Miss Detection vs. False Alarm: Adversarial Learning for Small Object Segmentation in Infrared Images

Huan Wang
Nanjing University of Science &
Technology, Nanjing, P.R.China

Luping Zhou
University of Sydney
Sydney, Australia

Lei Wang*
University of Wollongong
Wollongong, Australia

Abstract

A key challenge of infrared small object segmentation (ISOS) is to balance miss detection (MD) and false alarm (FA). This usually needs "opposite" strategies to suppress the two terms and has not been well resolved in the literature. In this paper, we propose a deep adversarial learning framework to improve this situation. Departing from the tradition of relying on a single objective to jointly reduce both MD and FA, we decompose this difficult task into two sub-tasks handled by two models trained adversarially, with each focusing on reducing either MD or FA. Such a new design brings forth at least three advantages. First, as each model focuses on a relatively simpler sub-task, the overall difficulty of ISOS is somehow decreased. Second, the adversarial training of the two models naturally produces a delicate balance of MD and FA, and low rates for both MD and FA could be achieved at Nash equilibrium. Third, this MD-FA detachment gives us more flexibility to develop specific models dedicated to each sub-task. To realize the above design, we propose a conditional Generative Adversarial Network comprising of two generators and one discriminator. Each generator strives for one sub-task, while the discriminator differentiates the three segmentation results from the two generators and the ground truth. Moreover, in order to better serve the sub-tasks, the two generators, based on context aggregation networks, utilize different size of receptive fields, providing both local and global views of objects for segmentation. As verified on multiple infrared image data sets, our method consistently achieves better segmentation than many state-of-the-art ISOS methods.

1. Introduction

Segmenting small objects or targets in infrared images is an important computer vision task. It plays a fundamental role in many practical applications such as defect inspection [22, 26], organ segmentation [24], cell counting [1], maritime surveillance [25] and early warning systems [9, 13], to name but a few. With respect to commonplace object segmentation, infrared small object segmentation (ISOS) has its special characteristics (see some examples in Fig. 1). First, due to either their sizes or the long distances from infrared sensors, the objects usually appear to be very small in infrared images, with the extreme case of one pixel only. Second, the infrared radiation energy decays markedly over distances, making the objects appear to be extremely dim. Consequently, they are easily submerged in background clutters and sensor noises. Third, different from dense small object instance segmentation [18, 15, 2], the objects in ISOS are usually very *sparse*, e.g., containing a single instance only. This leads to a severe imbalance between the object area and the background area. These three factors significantly complicate ISOS.

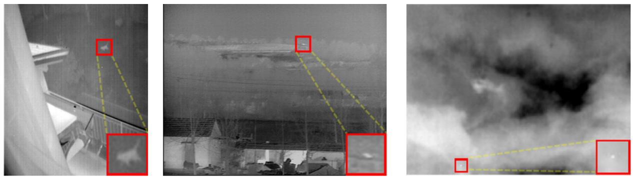

Figure 1. Illustration of ISOS examples with the objects or targets indicated by the red bounding boxes. A close-up is displayed at the bottom right corner of each example. Left: a cat whose silhouette is recognizable in a backyard background; Middle: a car which is hard to tell in a mountain scene; Right: a dim target in the size of three pixels submerged in a cloud background.

For many infrared image related applications, high quality segmentation is essential for the precise measurement, localization and recognition tasks in the sequel. The segmentation errors in ISOS boil down to miss detection (MD in short, i.e., the pixels of an object are wrongly segmented into the background) and false alarm (FA in short, i.e., the pixels of the background are wrongly segmented into the object). An optimal segmentation should minimize both MD and FA. However, reducing MD and FA usually involves "opposite" strategies, e.g., the former prefers a low threshold while the latter prefers a high one on the confidence maps used in many ISOS methods [12, 29, 4] and

*Lei Wang is the corresponding author.(leiw@uow.edu.au)

2380-7504/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCV.2019.00860

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

they are difficult to balance. Conventional signal processing based ISOS methods apply an adaptive threshold on objects' confidence maps to balance MD and FA. These methods do not involve any feature learning and cannot effectively handle the complexity in real scenarios. Recently, deep learning based methods have been developed for small object segmentation [15, 18, 28]. Considering the special characteristics of that task, they remove the common CNN network components that are not suitable and design special structures to cater for small object segmentation. However, the scenarios addressed by these methods are still significantly different from the ISOS task focused in this paper, e.g, the work on remote sensing imagery [15] is proposed for *dense* small object segmentation. More importantly, all of these deep learning methods crucially rely on a single objective to minimize the overall segmentation error, instead of separating MD and FA as in our work. Given the complicated nature of the ISOS task, these methods still perform less satisfactorily, as demonstrated in the experiment later.

This work is motivated by the following thought that we gain from ISOS tasks. Due to the small size of the objects to segment and their sparsity in an infrared image, producing high quality segmentation requires a delicate balance of MD and FA. However, such a balance may not be sufficiently achieved by merely applying a threshold or solely employing a loss function in the form of a weighted linear combination of MD and FA. Inspired by the recent success of adversarial learning, we realize that a better approach may be to let the two "opposite" tasks, minimizing MD and minimizing FA, compete with each other, with the expectation that the delicate balance could naturally arise when such a competition achieves its stable state.

