedding techniques to convert text-format bounding box annotations into vectors for conditional input, we attain bounding box mask images from annotations and frame the DA process as mask-image-to-raw-image translation with our proposed DOCI-GAN.

As depicted in Fig. 1, in DOCI-GAN, starting from bounding box masks $x_m \in \mathbf{R}^{H\times W\times 1}$, where H, W represent height, width of the mask image respectively, the generator is to translate these mask images to IR object images $y_f \in \mathbf{R}^{H\times W\times 1}$. The bounding box masks delineate rectangle areas and the categories of objects. Bounding box masks is obtained by a text-to-image converter, which will be elaborated in next subsection. To streamline the generation process, we also incorporate background images $x_b \in \mathbf{R}^{H\times W\times 1}$ into the input, obviating the need for the model to fabricate intricate backgrounds. Generated images with inferior or incorrect background are thus avoided.

With background images as conditional input, the mask-image-to-raw-image translation actually becomes an image inpainting process. The objective of the generator is to accurately inpaint multi-class objects within specified rectangular areas and categories onto the given backgrounds. Hence, we name our DA model for IR images as **d**etection **o**riented **c**onditional **i**npainting GAN, or DOCI-GAN.

Detailed pipeline of DOCI-GAN is displayed in Fig. 1. The generator takes random noise $x_n \in \mathbf{R}^{H\times W\times 1}$ as input, together with bounding

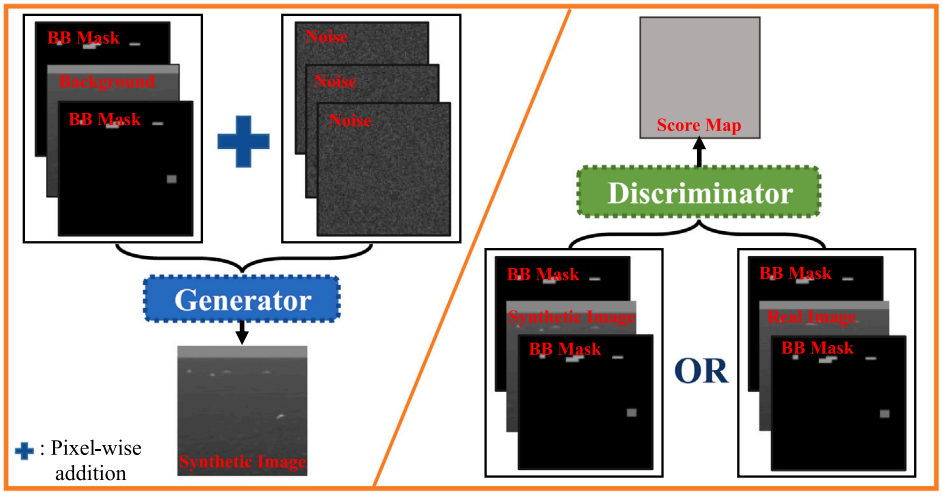

Fig. 1. Illustration of proposed DOCI-GAN pipeline. The generator is to generate IR images with specified objects based on provided bounding box (BB) masks and background images. The discriminator is to distinguish between the generated IR object images and real images.

box mask x_m and background image x_b as condition, to generate IR image y_f. Inspired by the work [26], we adopt a symmetric structure for the input of generator. This symmetric setup involves replicating the bounding box mask on both sides of the background image, creating a sandwich-like conditional input denoted as $x_{c-sandw} = [x_m, x_b, x_m] \in \mathbf{R}^{H\times W\times 3}$. By this symmetric structure, influence from intensity discrepancies between the bounding box mask and the background image can be mitigated. $x_{c-sandw}$ is then combined with three-channel random noise $x_{n-sandw} = [x_n, x_n, x_n] \in \mathbf{R}^{H\times W\times 3}$ and fed into the generator.

Upon the completion of object generation by the generator within IR background images, the discriminator steps in to distinguish between the synthetic fake image y_f and original real image $y_r \in \mathbf{R}^{H\times W\times 1}$. Adhering to the conventional setting in conditional GAN, conditional information, that is the bounding box mask x_m, is inputted into discriminator meanwhile. x_m enables the discriminator to disregard the background and concentrate solely on assessing the quality of the generated objects in comparison to the real ones. Moreover, in order to enhance the convergence stability of DOCI-GAN, we employ the symmetric setup for the discriminator input. Concretely, as illustrated in Fig. 1, the discriminator receives synthetic IR image along with bilateral bounding box masks as the fake input $z_{f-sandw} = [x_m, y_f, x_m]$, and the real image with bilateral masks as the real input $z_{r-sandw} = [x_m, y_r, x_m]$.

In traditional conditional GANs, the discriminator typically produces a single scalar value as an assessment of authenticity. Yet, relying solely on such a scalar for the adversarial loss of both generator and discriminator causes unreliability. This single scalar output indicates the authenticity of an image globally but lacks attention on every subregion. In contrast, the discriminator in DOCI-GAN addresses this limitation by producing a score map as the output to discern between fake and real inputs. With the score map as output, the discriminator empowers the generator to generate high-resolution, fine-grained IR multi-class object images. This approach ensures that attention is paid to the local authenticity of every object within the synthesized images.

3.2. Network architecture of generator and discriminator

In DOCI-GAN, the generator adopts a U-shaped architecture, a design proven to be effective in image inpainting [27]. Leveraging the U-shaped framework, the generator encodes contextual information from bounding box masks and IR background images into multi-level features. These features are then integrated through skip connections and decoded to produce IR images with reasonable global context and sufficient local details. Additionally, we introduce residual connections within the generator. These connections link consecutive convolutional layers in every stage of the U-shaped architecture, forming residual blocks. With residual blocks, the generator ensures robust training while preserving crucial structural and contextual information, thereby contributing to the generation of excellent inpainting results.

Detailed architecture of the generator is displayed in Fig. 2. Hyperparameter setting of the generator can be found in Table 1. Moreover, we integrate non-local attention module [28] in the generator. This module enables the capturing of long-range correlations among different regions, boosting the generator to learn global contextual features for inpainting reasonable objects and their surroundings.

Architecture of the discriminator is also illustrated in Fig. 2. It comprises five successive convolutional layers to produce a score map differentiating real input $z_{r-sandw}$ from fake input $z_{f-sandw}$. With a shallow model structure, the discriminator efficiently captures detail difference between forged objects and real ones, encouraging the generator to synthesize plausible IR object images.

3.3. Bounding box mask generation

Each original IR object image has the bounding box annotations stored as text, following the COCO format [29]. In this format, a JSON structure collects information on labels and metadata, including the coordinates of bounding boxes and the category of each object. Given a COCO file, we utilize a text-to-image converter, named T2I converter, to generate bounding box masks as conditional information for DOCI-GAN. Concretely, T2I extracts the coordinates of bounding boxs along with the category labels of objects, and then convert this information into a bounding box mask image.

In the bounding box mask images as shown in Fig. 3, each bounding box delineates an area containing an object. These bounding boxes are then assigned different gray values, corresponding to different categories. The rest areas in every bounding box mask image represent the background and are assigned a value of zero.

When augmenting IR object image data, we artificially introduce new bounding boxes and add them to the bounding box masks of original IR images. To ensure the relationality of augmented objects, we constrain the object to generate is spatially close to a real object of the same category. Specifically, these additional bounding boxes are randomly positioned near the original IR objects belonging to the same category. Thereby, generated objects of each class remain nearby authentic samples of the corresponding class, preventing the occurrence of illogical object positions. With these artificial bounding box masks, DOCI-GAN generates IR multi-class object image with increased diversity, realizing effective DA for object detection.

