PixelGame: Infrared Small Target Segmentation as a Nash Equilibrium

Heng Zhou\textsuperscript{a}, Graduate Student Member, IEEE, Chunna Tian\textsuperscript{a}, Member, IEEE, Zhenxi Zhang\textsuperscript{a}, Graduate Student Member, IEEE, Chengyang Li, Graduate Student Member, IEEE, Yongqiang Xie\textsuperscript{a}, and Zhongbo Li\textsuperscript{a}

Abstract—A key challenge of infrared small target segmentation (ISTS) is to balance false negative pixels (FNs) and false positive pixels (FPs). Traditional methods combine FNs and FPs into a single objective by weighted sum, and the optimization process is decided by one actor. Minimizing FNs and FPs with the same strategy leads to antagonistic decisions. To address this problem, we propose a competitive game framework (pixelGame) from a novel perspective for ISTS. In pixelGame, FNs and FPs are controlled by different player whose goal is to minimize their own utility function. FNs-player and FPs-player are designed with different strategies: One is to minimize FNs, and the other is to minimize FPs. The utility function drives the evolution of the two participants in competition. We consider the Nash equilibrium of pixelGame as the optimal solution. In addition, we propose maximum information modulation (MIM) to highlight the target information. MIM effectively focuses on the salient region including small targets. Extensive experiments on two standard public datasets prove the effectiveness of our method. Compared with other state-of-the-art methods, our method achieves better performance in terms of F1-measure ($F_1$) and the intersection of union.

Index Terms—Deep learning, game theory, infrared image, small target segmentation.

I. INTRODUCTION

Infrared (IR) target segmentation plays a fundamental role in many applications, including surveillance and reconnaissance [1], [2], precise strike, and guidance [3], [4]. To the long distance in real applications, IR targets are usually “dim,” “small,” and “sparse” compared with RGB images. The top of Fig. 1 is a typical IR image [5], and the bottom is a common RGB image of natural scene [6]. As can be seen from Fig. 1, the infrared small target segmentation is more challenging than visible light images. To be specific, 1) dim: The noisy and clutter background results in that the IR targets have low contrast and low signal-to-clutter ratio; 2) small: The pixels of IR targets only account for a small proportion in the image. Most of the pixels in IR image are background pixels; 3) sparse: The layout of IR small targets is sparse. The above three properties cause difficulties in obtaining robust and generalized feature representations of small infrared targets.

The traditional methods rely on hand-crafted features. They mainly either simplify the small target as a bright spot [7], [8], or model the background, target, and the relationship between them [9], [10] in a particular scene. Most traditional methods are designed with specific yet limited features, which failed to cover multiple scenarios, and thus, resulted in degraded performance in open and diverse environments. With the rapid development of convolutional neural networks (CNN), more and more researchers employed deep learning networks to ISTS and detection [11], [12], [13], [14], [15]. Instead of designing IR target features manually, CNN can automatically mine IR target representations in an end-to-end manner, and consequently, achieve better results. Nevertheless, most of the
existing deep learning methods extract general features directly from the whole image and do not take into account the specific imaging processes involved in small IR targets.

The foreground–background imbalance problem is one of the main challenges in the infrared small target segmentation (ISTS). As shown in Fig. 1(a), the foreground pixels in the image are far fewer than the background pixels. Specifically, a large number of background pixels are incorrectly segmented as targets (false positive pixels, FPs). A small number of target pixels are submerged by clutter (false negative pixels, FNs).

In order to balance false negative pixels (FNs) and false positive pixels (FPs), the previous CNN-based methods [16] combine the two objectives into one function through weighted sum. The combined objective functions include Jaccard loss [17], Dice loss [18], F1 score [19], asymmetric similarity loss [20], sensitivity-specificity loss [21], and penalty loss [22]. The methods based on the combined loss function tune up the weights to get acceptable solutions, the selection being done by the one actor. The weights as hyperparameters of the loss function control the tradeoff between FNs and FPs. However, such training objective designs mainly suffers from two limitations: 1) The same strategy to minimize FNs and FPs simultaneously lead to antagonistic decisions. The former aims to predict as many target pixels as possible, while the latter tends to predict a small number of target pixels with high confidence; 2) extensive studies on loss function show that the setting of their hyperparameters is an experienced and difficult work. Therefore, it is intuitive and rational to optimize FNs and FPs as two objectives independently, as in our work.