Following the above idea, we propose a deep adversarial learning framework to improve the performance of ISOS. In this framework, an ISOS task is decomposed into two sub-tasks, i.e., minimizing MD and minimizing FA. Two deep neural networks are constructed to focus on each of the two tasks, respectively. The two networks play the role of generator and each outputs a segmentation result. To make the two segmentation results align with the ground truth segmentation result, a discriminator network is constructed to classify the above three results. In this way, the two generators work in an interesting "competitive and cooperative" manner. By competition, they strive to maximally segment the pixels into the objects *or* the background, respectively. By cooperation, they negotiate (i.e., balance) with each other to both converge towards the ground truth segmentation in order to fool the discriminator. Given a test image, the output of either generator (or their average) will be the segmentation result. The whole framework can be readily implemented by expanding a conditional Generative Adversarial Network, as illustrated in Fig. 2.

The contributions of this paper can be summarized as follows. **First**, we propose a novel framework for infrared small object segmentation, by employing the adversarial learning paradigm. It removes the burden of explicitly balancing MD and FA and can achieve a delicate balance in an implicit and natural manner; **Second**, taking advantage of the separability of MD and FA minimization, an ISOS task is decomposed to two individual and simpler sub-tasks. Compared with the existing methods that use a single network for segmentation, our approach could reduce the overall difficulty of model and network design. **Third**, an immediate advantage resulted from the above separation is the extra flexibility to develop the model that best suits a subtask, and this has been demonstrated in our work as follows. We find that in ISOS, the segmentation of objects prefers local visual information, while the suppression of false alarm benefits from global visual information. To meet this requirement, we utilize different size of receptive fields in the two generators, via context aggregation networks [15]. Without the separation of the two sub-tasks, implementing this special setting could be awkward if not impossible. **Last**, we compare our method with relevant state-of-the-art small object segmentation methods on multiple infrared image datasets. The results well demonstrate the superiority of the proposed method and its interesting properties.

2. Related Work

2.1. Generic Small Object Segmentation

Recently, deep network models have been successfully applied to semantic segmentation, e.g., the Fully Convolutional Network (FCN) [21] and its numerous variants that dominate the literature. Unfortunately, directly applying generic FCN models to ISOS will not work effectively because they do not sufficiently consider the special characteristics of small objects in that task. For example, consecutive max-pooling may suppress or even eliminate the important features of small objects [15]. Also, due to the considerable disproportion between the object area and the background area, these models could be easily confused by the background regions containing complicated content [28]. To resolve these issues, some deep learning models have been proposed to adapt FCN to small object segmentation (but not to ISOS though). For example, the Front-end-module [15] (denoted as Front) replaces the max-pooling layers with dilated convolutional layers for large receptive fields; the recurrent saliency transformation network [28] (denoted as FCN-RSTN) adopts a multi-stage strategy for coarse-to-fine small organ segmentation; and the FCN in [18] (denoted as FCN-MFB) weighs its loss function with median frequency to balance small classes. These models work well for their targeted scenarios, e.g., Front and FCN-MFB for dense small object segmentation in remote sensing images, and FCN-RSTN for CT organ seg-

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

mentation with relatively fixed locations. However, those scenarios are still significantly different from ISOS where the objects to segment are tiny, dim and sparse with unpredictable locations. Moreover, all of the above models attempt to minimize the overall segmentation errors by relying on a single objective. As we have argued previously, achieving this in ISOS tasks could be awkward due to their more complicated nature.

2.2. Infrared small object segmentation (ISOS)

For infrared images, many ISOS methods in the literature are rooted in detection frameworks using a segmentation-before-detection strategy, and most of them are based on traditional image processing techniques. A common pipeline is to apply image filtering [12, 29, 3], contrast and saliency detection [4, 9, 11] or low-rank recovery [13, 8, 7] to suppress the background and enhance the objects to obtain a confidence map. After that, an adaptive thresholding is conducted on this map to segment the objects out. These methods perform well for relatively simple backgrounds, but tend to fail for complex ones, since they do not involve any feature learning and therefore are not capable enough to handle varied real scenarios.

Sporadically, learning-based methods [20, 6] are also used for ISOS. The work in [20] densely samples regions using a sliding window and classifies the region centroids into background or object. It uses a CNN model with the features extracted from the region proposals for segmentation. The work in [6] uses a two-stage strategy, where object regions are proposed in the first stage and verified in the second stage via an SVM classifier. The relatively simple learning strategies used in these methods are not sufficient to handle the real complexity in ISOS and they only produce mediocre performance, as shown in our experiment later.

2.3. Conditional GAN

Our model utilizes the conditional Generative Adversarial Network (GAN). GAN [14] has recently achieved great success in numerous visual recognition tasks [23]. Briefly, different from common deep models, a basic GAN model consists of two players: a generator G and a discriminator D. These two players compete in a zero-sum game, in which G aims to produce a realistic image given an input random vector, while D attempts to distinguish the fake images generated by G from the real ones. Such an adversarial competition progressively boosts the performance of both G and D, until a Nash equilibrium is reached. Conditional GAN (cGAN) [17] extends GAN by introducing an additional conditional variable to both G and D. For example, in many visual tasks, this conditional variable usually corresponds to an original input image as the reference. GANs have also been used for image segmentation [17], where the generator attempts to produce segmentation labels close to the ground truth as much as possible to fool the discriminator. Our work utilizes the basic cGAN framework, but makes substantial changes to accommodate the proposed separation of MD and FA minimization tasks.

3. The proposed model

This section begins with an overview of the proposed model. After that, the loss functions for the generators, discriminator, and adversarial learning are presented. Following that, we describe how each generator is individually designed to better handle the characteristics of ISOS, with the key implementation details provided.