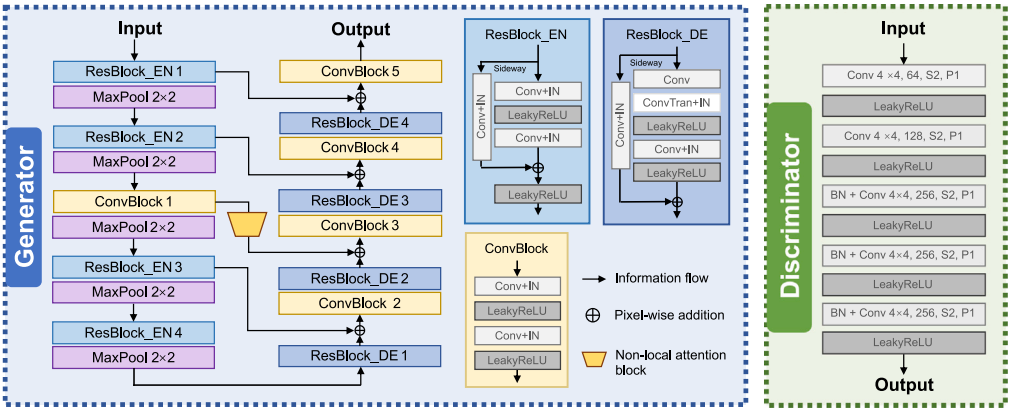

Fig. 2. Architectures of DOCI-GAN's generator and discriminator. Conv is the convolutional layer. IN is the in-stance normalization layer. ConvTran is the transposed convolutional layer. LeakyReLU is the activation layer. S represents stride length and P represents padding value.

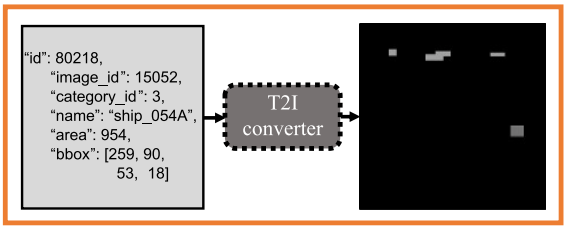

Fig. 3. Illustration of T2I converter. It transforms COCO-format bounding box annotations (displayed in the left rectangle) to bounding box mask images (displayed in the right rectangle).

Table 1
Parameters of DOCI-GAN's generator. Conv $m \times m, n$ means convolutional layer with kernel size $m \times m$ and channel number n. S means stride length and P means padding size.

Block	Convolutional layers setting
ResBlock_EN 1	[Conv 5 × 5, 64, S1, P2] × 2 *Sideway*: Conv 1 × 1, 64, S1
ResBlock_EN 2	[Conv 5 × 5, 128, S1, P2] × 2 *Sideway*: Conv 1 × 1, 128, S1
ResBlock_EN 3	[Conv 5 × 5, 256, S1, P2] × 2 *Sideway*: Conv 1 × 1, 256, S1
ResBlock_EN 4	[Conv 5 × 5, 512, S1, P2] × 2 *Sideway*: Conv 1 × 1, 512, S1
ResBlock_DN 1	Conv 1 × 1, 256, S1 ConvTran 4 × 4, 256, S2, P1 Conv 3 × 3, 256, S1, P1 *Sideway*: Conv 1 × 1, 256, S1
ResBlock_DN 2	Conv 1 × 1, 128, S1 ConvTran 4 × 4, 128, S2, P1 Conv 3 × 3, 128, S1, P1 *Sideway*: Conv 1 × 1, 128, S1
ResBlock_DN 3	Conv 1 × 1, 128, S1 ConvTran 4 × 4, 128, S2, P1 Conv 3 × 3, 128, S1, P1 *Sideway*: Conv 1 × 1, 128, S1
ResBlock_DN 4	Conv 1 × 1, 64, S1 ConvTran 4 × 4, 64, S2, P1 Conv 3 × 3, 64, S1, P1 *Sideway*: Conv 1 × 1, 64, S1
ConvBlock 1	[Conv 5 × 5, 128, S1, P2] × 2
ConvBlock 2	[Conv 5 × 5, 256, S1, P1] × 2
ConvBlock 3	[Conv 5 × 5, 128, S1, P1] × 2
ConvBlock 4	[Conv 5 × 5, 128, S1, P2] × 2
ConvBlock 5	[Conv 5 × 5, 64, S1, P2] × 2

3.4. Loss function

Let G and D represent the generator and discriminator of DOCI-GAN. To mitigate unstable training and mode collapse of deep generative model, we leverage the adversarial training loss presented in Wasserstein GANs [30] in DOCI-GAN. As an innovation to traditional GANs, Wasserstein GANs employ a smooth metric, the Wasserstein distance, to measure the dissimilarity between distributions of fake and real data. The adversarial training objective of DOCI-GAN is consequently formulated as Eq. (1).

$$L_{DOCI-GAN}(G,D) = -\mathbb{E}_{x_m,y_r} D(x_m,y_r) + \mathbb{E}_{x_m,y_f} D(x_m,y_f) + \lambda \mathbb{E}_{x_m,\hat{y}} (\| \nabla_{\hat{y}} D(x_m,\hat{y}) \|_2 - 1)^2 \tag{1}$$

In Eq. (1), x_m represents the bounding box mask. y_f represents the forged image while y_r represents the real image. $\hat{y}$ is uniformly sampled along straight lines connecting pairs of forged and real IR images. $\|\|_2$ represents 2-norm. $\lambda > 0$ is a trade-off parameter. In our experiments, λ is set to 0.05. Utilizing the above adversarial loss, G is trained to translate bounding box masks along with background images into IR images that closely resemble real ones.

During the training of G, we incorporate a perceptual loss $L_{perceptual}$ to assist G in generating IR object images semantically consistent with real ones. Here, we employ a perceptual loss defined in feature space as follows:

$$L_{perceptual}(G) = \mathbb{E}_{y_f,y_r} \frac{1}{HW} \|\varphi(y_f) - \varphi(y_r)\|_2 \tag{2}$$

In Eq. (2), φ is a feature extractor. H and W represent height and width of paired forged image y_f and real image y_r. In our implementation, we choose VGG-19 pretrained on the ImageNet dataset as φ. Since VGG-19 network takes RGB images as input while IR images are grayscale, we introduce the symmetric structure input comprising the bounding box mask along with IR image for the pretrained VGG-19. Features extracted by the 16th convolutional layer in VGG-19 are used for computing $L_{perceptual}$.

Additionally, we propose a bounding box loss L_{bbox} that leads G to focus on producing plausible IR objects within specified bounding boxes. The proposed bounding box loss is formulated as follows:

$$L_{bbox}(G) = \mathbb{E}_{x_m,y_f,y_r} \sum_{n=1}^{N} \frac{1}{H_m^n W_m^n} \|x_m^n y_f - x_m^n y_r\|_1 \tag{3}$$

In Eq. (3), N represents number of bounding boxes in a bounding box mask image x_m. H_m^n and W_m^n are respectively the height and width of the n_{th} bounding box. x_m^n represents the binary mask of n_{th} bounding box. $\|\|_1$ represents 1-norm. The bounding box loss emphasizes the penalty on generated object regions, so it prompts DOCI-GAN to generate realistic objects.

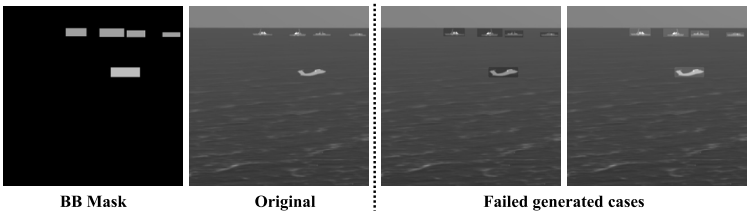

Fig. 4. Bounding box artifacts exist in generated IR images.

Multiscale Erosion-based MSE Loss: Since DOCI-GAN realizes DA as mask-image-to-raw-image translation, those bounding box masks as conditional input can cause artifacts in generated IR images. As shown in Fig. 4, the grayscale contrast between the bounding boxes and the surrounding areas in the mask image aids in distinguishing objects from the background during inference. As a result, there is a notable intensity disparity between the local background within the bounding boxes of the generated image and its overall background. Even though the generated objects within the bounding boxes may be of high quality, this discrepancy can make the entire IR image appear artificial, looking like objects simply pasted onto the background. To address this issue, we define a novel mean square error (MSE) loss based on multi-scale morphological erosion operation. Concretely, multi-scale erosion is applied to both the original and generated images, expanding the local backgrounds of objects to multiple spatial ranges. By minimizing the distance between the multi-scale eroded results of the original and generated images, the local and global backgrounds tend to be consistent. The proposed multi-scale morphological erosion-based MSE loss is formulated as follows:

$$L_{multi-erosion}(G) = \mathbb{E}_{y_f, y_r} \sum_{e \in E}^{E} \| y_f \ominus e - y_r \ominus e \|_2 \tag{4}$$

In Eq. (4), $\ominus$ represents erosion operation and E represents the set of multi-scale structural elements used for erosion. In experiments, E includes structural elements of size $(11, 21, 31, 41, 51, 61, 71, 81, 91)$.

4. Experiments

Below, we present a series of experimental results. Initially, we detail the dataset and experimental settings in our work. Then, we quantitatively and visually compare the performance of our proposed DOCI-GAN with other advanced deep generation methods. Finally, we train and test object detection models using the augmented datasets to validate the effectiveness of DOCI-GAN.

4.1. Dataset and experiment settings

The motivation of this work is to find a DA solution for multi-class object detection when confronted with limited training data. To this end, we constructed a dataset containing 64 images for training and 320 images for testing, with a uniform size 360 × 360. Objects within these IR images occupy an area ratio ranging from 0.001 to 0.01 of the entire image. These objects fall into 9 categories, including cruise missile, ship, surveillance plane, cargo plane, helicopter, fighter, bomber, big unmanned aerial vehicle (big UAV) and revolve unmanned aerial vehicle (revolve UAV). Every IR image has corresponding background image, with object annotation stored in a COCO-format JSON file. We name the dataset as multi-class thermal infrared few-shot detection (MCTIFD) dataset and make it publicly available in https://github.com/RonaldoPeng/MCTIFD.git.

In our experiments, DOCI-GAN was trained from scratch using Adam optimizer. We utilized a mini-batch size of 128, and the hyper-parameters for Adam were set as follows: $\alpha = 1e-4$, $\beta_1 = 0.5$, $\beta_2 = 0.9$. Throughout the training process, input noise maps for the generator were sampled from a uniform distribution within the interval [0, 0.05). We implemented DOCI-GAN with Pytorch library. All experiments were conducted on a platform equipped with 2 Intel Core e5-2640 CPU, and 128 GB RAM. For training purposes, 4 NVIDIA GeForce GTX 3090 GPUs were used.

4.2. Visual comparison of different methods

In this subsection, we compare forged IR object images generated by DOCI-GAN and other deep generative models, including conventional GAN [31], VAE-GAN [32], DDPM [9], BicycleGAN [33], Pix2Pix [25] visually. All methods were trained using the same training set. Fig. 5 showcases examples of generated images from different methods utilizing the MCTIFD dataset. For good visual effect, all generated IR images are adjusted with automatic contrast enhancement. Original IR images are in the first column for reference.

Among the compared methods, GAN, VAE-GAN, and DDPM were fed solely with random noise as input, while BicycleGAN, Pix2Pix, and DOCI-GAN received additional inputs of bounding box mask images and background images. GAN, VAE-GAN and DDPM generate images from scratch. In the images generated by GAN, while the texture looks plausible, the background is quite noisy, and the quality of generated objects are low. Instances of repeated objects of the same class often appear within a single generated image, which is illogical. VAE-GAN produces improved backgrounds, however, the generated objects exhibit poor quality, making them hard to classify. DDPM, as a diffusion model, generates better background but still struggles with object generation. These noise-input generation models are severely affected by the limited training dataset.

BicycleGAN, Pix2Pix and DOCI-GAN were inputted with bounding box mask images and background images, resulting in synthetic images with backgrounds resembling real ones, as shown in Fig. 5. However, the backgrounds produced by BicycleGAN exhibit considerable noise. The objects generated by BicycleGAN lack clarity, with some objects deviating from their original positions. In comparison, Pix2Pix produces images with significantly less noise. The forged IR objects generated by Pix2Pix exhibit improved quality and are precisely located at the positions of the original objects. Upon further comparison between Pix2Pix and our proposed DOCI-GAN, it is evident that the quality of both forged IR objects and backgrounds by DOCI-GAN surpasses that of Pix2Pix. The details of the background generated by DOCI-GAN closely resemble those of the original image. Moreover, the forged multi-class IR objects generated by DOCI-GAN are clear, featuring sharp boundaries and recognizable shapes.

Fig. 6 presents a series of examples showcasing multi-class IR objects generated by different methods. The first row in Fig. 6 shows real objects as a reference. Obviously, forged objects by GAN, VAE-GAN, DDPM and BicycleGAN are of low quality. Many of them fail to accurately reconstruct certain key characteristics, making classification difficult. In contrast, the forged objects generated by Pix2Pix and DOCI-GAN are more realistic, especially those generated by DOCI-GAN have best visual quality.

In summary, DOCI-GAN yields IR multi-class object images that exhibit superior visual quality compared to other methods. On one hand, it generates multi-class objects with sharp boundaries and intricate details. On the other hand, DOCI-GAN leverages the inputted background

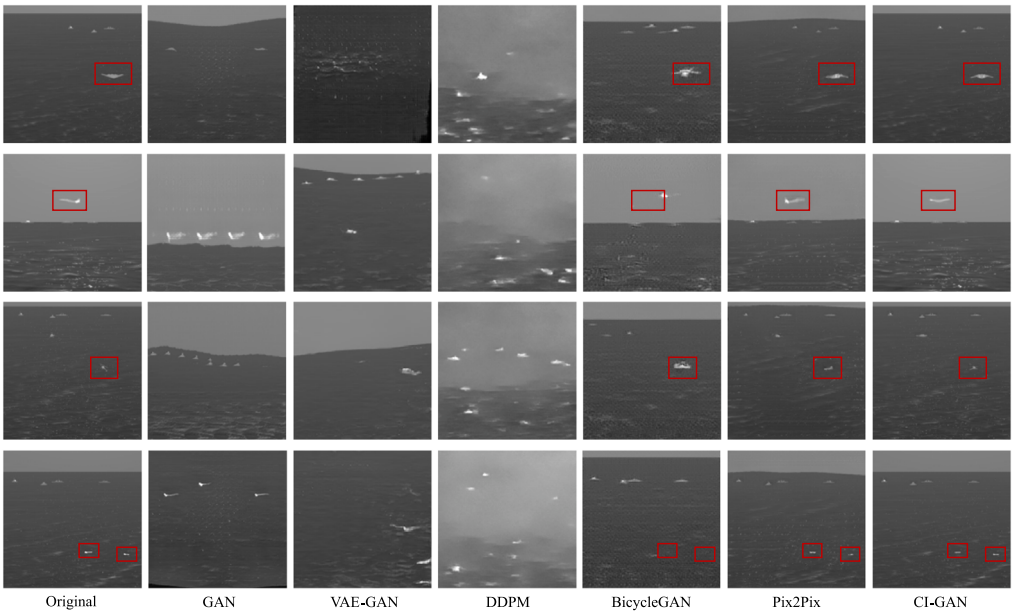

Fig. 5. Examples of generated IR object images by compared deep generative methods and DOCI-GAN, utilizing the MCTIFD dataset. The red rectangles highlight objects expected to appear in identical positions in both the original and forged images.