3.1. Model overview

As illustrated in Fig. 2, the proposed model consists of generator and discriminator components as in cGAN. However, different from cGAN, it has two generators, G_1 and G_2, and one discriminator D. Each of the generators maps an input image $\mathbf{I}$ to another image $\mathbf{S}$ showing the segmentation result, subject to the minimization of MD or FA. Formally, this can be represented as $G_1(\mathbf{I}) \rightarrow \mathbf{S}_1$ and $G_2(\mathbf{I}) \rightarrow \mathbf{S}_2$, where $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the segmentation results. To carry out adversarial learning, the discriminator is designed to distinguish three segmentation results, i.e. $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_0$, where $\mathbf{S}_0$ denotes the ground truth segmentation ("1" for objects and "0" for the background).

Intuitively, we could individually train the two generator networks and then fuse their segmentation results after they are well trained. For example, this strategy has been seen in the literature when addressing shadow segmentation [23]. However, this fusion-after-training strategy blocks the sharing of information between the training of the two generators, and this will result in inferior segmentation. This issue is well avoided in the proposed model, which jointly trains the two generators via the cGAN framework. Specifically, it leverages the discriminator D as a medium to connect G_1 and G_2, so that information can flow between them. This information exchange could in turn boost the ability of G_1 (originally designed for minimizing MD) in reducing FA, and boost the ability of G_2 (originally designed for minimizing FA) in reducing MD. Furthermore, both generators receive strong supervision signals from D, as the adversarial mechanism forces them to converge towards the ground truth in order to fool D. Through this process, the two generators will end up with producing consistent segmentation and becoming highly similar to the ground truth. Fig. 3 gives an example of the output evolution of G_1 and G_2.

After training the whole model, either generator can be applied to a test image to produce segmentation result, since the two generators have been trained to converge through the adversarial learning process. In practice, for the sake of robustness, we apply both generators and use the average of their outputs as the final segmentation result.

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

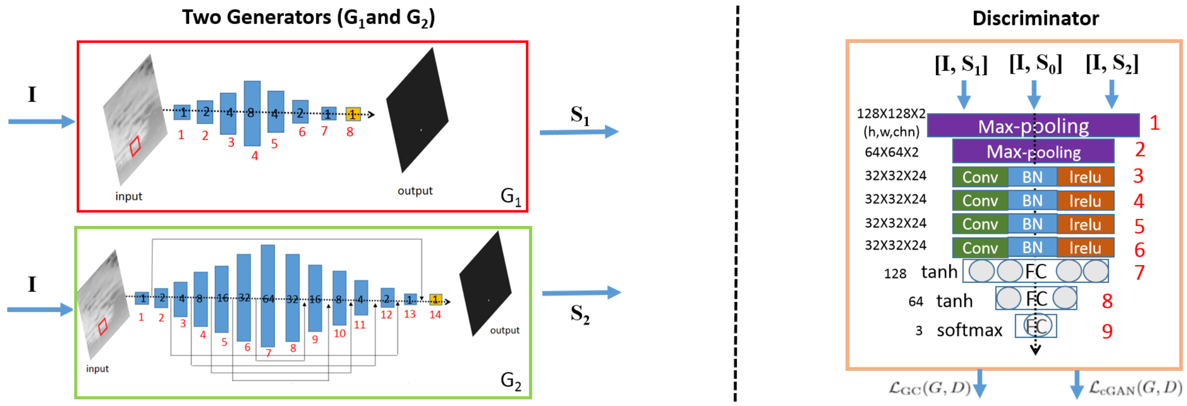

Figure 2. System overview and network architectures. On the left are the two generators. The generator one (i.e., G_1) is illustrated in the red box, while the generator two (i.e., G_2) is shown in the green one. On the right is the discriminator plotted in the orange box. Layers are indexed in red numbers. For the generators, the black number within each layer is the dilation factor used by that layer. For the discriminator, the width, height and channel number of feature maps in each layer are given besides that layer.

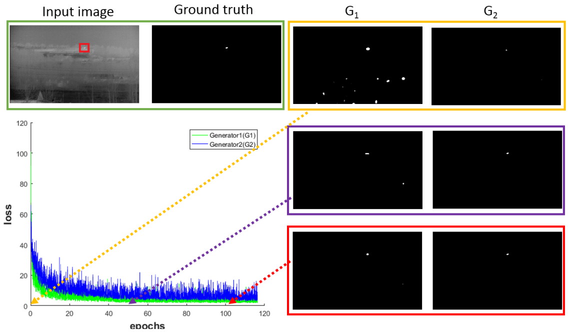

Figure 3. The output evolution of G_1 and G_2 during a training process. The top left shows an input image and its corresponding ground truth. The bottom left shows the loss curves of G_1 and G_2. The right shows the outputs of G_1 and G_2 at three different epochs.

3.2. Loss formulation in the proposed model

The objective of the proposed model consists of three parts: the adversarial loss, a generator consistency loss, and a data loss treating MD and FA[1], described as follows.

Adversarial loss. Different from the one in the common cGAN, this loss now consists of three terms due to the use of two generators. It can be expressed as

$$\begin{aligned}\mathcal{L}_{\mathrm{cGAN}}(G, D) &= \mathbb{E}_{\mathbf{I},\mathbf{S}_0}\left[\log D(\mathbf{I}, \mathbf{S}_0)\right] \\ &+ \mathbb{E}_{\mathbf{I}}\left[\log(1 - D(\mathbf{I}, G_1(\mathbf{I})))\right] \\ &+ \mathbb{E}_{\mathbf{I}}\left[\log(1 - D(\mathbf{I}, G_2(\mathbf{I})))\right],\end{aligned} \tag{1}$$

where the last two terms correspond to two generators. Minimizing this objective with respect to the network weights of G_1 and G_2 encourages their outputs (i.e., $\mathbf{S}_1$ and $\mathbf{S}_2$ previously defined) to become similar as the ground truth $\mathbf{S}_0$. Maximizing it with respect to the network weights of D enhances its discrimination in the three segmentation results.