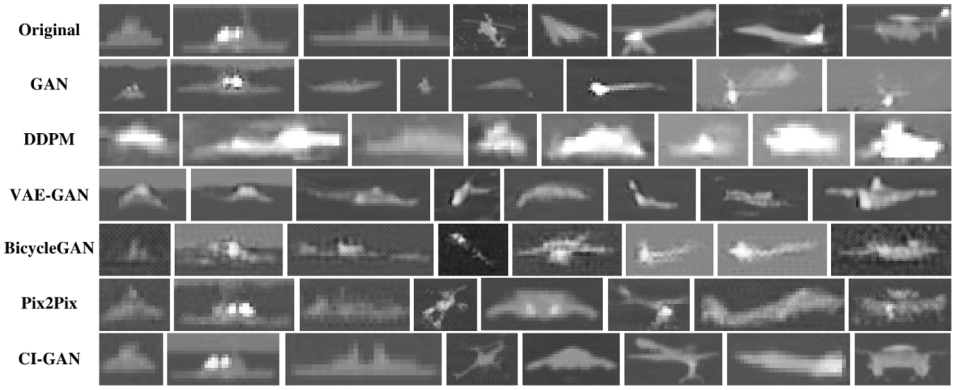

Fig. 6. Examples of multi-class IR objects generated by different generative models alongside original real objects from MCTIFD dataset.

images, resulting in forged images with more natural backgrounds and less noise. Regarding these two aspects, the visual comparison between the results of DOCI-GAN and other methods confirms the superiority of DOCI-GAN.

4.3. Evaluation metrics for augmented images

In the task of augmenting IR object images, when background images and bounding box masks are provided, the main aim of DOCI-GAN is to generate realistic objects. However, the objects we studied in this work usually occupy a small area in IR images. Hence, when assessing the performance of different methods for IR object image generation, we propose to primarily compare the quality of the generated objects. To be concrete, objects were extracted from generated IR images, and image quality assessment metrics were employed to evaluate these generated objects across different methods.

Quality of generated images is evaluated by objective metrics. Common full-reference metrics for image generation applications are based on assessing the difference between generated image and its corresponding real image, i.e. ground truth, such as MSE and peak signal to noise ratio (PSNR). Spatial alignment between generated image and ground truth is crucial for these metrics. However, during testing, as artificial bounding boxes were added to original bounding box masks, the generated IR object images have no exact ground truths that are spatially aligned. Therefore, in our experiments, we adopted reference-free image quality assessment (IQA) metrics to evaluate the quality of generated IR images. Specifically, seven metrics were utilized, including gray mean gradient (GMG), Tenenbaum gradient (G_T), energy gradient (G_E), Brenner gradient (G_B), sum of adjacent difference (SAD), spatial entropy (E_S), Inception Score (IS) [34] and Fréchet Inception Distance (FID) [35].

GMG is defined as follows:

$$GMG = \frac{1}{(H-1)(W-1)} \cdot \sum_{i=1}^{H-1} \sum_{j=1}^{W-1} \sqrt{\frac{\left|y_f^{(i,j)} - y_f^{(i+1,j)}\right|^2 + \left|y_f^{(i,j)} - y_f^{(i,j+1)}\right|^2}{2}} \tag{5}$$

H and W represent the height and width of the forged IR object image y_f. i and j denote the spatial indexes of pixels within the image.

G_T uses Sobel operators to compute gradient information in both the horizontal and vertical directions of an image. The calculation is as

follows:

$$G_T = \frac{1}{(H-2)(W-2)} \cdot \sum_{i=2}^{H-1} \sum_{j=2}^{W-1} \sqrt{S_x * y_f^{(i,j)} + S_y * y_f^{(i,j)}}$$
$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, S_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (6)$$

In Eq. (6), $*$ represents convolution operation.

G_E is the sum of square of grayscale difference between adjacent pixels, calculated as follows:

$$G_E = \frac{1}{(H-1)(W-1)} \cdot \sum_{i=1}^{H-1} \sum_{j=1}^{W-1} \left(\left| y_f^{(i,j)} - y_f^{(i+1,j)} \right|^2 + \left| y_f^{(i,j)} - y_f^{(i,j+1)} \right|^2 \right) \quad (7)$$

G_B computes the grayscale difference between a pixel and its neighbors with a horizontal or vertical distance of 2 as follows:

$$G_B = \frac{1}{(H-2)(W-2)} \cdot \sum_{i=1}^{H-2} \sum_{j=1}^{W-2} \left(\left| y_f^{(i,j+2)} - y_f^{(i,j)} \right|^2 \right) \quad (8)$$

SAD computes the absolute sum value of grayscale differences between adjacent pixels, as defined in Eq. (9).

$$SAD = \frac{1}{(H-1)(W-1)} \cdot \sum_{i=1}^{H-1} \sum_{j=1}^{W-1} \left(\left| y_f^{(i,j)} - y_f^{(i+1,j)} \right| + \left| y_f^{(i,j)} - y_f^{(i,j+1)} \right| \right) \quad (9)$$

GMG, G_T, G_E, G_B, SAD all measures image sharpness and contrast, where a higher value suggests better image quality. E_S quantifies the amount of information present in an image, with a higher E_S indicating richer content. E_S can be defined as follows:

$$E_S = -\sum_{k=0}^{255} p(k) \log_2 p(k), p(k) = \frac{h(k)}{HW} \quad (10)$$

In Eq. (10), $h(k)$ is the histogram of a forged image where k is the intensity index.

IS and FID are commonly used to assess the quality and diversity of images generated by generative models. IS correlates with human perceptual realism scores of images. It computes a statistic of network output by feeding generated images into a pretrained Inception model, and it is defined as follows:

$$IS = e^{\left[\mathbb{E}_{y_f}\left[D_{KL}(p(z|y_f)\|p(z))\right]\right]} \quad (11)$$

In Eq. (11), $D_{KL}(p \parallel q)$ represents the KL-divergence between two probability distributions p and q. Here, z denotes the class label for a generated image y_f. $p(z|y_f)$ represents the posterior probability of a label z conditioned on y_f, computed by the Inception model. A high IS suggests the generated images are of high quality and diversity.

In FID, both forged images and real images are inputted into a pretrained Inception model to extract visually relevant features. Assuming that the feature distribution follows a multivariate Gaussian distribution, the distance between the feature distributions of forged images and real images is defined as follows:

$$FID = \left\| \mu_r - \mu_f \right\|_2^2 + Tr\left(\Sigma_r + \Sigma_f - 2\sqrt{\Sigma_r \Sigma_f} \right) \quad (12)$$

In Eq. (12), (μ_r, Σ_r) and (μ_f, Σ_f) denote the mean and covariance of features extracted from real images and forged images, respectively. Tr denotes the trace of a matrix. A low FID suggests that the forged images closely resemble real images.

4.4. Quantitative comparison of different generative models

Table 2 presents a performance comparison between our DOCI-GAN and other deep generative models for image generation on the testing set of the MCTIFD dataset. This table shows that our proposed DOCI-GAN outperformed other models in terms of metrics G_T and G_B. This quantitative result indicates that DOCI-GAN generated IR object images with high image contrast and clarity. Moreover, from the perspective of FID metric, DOCI-GAN achieved a significantly lower value compared to the other models, indicating that the generated IR object images by DOCI-GAN closely resemble real ones.

Table 2
IQA results of objects generated by different deep generative models and the original objects from the MCTIFD dataset. The highest IQA results are highlighted in bold.

Method\measures	GMG	G_T	G_E	G_B	SAD	E_S	IS	FID
Original	4.98	174.70	72.09	973.40	89.56	0.29	1.61	–
GAN	2.89	81.56	32.85	496.14	55.66	0.24	**1.74**	123.60
VAE-GAN	3.13	89.52	39.03	390.06	62.91	0.23	1.42	87.08
DDPM	2.06	20.69	13.71	40.21	62.87	0.90	1.22	158.98
BicycleGAN	**6.61**	114.04	**102.24**	703.73	94.85	**0.48**	1.52	99.69
Pix2Pix	5.64	161.41	81.89	750.70	**104.79**	0.36	1.69	35.45
DOCI-GAN	4.91	**167.04**	70.48	**834.55**	91.66	0.29	1.53	**18.93**

Table 2 also highlights that BicycleGAN achieved favorable results in terms of GMG, G_E and E_S. High GMG, and G_E values obtained by BicycleGAN indicates that BicycleGAN also generated IR object images with high image contrast and sharpness. Additionally, a high E_S value obtained by BicycleGAN suggests that generated IR objects by BicycleGAN contain rich content. However, it is worth noting that noisy images could also yield high E_S values. In general, superior IQA results cannot confirm that the generated IR objects are more plausible. However, it is essential for the generated objects to approximate real ones. Therefore, we propose to compute the difference between IQA values of forged objects, denoted as M_f, and IQA values of original real objects, denoted as M_r. Specifically, the difference is expressed as a ratio, computed as $\frac{|M_f - M_r|}{M_r}$. A smaller difference value indicates that the forged IR objects are more realistic.

Table 3 presents the difference between IQA values obtained by various deep generative models and the IQA values of real objects. From Table 3, we can observe that the forged IR objects generated by our proposed DOCI-GAN exhibit the minimum difference in IQA values compared to the IQA values of real IR objects. The difference results in Table 3 underscore the superiority of DOCI-GAN in generating realistic IR objects.

4.5. Ablation study of multiscale erosion-based MSE loss

We designed the multiscale erosion-based MSE loss, $L_{multi-erosion}$, for our DOCI-GAN, which generates object images through inpainting missing regions within bounding boxes. Here, $L_{multi-erosion}$ plays an important role in ensuring consistency between the local backgrounds of inpainted object regions and the global background of the entire image. To validate our design, we conducted an ablation experiment in which DOCI-GAN was trained without $L_{multi-erosion}$.

We measured the difference between images generated by DOCI-GAN with and without $L_{multi-erosion}$ using IQA metrics as described in Section 4.3. Moreover, we examined the transitional region from the local background around the object to the global background from a regional perspective. Specifically, we cropped out 3-pixel-width regions inwards and outward from the bounding box. These two regions combined to form a ring, showing the transition from local background inside the bounding box to the surrounding background. Then, we computed the IQA metrics of this transitional region in the generated images.

Table 4 presents the IQA results for both whole images and transitional regions. We compared these IQA values of generated images to those of original real images, as in Table 3. Since IS and FID reflect image diversity, which is not the focus of $L_{multi-erosion}$, they are excluded here. From Table 4, we can find that with $L_{multi-erosion}$,

Table 3
Difference between IQA results of forged objects by different generative methods and IQA results of original objects in MCTIFD dataset. The minimal difference is highlighted in bold.

Method\measures	GMG	G_T	G_E	G_B	SAD	E_S	IS	FID
GAN vs. Original	41.96%	53.31%	54.43%	49.03%	37.85%	16.45%	8.64%	123.60
VAE-GAN vs. Original	37.18%	48.76%	45.86%	59.93%	29.76%	21.30%	11.82%	87.08
DDPM vs. Original	58.63%	88.16%	80.98%	95.87%	29.80%	210.34%	31.06%	158.98
BicycleGAN vs. Original	32.76%	34.72%	41.82%	27.70%	5.91%	65.13%	5.37%	99.69
Pix2Pix vs. Original	13.23%	7.60%	13.60%	22.88%	17.00%	23.58%	5.16%	35.45
DOCI-GAN vs. Original	**1.52%**	**4.38%**	**2.24%**	**14.26%**	**2.35%**	**0.63%**	**4.39%**	**18.93**

Table 4
IQA results of generated images and original images from both the whole image and transitional region views revealing the effectiveness of $L_{multi-erosion}$.

Method (whole image)\measures	GMG	G_T	G_E	G_B	SAD	E_S
Original	3.10	106.32	40.20	554.44	55.59	0.15
DOCI-GAN	3.17	99.42	40.16	428.97	61.23	0.16
DOCI-GAN wo $L_{multi-erosion}$	3.21	101.01	40.63	425.46	62.07	0.17
DOCI-GAN vs. Original	**2.17%**	6.49%	**0.10%**	**22.63%**	**10.15%**	**9.77%**
DOCI-GAN wo $L_{multi-erosion}$ vs. Original	3.42%	**4.99%**	1.08%	23.26%	11.66%	12.31%
Method (transitional region)\measures	GMG	G_T	G_E	G_B	SAD	E_S
Original	2.21	132.78	28.04	1418.07	33.47	0.15
DOCI-GAN	2.28	131.13	28.47	1351.82	38.20	0.16
DOCI-GAN wo $L_{multi-erosion}$	2.29	130.99	28.24	1340.83	38.94	0.17
DOCI-GAN vs. Original	**3.49%**	**1.24%**	1.55%	**4.67%**	**14.15%**	**10.44%**
DOCI-GAN wo $L_{multi-erosion}$ vs. Original	3.65%	1.35%	**0.70%**	5.45%	16.35%	12.98%

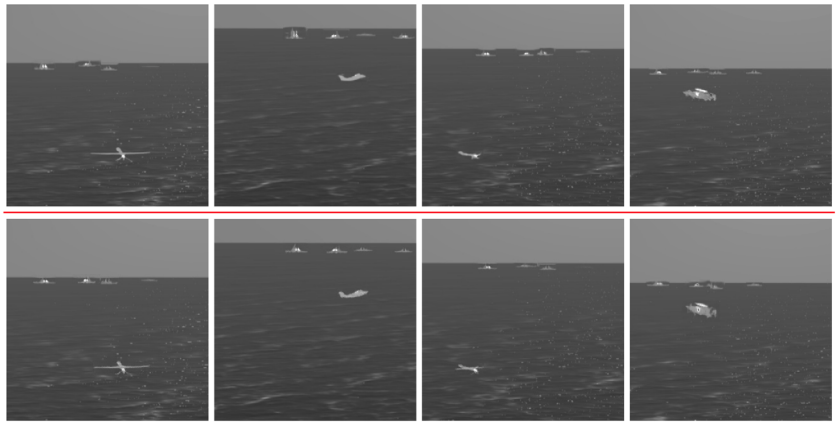

Fig. 7. Examples of images generated by DOCI-GAN without $L_{multi-erosion}$ (first row) or with it (second row).

DOCI-GAN achieves lower GMG, G_E, SAD and E_S values from both the whole image and transitional region views. This indicates that $L_{multi-erosion}$ successfully smoothens the intensity contrast between the local backgrounds of objects and the global background. Meanwhile, with $L_{multi-erosion}$, DOCI-GAN generates images with less difference from original images in terms of most metrics.

Fig. 7 compares generated images by DOCI-GAN with and without $L_{multi-erosion}$. It is evident that with the proposed $L_{multi-erosion}$, DOCI-GAN effectively inpaints objects and their local backgrounds in harmony with the surroundings, resulting in more plausible generated images overall.