Generator consistency loss. The above adversarial loss forces $\mathbf{S}_1$ and $\mathbf{S}_2$ to approach $\mathbf{S}_0$. However, this is insufficient. As we observe, $\mathbf{S}_1$ and $\mathbf{S}_2$ could move towards $\mathbf{S}_0$ in their own ways. As a result, their discrepancy could still remain significant after training. This cannot effectively force them to compete on every pixel to strike a balance between MD and FA. To address this issue, we impose an extra content consistency loss to bind the two generators tighter (i.e., enhance the information flow between them). This loss is defined as the L_2 norm of the difference between the convolutional feature maps in the discriminator D with respect to the input pairs $(\mathbf{I}, \mathbf{S}_1)$ and $(\mathbf{I}, \mathbf{S}_2)$ as

$$\mathcal{L}_{\mathrm{GC}}(G, D) = \frac{1}{w \cdot h \cdot d}\|\phi(\mathbf{I}, \mathbf{S}_1) - \phi(\mathbf{I}, \mathbf{S}_2)\|_2^2 \tag{2}$$

where $\phi(\cdot)$ denotes the corresponding feature mapping, and w, h and d are the three dimensions of the convolutional feature maps.

Data loss. In image-to-image conditional GAN [17], an L_1 or L_2 loss is commonly used to indicate the difference between the prediction and the ground truth. However, simply using an L_1 or L_2 loss only accounts for pixel-level discrepancy, while ignoring the measurement of MD or FA. To handle this, we instead define the data loss as follows. The losses for two generators G_1 and G_2 are, respectively,

$$\begin{aligned}\mathcal{L}_{MF1}(G_1, D) &= \frac{1}{n}\sum_{i=1}^{n}(\lambda_1 MD_{1i} + FA_{1i}) \\ \mathcal{L}_{MF2}(G_2, D) &= \frac{1}{n}\sum_{i=1}^{n}(MD_{2i} + \lambda_2 FA_{2i}),\end{aligned} \tag{3}$$

[1] As will be clarified shortly, different from the existing methods that significantly reply on a single objective jointly considering MD and FA, this data loss only plays an auxiliary role in training the proposed framework.

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

where MD_{1i} and FA_{1i} are the miss detection and false alarm rates computed based on $\mathbf{S}_1$ (i.e., the segmentation result produced by G_1) for the i-th image in the training set containing n images. Similarly, MD_{2i} and FA_{2i} are computed based on $\mathbf{S}_2$.[2] λ_1 and λ_2 are the parameters to weigh MD and FA, so that G_1 focuses on MD and G_2 focuses on FA. The final data loss is

$$\mathcal{L}_{MF}(G_1, G_2, D) = \mathcal{L}_{MF1}(G_1, D) + \mathcal{L}_{MF2}(G_2, D). \quad (4)$$

It is worth **clarifying** that this data loss does not contradict our previous claim that the proposed model does not crucially reply on a single objective to balance MD and FA. We highlight that regularizing MD with a small FA term in G_1 (or vice versa in G_2) is to achieve a good initialization for training. It helps the two generators quickly enter their roles and speed up the convergence of training process. The proposed method is insensitive to the specific values of λ_1 and λ_2, and they can vary in a relatively large range, as demonstrated in our experiment. This is significantly different from the traditional use of this kind of loss in the literature, in which how to set λ_1 and λ_2 has a direct and substantial impact on the segmentation performance.

Now, the complete objective of our proposed model is

$$(G_1^*, G_2^*, D^*) = \arg\min_{G_1, G_2}\max_{D}\left(\mathcal{L}_{\mathrm{GC}} + \alpha_1\mathcal{L}_{MF} + \alpha_2\mathcal{L}_{\mathrm{cGAN}}\right), \quad (5)$$

where α_1 and α_2 are the algorithmic coefficients. The setting will be provided in the experimental part.

3.3. Network Architecture

Since the reductions of MD and FA are conducted by different generators in our model, they can enjoy different network architectures or parameters that better suit the specific objective. As observed, the detection of small objects may prefer local receptive fields to preserve the footprints of the objects, while the suppression of FA seems to need the context clues provided by more global receptive fields. In our model we build our generators using Context Aggregation Network (CAN)[5] and assign different receptive fields to different generator.

Specifically, as shown in Fig. 2, to form the backbone of G_1 or G_2, we concatenate two CANs back to back, where the first one has an exponentially (i.e., at the power of 2) increasing dilation factor from 1 to the maximum M_{DF} and the second one has an exponentially (i.e., at the power of 2) decreasing dilation factor from M_{DF} to 1. The generator G_1 prefers local receptive fields to reduce MD and thus sets $M_{DF} = 8$, while the generator G_2 prefers global receptive fields to suppress FA and thus sets $M_{DF} = 64$. The total receptive field is 31×31 for G_1 and 257×257 for G_2. Compared with G_1, in addition to having more layers, G_2 also uses skip connections to connect layers with the same dilation factors to mitigate the gradient vanishing problem when the model goes deeper.

The discriminator is a CNN network for classification. It consists of two max-pooling layers (Layers 1 and 2) to downsample the input images, four layers (Layers $3 \sim 6$) of network modules consisting of conv-layer + BN (batch normalisation) + leaky-ReLU activation, two fully connected layers (Layers 7 and 8) and the output layer with softmax activation to classify the source of the input images (either from G_1, G_2 or the ground truth). The size of feature maps for each layer is given in Fig. 2.

4. Experimental Result

This experiment will compare the proposed method with the related state-of-the-art small object detection methods and ISOS methods. Also, it will conduct ablation study to clearly show its advantage and appealing properties.

4.1. Datasets

Lacking public benchmark datasets for ISOS tasks, we collect real infrared images and generate synthetic ones to validate the proposed model. The real infrared images come from two bespoke datasets containing small objects, denoted as "AllSeqs" and "Single", respectively. The dataset "AllSeqs" contains 11 real infrared sequences with 2098 frames in total, and the dataset "Single" contains 100 real individual infrared images with different small objects. The detailed description about the real infrared image datasets is given in Table 1, and example images are given in Fig. 4. In addition, to augment the datasets, synthetic infrared images with small objects are generated. For this purpose, we collect infrared high-resolution natural scene images from the Internet and crop different regions from these images to form different backgrounds; then small target objects either separated from the real infrared images or synthesised using the two dimensional Gaussian function in [20] are overlaid on the attained backgrounds to form new images. Detailed procedure to produce the synthetic dataset is given in the supplement material. Both the real and synthetic datasets used in this experiment will be released for public use.

Our experiment is conducted under two training-test configurations. In Configuration I, the "Single" data set is used as the test set, while the "AllSeqs" dataset and the synthetic images are used as the training set. In Configuration II, the "AllSeqs" dataset is used as the test set , while the "Single" dataset and the synthetic images are used as the training set. Note that under either configuration, the backgrounds in the test images are not seen in the training images, increasing the difficulty of the tasks and ensuring an accurate evaluation of generalization capability of each

[2] The calculation of MD and FA is simple. Given a binary segmentation result $\mathbf{S}$ and the ground truth $\mathbf{S}_0$, $MD = \|(\mathbf{S} - \mathbf{S}_0) \otimes \mathbf{S}_0)\|_2^2$ and $FA = \|(\mathbf{S} - \mathbf{S}_0) \otimes (1 - \mathbf{S}_0)\|_2^2$, where the L_2 norm is matrix-based and the operator $\otimes$ denotes the element-wise multiplication.

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

method. To increase the number of the training samples, we randomly sample 128×128 images patches from the original images as our input, which adds up to $10,000$ patches for training under each configuration.

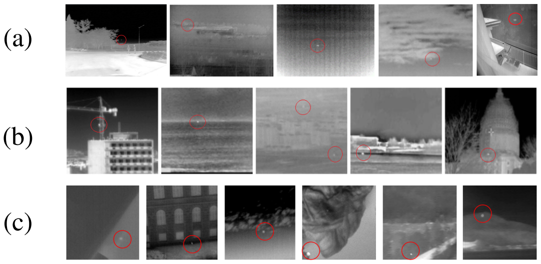

Figure 4. Representative real images from (a) ”Allseqs” and (b) ”Single”, and (c) synthesised images used for model training. Small objects to segment are indicated by red circles.

Table 1. Datasets: No. 1∼11 are the eleven sequences in “AllSeqs” dataset, and No. 12 is the single frame image dataset “Single”.

No.	Name	Size	Frames/images
1	Cannonball	352×288	30
2	Car	344×256	116
3	Plane	320×240	298
4	Bird	640×480	232
5	Cat	216×256	292
6	Rockets	320×240	242
7	Drone	384×288	396
8	Target1	480×360	361
9	Target2	256×200	30
10	Target3	352×240	50
11	Target4	384×288	51
12	Single-frame image set	Min:173×98, Max:407×305	100

4.2. Experimental Settings

The experiment is conducted on a computer with 2.50GHz CPU, 8GB RAM and GeForce GTX 1080ti GPU. Our model is implemented by Python and Tensorflow. We use Adam algorithm for optimization. Key parameters are empirically set as $\alpha_1 = 100$, $\alpha_2 = 10$, and they are uniformly applied to all experiments. The mini-batch size is set to be 10. The learning rate is set to be 10^{-4} for the two generators and 10^{-5} for the discriminator. The weights of the generators are initialized using the identity initialization technique [5], while the weights of the discriminator is initialized by following the literature [16]. All models are trained from the scratch, and the whole training process terminates in 30,000 iterations, which equals 30 epochs.

4.3. Methods in comparison

This experiment compares the proposed model (denoted as MDvsFA-cGAN) with two groups of related methods: generic small object segmentation and the ISOS methods.

The state-of-the-art generic small object segmentation methods are based on deep learning, and they are not originally designed to solve ISOS problems. We choose the most related models for our comparison, including the methods of Front [15] and FCN-MFB [18] originally for remote sensing, and FCN-RSTN[28] for small organs in CT scans. Although scGAN [23] is designed for shadow segmentation, it is also included in the comparison because it provides a fusion strategy to integrate different sensitivity parameters in a single loss function of the generator, which is also related to our task. Moreover, as our model is trained with the cGAN framework, it is also compared with the well-known pix2pix-cGAN [17] model for segmentation.

For the ISOS methods, we compare our model with 14 methods, covering the state-of-the-art ones in this field. These methods can be categorized into four groups, including i) background suppression methods (Max-Median [12], Top-hat [29] and DSVT [27]); ii) contrast and saliency based methods (LCM [4], WLDM [9], PatchSim [3], MSLH [11], LDM [10] and MFMM [11]); iii) decomposition based methods (CLSDM [19], IPI [13], NIPPS [8] and RIPT [7]); and iv) learning based method FCnet [20].

4.4. Evaluation Metrics

The common metric to evaluate the (binary) segmentation result by considering the balance of MD and FA is the F-measure. It is the harmonic mean of Precision and Recall. This measure is used in our experiment. In addition, Precision and Recall are also displayed. It is worth highlighting that merely achieving high Precision *or* Recall does not necessarily indicate a good method. F-measure shall be the primary evaluation metric to compare the methods.