4.6. IR multi-class object detection based on augmented dataset

To assess the practicality of DOCI-GAN for augmenting IR object image datasets, we employed DOCI-GAN to enlarge the original MCTIFD dataset. Fig. 8 displays examples of IR images with augmented objects generated by DOCI-GAN. Instances in the first column are original IR images. For the purpose of augmentation, we added bounding boxes with specified classes, for generating objects with greater diversity. As shown in Fig. 8, DOCI-GAN produces plausible and fine-grained IR object images. After 10 times augmentation, the augmented training dataset was utilized to train classical object detection models, including Faster RCNN [1], SSD [3], RetinaNet [36], to perform multi-class IR object detection. We compared the object detection accuracy of models trained respectively with original dataset and the augmented dataset by DOCI-GAN. The Mean Average Precision (mAP) metric was employed to evaluate object detection accuracy. To mitigate potential biases arising from unequal sample sizes between the original dataset and the augmented dataset, we duplicated the original IR images 10 times when the training object detection models with the original dataset.

Table 5 presents the mAP scores of Faster RCNN, SSD and RetinaNet on our MCTIFD dataset. Comparing models trained with the original dataset to those trained with augmented dataset, we observe a notable improvement in the accuracy of multi-class object detection. Specifically, when Faster RCNN, SDD and RetinaNet were trained with the augmented dataset, they demonstrated an increase in mAP scores for IR objects across most categories, resulting in an overall improvement across all categories.

To assess the significance of DOCI-GAN in enhancing object detection accuracy, we conducted a class-level hypothesis test. Given the variation in object numbers in our dataset, ranging from 40 to

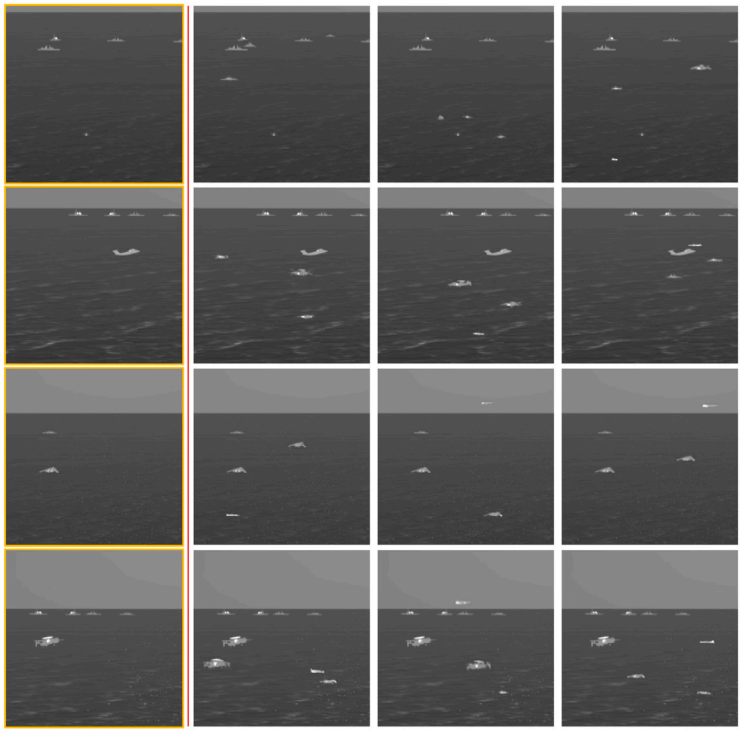

Fig. 8. Examples of object augmentation using our proposed DOCI-GAN for the MCTIFD dataset. Multi-class objects are generated with specified categories and positions.

Table 5
The mAP scores of different detection models for IR multi-class object detection. Models were trained using either the original MCTIFD dataset or the augmented dataset generated by DOCI-GAN.

AP\method	Faster RCNN (ori)	Faster RCNN (aug)	SSD (ori)	SSD (aug)	RetinaNet (ori)	RetinaNet (aug)
Helicopter	0.2820	0.5050	0.2560	0.6000	0.1030	0.2110
CruiseMissle	0.4110	0.3160	0.1810	0.2990	0.4070	0.2820
Ship	0.6140	0.6060	0.3450	0.3420	0.2690	0.4550
Surveillance	0.6290	0.7330	0.1990	0.2860	0.2040	0.4020
Cargo	0.7200	0.7000	0.2630	0.3010	0.3250	0.0610
Bomber	0.4510	0.4450	0.1770	0.2880	0.3020	0.3460
Big UAV	0.5180	0.5690	0.2160	0.2500	0.1740	0.0360
Revolve UAV	0.6120	0.7950	0.5940	0.6100	0.3060	0.6320
Fighter	0.3610	0.5690	0.2700	0.3470	0.1940	0.2730
mAP	0.5109	**0.5820**	0.2779	**0.3692**	0.2538	**0.2998**

Table 6
The paired-sample t-test for duplication accuracies.

H = 0: X = Y	X = ori, Y = aug (H/p-value)
Faster RCNN	1/1.2144e−10
SSD	1/1.97e−12
RetinaNet	1/2.22e−28

950, we aimed to mitigate the impact of class imbalance. We modified the vanilla t-test by duplicating category quantities. The accuracy duplication for the t-test is proposed as follows:

$$A_{duplication}(i) = \left[\underbrace{A(1), \ldots, A(i)}_{n(i)}\right], i \in C_0, \ldots, C_9 \tag{13}$$

In Eq. (13), $C_0, \ldots, C_9$ represent the nine classes for detection, $A(i)$ is the accuracy of the ith class, and $n(i)$ is the number of objects in the ith class. The t-test is performed on duplicated accuracies. The results of the paired-sample t-test are presented in Table 6. In Table 6, all null hypotheses ($H = 0$) are rejected with low p values. This indicates that the quantitative performance of detection models based on the augmented training set was significantly improved.

4.7. Further experiments on public IR single-class object datasets

We conducted experiments on public IR single-class object datasets, including OSU Thermal Pedestrian Database [37], small target datasets IRSTD-1k [38] and SIRST [18], to evaluate the data augmentation effectiveness of DOCI-GAN in various scenarios. The OSU dataset comprises 10 IR sequences of pedestrians. As neighboring frames within a sequence are similar, we selected frames with significant appearance differences. Subsequently, we collected 9 images for training and 111 images for testing for the OSU dataset experiment. Both the IRSTD-1k and SIRST datasets contain over 1000 annotated images. Since DOCI-GAN is designed for object detection with limited data, only 24 images were randomly chosen for training and augmentation inference when experimenting with the IRSTD-1k and SIRST datasets. After training of detection models with the augmented dataset, 201 original images from IRSTD-1k and 96 original images from the SIRST dataset were used for testing.

Fig. 9 showcases examples of augmented images generated by DOCI-GAN for the OSU, IRSTD-1k, and SIRST datasets. These images demonstrate that DOCI-GAN can produce plausible and clear IR pedestrian or small target images. Respectively based on the augmented OSU, IRSTD-1k and SIRST datasets, we trained general object detection models including Faster RCNN, SSD, RetinaNet, and small target detection

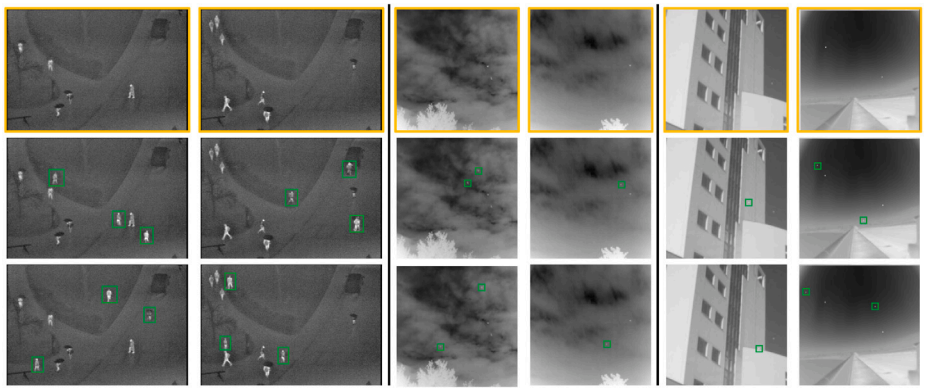

Fig. 9. Examples of object augmentation using our proposed DOCI-GAN for OSU dataset (left), IRSTD-1k dataset (middle) and SIRST dataset (right). The first row shows original images, while the second and third rows display the corresponding generated images, with green rectangles highlighting the generated objects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
IR single-class object detection accuracy using general object detection models trained on either original or augmented datasets.

Dataset	Faster RCNN			SSD			RetinaNet		
	mAP	P_d	F_a	mAP	P_d	F_a	mAP	P_d	F_a
OSU(ori)	0.471	0.514	0.239	0.403	0.400	0.228	0.313	0.488	0.232
OSU(aug)	0.536	0.584	0.171	0.426	0.462	0.199	0.396	0.499	0.222
IRSTD-1k(ori)	0.061	0.133	0.744	0.290	0.367	0.351	0.107	0.333	0.660
IRSTD-1k(aug)	0.340	0.449	0.211	0.338	0.459	0.285	0.203	0.428	0.484
SIRST(ori)	0.076	0.132	0.756	0.306	0.416	0.193	0.253	0.419	0.211
SIRST(aug)	0.327	0.408	0.187	0.361	0.449	0.193	0.294	0.442	0.272

Table 8
IR single-class object detection accuracy using small target detection models trained on either original or augmented datasets.

Dataset	ISNet			ACM		
	mAP	P_d	F_a	mAP	P_d	F_a
OSU(ori)	0.642	0.581	1.301e−1	0.557	0.432	1.885e−1
OSU(aug)	0.854	0.689	7.696e−2	0.671	0.684	9.105e−2
IRSTD-1k(ori)	0.445	0.744	1.072e−4	0.451	0.798	1.895e−4
IRSTD-1k(aug)	0.649	0.932	5.810e−5	0.586	0.795	7.640e−5
SIRST(ori)	0.352	0.830	2.898e−4	0.267	0.820	1.836e−4
SIRST(aug)	0.521	0.950	8.210e−5	0.549	0.980	5.910e−5

models including ISNet [38], ACM [18]. Tables 7 and 8 respectively present the detection accuracies of these models trained with original datasets or augmented datasets. In addition to mAP, we also employed Probability of Detection P_d and False-Alarm Rate F_a metrics for evaluating single-class object detection accuracy. The results from Tables 7 and 8 indicate that data augmentation by DOCI-GAN led to an increase in object detection accuracy of IR pedestrians and small targets for both general object detection models and specific small target detection models. This suggests the feasibility and effectiveness of DOCI-GAN in IR single-class object image applications.

Furthermore, we investigated the impact of augmented datasets generated by DOCI-GAN on the generalization ability of trained detection models. Specifically, for small target detection, detection models were trained on the augmented IRSTD-1k/SIRST datasets and then tested on images from the SIRST/IRSTD-1k dataset. Tables 9 and 10 demonstrate that models trained with augmented data achieved higher mAP and P_d scores on testing data drawn from a different distribution compared to models trained with original data. Regarding the F_a metric, the scores of some models decreased using augmented training dataset generated by DOCI-GAN when confronting the domain shift, but the models trained with augmented datasets still maintained competitive performance. These results indicate the enhancement in the generalization of detection models trained with augmented datasets compared to original datasets, further underscoring the advantages of utilizing DOCI-GAN for DA.

5. Conclusion

In this work, we propose DOCI-GAN, a generative adversarial networks model tailored for augmenting IR object images to enhance IR multi-class object detection with limited data. Inputted with bounding box masks and background images, DOCI-GAN, structured as an image inpainting framework, learns to infer IR objects across the given background. A text-to-image converter was developed to convert COCO-format object annotations into bounding box mask images, delineating the position and category of each object. In DOCI-GAN, the generator adopts a U-shaped architecture and incorporates the self-attention mechanism to ensure the generation of clear and reasonable objects. Additionally, the discriminator evaluates image authenticity at a local level, focusing on objects within each IR image, thereby guiding the generator to enhance the visual quality of the generated objects. Since DOCI-GAN realizes DA through image inpainting based on bounding box masks, noticeable intensity differences may arise between the local background within bounding boxes and the global background. To address this, we employ multi-scale morphological erosion to extend the local backgrounds of objects, subsequently optimizing the generator to minimize the distance between the multi-scale eroded results of the original and generated images, thereby ensuring consistency between local and global backgrounds.

We constructed the MCTIFD dataset, comprising annotated infrared multi-class object images, to test our DOCI-GAN. Experimental results indicated that among various deep generative models, DOCI-GAN yielded the highest-quality IR object images. The synthetic images generated by DOCI-GAN exhibited distinct advantages, having not only IR multi-class objects with sharp boundaries and intricate details but also realistic backgrounds with minimal noise.

We applied the proposed DOCI-GAN to augment the MCTIFD and extended its validation to single-class object datasets, including the OSU pedestrian dataset, IRSTD-1k and SIRST small target datasets. Our experiments confirmed that DOCI-GAN effectively performed data augmentation for diverse IR object detection datasets. The augmented datasets led to a substantial increase in the accuracy of various deep learning detection models, benefiting both IR multi-class object detection and single-class object detection tasks. These results underscore the significance of DOCI-GAN in addressing multiple IR object detection tasks with limited training data.

One notable limitation of our work is its inability to augment certain IR object datasets, such as IR street scene datasets. These datasets often feature large areas occupied by vehicles or pedestrians in an image,

Table 9
IR small target detection accuracy using general detection models when trained on one dataset (either the original or the augmented) and tested on another dataset.

Traing data - Testing data	Faster RCNN			SSD			RetinaNet		
	mAP	P_d	F_a	mAP	P_d	F_a	mAP	P_d	F_a
IRSTD-1k(ori) - SIRST	0.109	0.224	0.576	0.287	0.371	0.233	0.162	0.341	0.447
IRSTD-1k(aug) - SIRST	0.500	0.571	0.067	0.383	0.488	0.166	0.334	0.463	0.214
SIRST(ori) - IRSTD-1k	0.031	0.069	0.865	0.206	0.346	0.429	0.094	0.271	0.681
SIRST(aug) - IRSTD-1k	0.208	0.296	0.448	0.234	0.385	0.427	0.109	0.347	0.682

Table 10
IR small target detection accuracy using small target detection models when trained on one dataset (either the original or the augmented) and tested on another dataset.

Traing data - Testing data	ISNet			ACM		
	mAP	P_d	F_a	mAP	P_d	F_a
IRSTD-1k(ori) - SIRST	0.480	0.890	5.110e−5	0.265	0.830	2.305e−4
IRSTD-1k(aug) - SIRST	0.510	0.910	3.080e−5	0.289	0.890	4.708e−4
SIRST(ori) - IRSTD-1k	0.418	0.721	1.871e−4	0.335	0.724	1.276e−4
SIRST(aug) - IRSTD-1k	0.574	0.886	9.710e−5	0.449	0.912	1.771e−4

with objects overlapping in an image or captured only partially. Since DOCI-GAN is tailored to inpaint whole objects, it is not suitable for such IR object datasets. Additionally, while DOCI-GAN can enhance detection models with limited training data, its impact may be limited in cases where training data is already sufficient. In future research, we plan to explore alternative deep generative methods capable of augmenting both small and large IR objects to address these limitations.