For those small object segmentation methods compared in Table 2, when their outputs are not strictly binary, a threshold of 0.5 will be applied for binarization before the above metrics are computed. For the ISOS methods compared in Table 3, they usually produce a gray-scale confidence map as output and then apply a threshold to localise objects in the map. F-measure, Precision and Recall are computed based on the binarized confidence map obtained by using the thresholding method suggested by the authors of those methods.

To give a more comprehensive evaluation, for the ISOS methods, the quality of their confidence map without thresholding is also measured. This is achieved via the ROC curve to evaluate their performance of localizing the objects or targets. Specifically, following the literature, two metrics, i.e., area under ROC curve (AUC) [10] and the detection probability (P_d) within a fixed false-alarm ([13]), are employed for this purpose.

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

4.5. Results and Discussion

4.5.1 Comparison with generic small object segmentation methods

In Table 2, the segmentation results from the proposed MDvsFA-cGAN and multiple deep learning models for small object segmentation are compared. As can be seen, on both the "AllSeqs" and "Single" datasets, our method achieves the highest F-measure on both datasets. This indicates its best balance to suppress the missed detection and the false alarms at the pixel-level. On top of this result, our method also achieves the highest precision with reasonably good recall. These are in contrast to the cases of some existing methods in comparison. For example, Front[15] and FCN-RSTN [28] show the highest recall on "AllSeqs" and "Single". However, the precision of Front is very low on both datasets, making its F-measure much lower than ours. Also, due to the inferior precision, FCN-RSTN loses to our method in F-measure on both datasets, and it loses in recall as well on the "Single" dataset to our method.

Table 2. Compare the proposed method with the generic small object segmentation methods. F-measure is the primary evaluation metric. Prec and Rec are for "precision" and "recall", respectively.

Dataset	**Method**	Prec	Rec	**F-measure**
AllSeqs	Front[15]	0.01	0.45	0.01
	FCN-RSTN[28]	0.13	**0.74**	0.22
	FCN-MFB[18]	0.09	0.66	0.15
	scGAN[23]	0.14	0.27	0.19
	pix2pix-cGAN[17]	0.01	0.02	0.01
	MDvsFA-cGAN (ours)	**0.17**	0.60	**0.27**
Single	Front[15]	0.10	**0.89**	0.18
	FCN-RSTN[28]	0.54	0.39	0.45
	FCN-MFB[18]	0.38	0.61	0.47
	scGAN[23]	0.47	0.55	0.50
	pix2pix-cGAN[17]	0.26	0.22	0.23
	MDvsFA-cGAN (ours)	**0.66**	0.54	**0.60**

4.5.2 Comparison with state-of-the-art ISOS methods

The comparison between the proposed MDvsFA-cGAN method and the state-of-the-art ISOS methods is reported in Table 3. In this experiment, we evaluate the performance at both the pixel-level (by F-measure, Precision and Recall for binary segmentation) and the object/target-level (by AUC and P_d for object detection).

Again, for segmentation, our MDvsFA-cGAN consistently achieves the best balance between the pixel-level missed detection and false alarms, as indicated by its highest F-measure obtained on both datasets. Also, it achieves the highest precision and recall on "AllSeqs" and the highest precision on "Single." Meanwhile, the proposed MDvsFA-cGAN demonstrates promising performance for small object/target detection. It outperforms all the existing ISOS methods in comparison in terms of both AUC and P_d. Specifically, it seems to better deal with background interference that affects the background filtering-based methods such as TopHat [29] and Max-median [12]. It learns to identify targets from complex background with fewer FAs, showing effectiveness in enhancing the targets compared with the contrast and saliency based methods such as LCM [4] and WLDM [9]. It demonstrates capability to suppress FAs from the rarely seen structures that often degrade the decomposition based methods such as CLSDM [19], IPI [13] and NIPPS [7]. Also, it aggregates both global and local contexts, superior to FCnet [20] that only learns from local features to detect targets.

By cross-referencing the results for small object/target segmentation (i.e., measured by F-measure, Precision and Recall) and detection (i.e., measured by AUC and P_d) in this table, it is revealed that although many existing ISOS methods have relatively good object/target detection performance, they indeed could not correctly label the exact pixels of the object/target. For example, on the "Single" dataset, MFMM [11] shows excellent target detection performance but poor target segmentation performance. As a result, it will not be applicable to the tasks requiring precise measurement, localization and recognition. This reinforces the necessity to improve small object segmentation in infrared images, which is the main purpose of this paper.

4.5.3 Ablation Study

In ablation study, we explore the following questions to understand the contributions of our model components.

Q1) *Does the proposed adversarial training between MD and FA outperform a single model targeting at reducing both through an integrated objective?*

Q2) *Whether using different network architectures for each sub-task contributes to the performance improvement?*

Q3) *Whether the proposed method is sufficiently insensitive to the value λ_1 and λ_2 in the data loss defined in Eq.(4)?*

We develop variants of our model to answer the first two questions. To facilitate expression, for our MDvsFA-cGAN model, let us denote the architecture of G_1 as CAN8 (CAN model with $M_{DF} = 8$), and that of G_2 as UCAN64 (CAN model with $M_{DF} = 64$ and using skip connections). To answer Q1, we develop two CAN models and two cGAN models for comparison, all of which optimise MD and FA in a single combined objective with the combination weight λ obtained by grid-search. Specifically, the two CAN models use either CAN8 or UCAN64 architecture and are denoted as CAN8-plain and UCAN64-plain, respectively. Each of the two cGAN models is composed of a single generator and a discriminator, while the generator tries to minimise MD and FA simultaneously. Depending on the architecture of their generator, the two cGAN models are denoted as CAN8-cGAN and UCAN64-cGAN, respectively.