CRediT authorship contribution statement

Peng Wang: Writing – original draft, Validation, Project administration, Methodology, Formal analysis, Conceptualization. **Zhe Ma:** Writing – review & editing, Conceptualization. **Bo Dong:** Visualization, Validation. **Xiuhua Liu:** Methodology, Investigation. **Jishiyu Ding:** Formal analysis, Conceptualization. **Kewu Sun:** Methodology, Investigation. **Ying Chen:** Writing – original draft, Visualization, Resources, Methodology, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We are organizing data and code now. We will release them on the github https://github.com/RonaldoPeng/MCTIFD later.

Acknowledgments

This work is supported by the National Natural Science Foundation of China [grant number 62102377], Young Elite Scientist Sponsorship Program by CAST (YESS) [grant number 2021QNRC001]. Ying Chen would like to thank the China Scholarship Council (CSC) for the financial support [No. 202206020082].

References

[1] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149.
[2] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 21–37.
[4] J. Redmon, A. Farhadi, YOLO9000: Better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6517–6525.
[5] N. Bustos, M. Mashhadi, S.K. Lai-Yuen, S. Sarkar, T.K. Das, A systematic literature review on object detection using near infrared and thermal images, Neurocomputing 560 (2023) 126804.
[6] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, Y. Akbari, Image inpainting: a review, Neural Process. Lett. 51 (2) (2020) 2007–2028.
[7] Y. Chen, H. Zhang, L. Liu, X. Chen, Q. Zhang, K. Yang, R. Xia, J. Xie, Research on image inpainting algorithm of improved GAN based on two-discriminations networks, Appl. Intell. 51 (2021) 3460–3474.
[8] X. Zhang, X. Wang, C. Shi, Z. Yan, X. Li, B. Kong, S. Lyu, B. Zhu, J. Lv, Y. Yin, Q. Song, X. Wu, I. Mumtaz, De-gan: Domain embedded gan for high quality face image inpainting, Pattern Recognit. 124 (2022) 108415.
[9] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 6840–6851.
[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.
[11] G. Zhang, J. Ji, Y. Zhang, M. Yu, T. Jaakkola, S. Chang, Towards coherent image inpainting using denoising diffusion implicit models, in: Proceedings of International Conference on Machine Learning, 2023, pp. 41164–41193.
[12] C. Corneanu, R. Gadde, A.M. Martinez, LatentPaint: Image inpainting in latent space with diffusion models, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 4334–4343.
[13] M. Zhang, C. He, J. Zhang, Y. Yang, X. Peng, J. Guo, SAR-to-optical image translation via neural partial differential equations, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022, pp. 1644–1650.
[14] B. Bosquet, D. Cores, L. Seidenari, V.M. Brea, M. Mucientes, A.D. Bimbo, A full data augmentation pipeline for small object detection based on generative adversarial networks, Pattern Recognit. 133 (2023) 108998.
[15] O. Bailo, D.S. Ham, Y. Min Shin, Red blood cell image generation for data augmentation using conditional generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 1039–1048.
[16] R. Kou, C. Wang, Z. Peng, Z. Zhao, Y. Chen, J. Han, F. Huang, Y. Yu, Q. Fu, Infrared small target segmentation networks: A survey, Pattern Recognit. 143 (2023) 109788.
[17] M. Zhang, K. Yue, J. Zhang, Y. Li, X. Gao, Exploring feature compensation and cross-level correlation for infrared small target detection, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1857–1865.
[18] Y. Dai, Y. Wu, F. Zhou, K. Barnard, Asymmetric contextual modulation for infrared small target detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 950–959.
[19] M. Zhang, H. Bai, J. Zhang, R. Zhang, C. Wang, J. Guo, X. Gao, Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1730–1738.
[20] B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, Y. Guo, Dense nested attention network for infrared small target detection, IEEE Trans. Image Process. (32) (2022) 1745–1758.
[21] M. Zhang, R. Zhang, J. Zhang, J. Guo, Y. Li, X. Gao, Dim2Clear network for infrared small target detection, IEEE Trans. Geosci. Remote Sens. (61) (2023) 1–14.
[22] S. Li, Y.J. Li, Y. Li, M. Li, X. Xu, Yolo-firi: Improved yolov5 for infrared image object detection, IEEE Access 9 (2021) 141861-141875.
[23] X. Dai, Y. Xue, X. Wei, TIRNet: Object detection in thermal infrared images for autonomous driving, Appl. Intell. 51 (2021) 1244–1261.
[24] C. Jiang, H. Ren, X. Ye, J. Zhu, H. Zeng, Y. Nan, M. Sun, X. Ren, H. Huo, Object detection from UAV thermal infrared images and videos using YOLO models, Int. J. Appl. Earth Obs. Geoinf. 112 (2022) 102912.
[25] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[26] P. Wang, X. Bai, Thermal infrared pedestrian segmentation based on conditional gan, IEEE Trans. Image Process. 28 (12) (2019) 6007–6021.

[27] H. Xiang, Q. Zou, M.A. Nawaz, X. Huang, F. Zhang, H. Yu, Deep learning for image inpainting: A survey, Pattern Recognit. 134 (2023) 109046.
[28] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[29] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick, Microsoft COCO: common objects in context, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 740–755.
[30] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of wasserstein GANs, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 5769–5779.
[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing System, 2014, pp. 2671–2680.
[32] A.B.L. Larsen, S.K. Sønderby, H. Larochelle, O. Winther, Autoencoding beyond pixels using a learned similarity metric, in: Proceedings of International Conference on Machine Learning, 2016, pp. 1558–1566.
[33] J. Zhu, R. Zhang, D. Pathak, T. Darrell, A.A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, in: Advances in Neural Information Processing System, 2017, pp. 465–476.
[34] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: Advances in Neural Information Processing Systems, Vol. 29, 2016, pp. 2234–2242.
[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 6629–6640.
[36] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[37] J. Davis, M. Keck, A two-stage approach to person detection in thermal imagery, in: IEEE Workshop on Applications of Computer Vision, 2005, pp. 364–369.
[38] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, J. Guo, ISNet: Shape matters for infrared small target detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 877–886.

Peng Wang received B.S., M.S. and Ph.D. degrees from the Beihang University in 2012, 2016 and 2020, respectively. He is an engineer in the Intelligent Science & Technology Academy of CASIC. He is involved in thermal infrared image analysis, few shot learning, object segmentation and GAN theory.

Zhe Ma received Ph.D. degree from Tsinghua University, Beijing, China in 2009. He is currently a research fellow in the Intelligent Science & Technology Academy of CASIC, with major research interests on intelligent systems.

Bo Dong received B.S. and M.S. degrees from the Beihang University(BUAA) in 2014 and 2017, respectively. He is an engineer in the Intelligent Science & Technology Academy of CASIC, and a Ph.D. student in Harbin Institute of Technology. He is involved in multi-vehicle mission planning and multi-agent theory.

Xiuhua Liu received B.E. degree from Beijing Forestry University in 2016 and Ph.D. degree from Peking University in 2021, respectively. She is an engineer in the Intelligent Science & Technology Academy of CASIC. Her research interests include wearable robotics, few shot learning and swarm intelligence.

Jishiyu Ding received the Ph.D. degree from Tsinghua University, Beijing, China in 2020. He is currently an engineer in the Intelligent Science & Technology Academy of CASIC with major research interests on multiagent reinforcement learning.

Kewu Sun received M.S. degree from Tsinghua University, Beijing, China in 2012. She is currently a senior engineer in the Intelligent Science & Technology Academy of CASIC. Her research interests include multi-agent reinforcement learning and system of systems (SoS).

Ying Chen received B.S. and M.S. degrees from Beihang University in 2019 and 2022, respectively. She is currently a Ph.D. student in the Institute for Stroke and Dementia Research, University Hospital, Ludwig Maximilians University Munich. She is involved in medical image analysis, infrared image analysis, object segmentation.