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

Table 3. Comparison with the state-of-the-art ISOS methods

Method	**AllSeqs**					**Single**				
	Precision	Recall	**F-measure**	AUC	P_d	Precision	Recall	**F-measure**	AUC	P_d
Max-median[12]	0.002	0.10	0.004	0.76	0.15	0.04	0.11	0.05	0.44	0.40
Tophat[29]	0.05	0.13	0.07	0.55	0.56	0.01	0.20	0.17	0.41	0.38
DSVT[27]	0.01	0.27	0.01	0.81	0.24	0.17	0.27	0.21	0.64	0.57
CLSDM[19]	0.15	0.25	0.18	0.57	0.57	0.34	0.35	0.34	0.48	0.14
LCM[4]	0.06	0.36	0.11	0.50	0.46	0.12	0.46	0.19	0.67	0.66
WLDM[9]	0.05	0.25	0.08	0.49	0.49	0.36	0.53	0.43	0.86	0.85
IPI[13]	0.13	0.34	0.19	0.64	0.65	0.43	**0.60**	0.50	0.86	0.87
NIPPS[8]	0.01	0.48	0.02	0.81	0.34	0.10	0.28	0.15	0.17	0.09
PatchSim[3]	0.16	0.34	0.22	0.77	0.77	0.57	0.51	0.54	0.85	0.84
FCnet[20]	0.04	0.40	0.07	0.63	0.10	0.15	0.34	0.21	0.71	0.74
MSLH[11]	0.01	0.19	0.02	0.73	0.17	0.12	0.33	0.17	0.56	0.54
LDM[10]	0.01	0.44	0.02	0.73	0.17	0.34	0.51	0.41	0.89	0.89
MFMM[11]	0.01	0.42	0.02	0.61	0.14	0.32	0.51	0.39	0.91	0.91
MDvsFA-cGAN (ours)	**0.17**	**0.59**	**0.27**	**0.84**	**0.79**	**0.66**	0.54	**0.60**	**0.91**	**0.92**

Table 4. Ablation study. "Prec" and "Rec" are for "precision" and "recall", respectively.

Test-set	**Method**	Prec.	Rec.	**F-measure**
AllSeqs	CAN8-plain	0.12	0.65	0.20
	UCAN64-plain	0.13	0.46	0.20
	CAN8-cGAN	0.12	0.62	0.21
	UCAN64-cGAN	0.13	0.60	0.22
	CAN8-double	0.10	**0.78**	0.17
	UCAN64-double	0.14	0.38	0.20
	MDvsFA-cGAN (ours)	**0.17**	0.59	**0.27**
Single	CAN8-plain	0.22	0.57	0.31
	UCAN64-plain	0.27	0.70	0.39
	CAN8-cGAN	0.26	0.61	0.36
	UCAN64-cGAN	0.28	**0.71**	0.40
	CAN8-double	0.26	0.69	0.37
	UCAN64-double	0.38	0.71	0.50
	MDvsFA-cGAN (ours)	**0.66**	0.54	**0.60**

To answer Q2, two models, denoted as CAN8-double and UCAN64-double are built. Each of them is composed of two generators and one discriminator in the same way as MDvsFA-cGAN. However, in those two models, the two generators share the same architecture of either CAN8 or UCAN64. The comparison results are reported in Table 4.

As can be seen, our model MDvsFA-cGAN outperforms the models (CAN8-plain, UCAN64-plain, CAN8-cGAN, and UCAN64-cGAN) that use a single objective to suppress MD and FA, no matter whether they use CAN or cGAN as backbones. We attribute this to two probable reasons. First, using a single network, these models have to deal MD and FA with the same network design, which however usually requires "opposite" strategies for suppression. Second, the dynamic balance between MD and FA in MDvsFA-cGAN provides a better way to nonlinearly integrate these two objectives, which is superior to the static linear combination used in the compared models. Moreover, the advantage of MDvsFA-cGAN over CAN8-double and UCAN64-double shows the benefits of using different network architectures for different sub-tasks (reducing MD or FA) for ISOS.

Table 5. Insensitivity of λ_1 and λ_2 in Eq.(4) on "Single" dataset.

λ_1	λ_2	**F-measure**	λ_1	λ_2	**F-measure**
100	10	0.60	500	1	0.59
100	5	0.60	200	1	0.59
100	1	0.60	100	1	0.60
100	0.5	0.60	50	1	0.60
100	0.1	0.60	10	1	0.58

To answer the third question, we test different values of λ_1 and λ_2. Note that, due to the small sizes of objects/targets in ISOS problems, the magnitudes of FA are usually 10 times more than those of MD. Taking this into account, we set $\lambda_1 = 100$ and $\lambda_2 = 1$ in Eq.(4) to make G_1 focus on MD and G_2 focus on FA. To show the insensitivity of our method to these two parameters, we fix $\lambda_1 = 100$ and vary λ_2 to test the performance change and then fix $\lambda_2 = 1$ and vary λ_1. As seen in Table 5, even when λ_1 and λ_2 are varied in a range where its max value is 50-100 times of its min, the fluctuation of F-measure is almost negligible.

5. Conclusion

In this paper, we decompose the ISOS problem into two sub-tasks of suppressing MD and FA, respectively, and jointly solve these two sub-tasks via adversarial learning. This new learning framework enables us to disentangle the suppression of MD and FA, design different models that better suit each sub-task, and provide a dedicated balance of MD and FA to lower the rates of both. It provides a new perspective to ISOS research and demonstrates its superiority over existing methods in this field. Also, it is our hope that the proposed method can inspire other computer vision research work that needs to strike a delicate balance between two mutually competitive criteria.

Acknowledgement. The author Huan Wang is supported by National Science Foundation of China (Grant No.61703209 and 61773215).

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

References

[1] Dwi Anoraganingrum, Sabine Kröner, and Björn Gottfried. Cell segmentation with adaptive region growing. *ICIAP Venedig, Italy*, pages 27–29, 1999. 1

[2] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. *Remote Sensing*, 9(4):368, 2017. 1

[3] Kun Bai, Yuehuan Wang, and Qiong Song. Patch similarity based edge-preserving background estimation for single frame infrared small target detection. In *IEEE International Conference on Image Processing*, pages 181–185. IEEE, 2016. 3, 6, 8

[4] CL Philip Chen, Hong Li, Yantao Wei, Tian Xia, and Yuan Yan Tang. A local contrast method for small infrared target detection. *IEEE Trans. geoscience and remote sensing*, 52(1):574–581, 2014. 1, 3, 6, 7, 8

[5] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. In *IEEE International Conference on Computer Vision*, volume 9, pages 2516–2525, 2017. 5, 6

[6] Zheng Cui, Jingli Yang, Shouda Jiang, and Changan Wei. Target detection algorithm based on two layers human visual system. *Algorithms*, 8(3):541–551, 2015. 3

[7] Yimian Dai and Yiquan Wu. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8):3752–3767, 2017. 3, 6, 7

[8] Yimian Dai, Yiquan Wu, and Yu Song. Infrared small target and background separation via column-wise weighted robust principal component analysis. *Infrared Physics & Technology*, 77:421–430, 2016. 3, 6, 8

[9] He Deng, Xianping Sun, Maili Liu, Chaohui Ye, and Xin Zhou. Small infrared target detection based on weighted local difference measure. *IEEE Trans. on Geoscience and Remote Sensing*, 54(7):4204–4214, 2016. 1, 3, 6, 7, 8

[10] He Deng, Xianping Sun, Maili Liu, Chaohui Ye, and Xin Zhou. Entropy-based window selection for detecting dim and small infrared targets. *Pattern Recognition*, 61:66–77, 2017. 6, 8

[11] He Deng, Xianping Sun, and Xin Zhou. A multiscale fuzzy metric for detecting small infrared targets against chaotic cloudy/sea-sky backgrounds. *IEEE Trans. on Cybernetics*, 2018. 3, 6, 7, 8

[12] Suyog D Deshpande, Meng Hwa Er, Ronda Venkateswarlu, and Philip Chan. Max-mean and max-median filters for detection of small targets. In *Signal and Data Processing of Small Targets 1999*, volume 3809, pages 74–84. International Society for Optics and Photonics, 1999. 1, 3, 6, 7, 8

[13] Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G Hauptmann. Infrared patch-image model for small target detection in a single image. *IEEE Trans. Image Processing*, 22(12):4996–5009, 2013. 1, 3, 6, 7, 8

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[15] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1442–1450. IEEE, 2018. 1, 2, 6, 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International conference on computer vision*, pages 1026–1034, 2015. 6

[17] Phillip Isola and et.al Zhu. Image-to-image translation with conditional adversarial networks. *IEEE conference on computer vision and pattern recognition*, 2017. 3, 4, 6, 7

[18] Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–9, 2016. 1, 2, 6, 7

[19] Li Li, Hui Li, Tian Li, and Feng Gao. Infrared small target detection in compressive domain. *Electronics Letters*, 50(7):510–512, 2014. 6, 7, 8

[20] Ming LIU, Hao-yuan DU, Yue-jin ZHAO, Li-quan DONG, and Mei HUI. Image small target detection based on deep learning with snr controlled sample generation. *Current Trends in Computer Science and Mechanical Automation*, pages 211–220, 2017. 3, 5, 6, 7, 8

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[22] Bojan Milovanović and Ivana Banjad Pečur. Review of active ir thermography for detection and characterization of defects in reinforced concrete. *Journal of Imaging*, 2(2):11, 2016. 1

[23] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *IEEE International Conference on Computer Vision*, pages 4520–4528. IEEE, 2017. 3, 6, 7

[24] Hairong Qi, Phani Teja Kuruganti, and Wesley E Snyder. Detecting breast cancer from thermal infrared images by asymmetry analysis. *Medicine and Medical Research*, 38, 2012. 1

[25] Michael Teutsch and Wolfgang Krüger. Classification of small boats in infrared images for maritime surveillance. In *Waterside Security Conference (WSS), International*, pages 1–7. IEEE, 2010. 1

[26] Ruben Usamentiaga, Yacine Mokhtari, Clemente Ibarra-Castanedo, Matthieu Klein, Marc Genest, and Xavier Maldague. Automated dynamic inspection using active in-

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.

frared thermography. *IEEE Trans. on Industrial Informatics*, 2018. 1

[27] Changcai Yang, Jiayi Ma, Shengxiang Qi, Jinwen Tian, Sheng Zheng, and Xin Tian. Directional support value of gaussian transformation for infrared small target detection. *Applied optics*, 54(9):2255–2265, 2015. 6, 8

[28] et al Yu Q., Xie L. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. *IEEE Conference on Computer Vision and Pattern Recognition.*, pages 8280–8289, 2018. 2, 6, 7

[29] Ming Zeng, Jianxun Li, and Zhang Peng. The design of top-hat morphological filter and application to infrared target detection. *Infrared Physics & Technology*, 48(1):67–76, 2006. 1, 3, 6, 7, 8

Authorized licensed use limited to: Wuhan University. Downloaded on June 26,2023 at 09:20:03 UTC from IEEE Xplore. Restrictions apply.