Inspired by game theory, we optimize the two objectives using an adversarial strategy. Game theory is a framework or paradigm to solve multiobjective optimization problems, especially in dealing with antagonistic criteria [23], [24], [25]. Game-theoretic learning uses an equilibrium state instead of the optimal solution. Following the above idea, we design two tailored subnetworks to act as the FNs-player and FPs-player in the game. FNs-player and FPs-player focus on FNs and FPs, respectively. The players, actions, and utility functions in the game theory are corresponding to subnetworks, the change of network parameters and loss functions in the proposed framework, respectively [26], [27]. In this way, the ISTS is transformed into a game paradigm. Under the constraints of the utility function, FNs-player and FPs-player flexibly choose the different optimization strategies. Finally, the division of strategy ensures that the whole network achieves Nash equilibrium, which is an ingenious tradeoff between FNs and FPs.

Moreover, the context of image is important for target segmentation [28], [29], [30]. A larger receptive field of traditional deep convolution can aggregate contextual information, but, they lose some spatial location information. The receptive field is mainly expanded by downsampling in deep CNN. Different from large-scale objects, the size of IR targets is usually small. Resolution degeneration may cause the IR small targets fail to be segmented. Therefore, to solve the contradiction between deep high-level semantics and shallow high-resolution feature maps, we adopt the dilated convolution modules [13] in both FNs-player and FPs-player to construct fully dilated convolution network (FDCN). FDCN can obtain larger receptive fields and maintain the spatial resolution at the same time. As shown in Fig. 2, as the receptive field of dilated convolution increases, the small targets are effectively retained and enhanced in the high-resolution features.

In addition, different from the obvious semantic dependencies between objects and backgrounds (e.g., boats and rivers, cars and roads, etc.) in RGB image segmentation, the IR targets are grouped into one broad class in ISTS. The semantic contrast between backgrounds and targets is weak. We observe that small IR targets are usually local salient in a specific region. Based on these observations, we propose maximum information modulation (MIM). MIM absorbs the advantages of attention mechanism [31], [32] in focusing on effective information. MIM effectively suppresses irrelevant information and enhances the representation of small target in the proposed framework.

In summary, the main contributions of this article are given as follows.

1) We present a novel perspective to model ISTS as a multiplayer strategy game (pixelGame). FNs-player and FPs-player focus on reducing FNs and FPs in pixelGame, respectively. At the same time, a new utility function of pixelGame is designed to encourage two players to conduct the game. The utility function ensures that the participants fully play the game and eventually reach Nash equilibrium.

2) To handle the small targets, we adopt the dilated convolution modules in both FNs-player and FPs-player. FDCN takes both large receptive field and high-resolution feature map into account.

3) Due to the fact that IR targets are usually local salient regions in images, we propose MIM to suppress irrelevant background information by calculating local maxima, which improves the feature discrimination ability on small targets.

The rest of this article is organized as follows. In Section II, we briefly review the related work on ISTS, deep learning based game theory, and two benchmark datasets. Section III gives a meticulous description of proposed pixelGame model on ISTS. In Section IV, we conduct extensive experiments on ablation study and comparison with state-of-the-art (SOTA) methods on.
the benchmark datasets. The results prove the effectiveness of our method. Finally, Section V concludes the article.

II. RELATED WORKS

In this section, we briefly review related works on ISTS, deep learning-based game theory, and two benchmark IR dim small target datasets.

A. Infrared Small Target Segmentation (ISTS)

ISTS models are divided into two categories: 1) Traditional methods based on mathematical modeling with strong prior assumptions; and 2) CNN-based methods emerging in recent years. Specifically, the traditional ISTS methods mainly include spatial domain filtering methods and optimization-based methods.

Traditional spatial domain-based methods, such as Top-Hat filtering [7] and Max-Median filtering [33], focus on suppressing the background. Compared with the classic Top-Hat filter, Bai et al. [8] constructed a ring filter template through two related but different structural elements to better suppress the background and noise. Deng et al. [34] weighted the image entropy by multiscale gray difference to improve the signal-to-clutter ratio (SCR) of small targets. Gao et al. [35] treated the target as a special sparse component in the noise, so as to distinguish the target from similar background noise. Huang et al. [36] used a relatively large density gap between targets and their neighbors to eliminate the interference caused by clutters in complex backgrounds. Moradi et al. [37] proposed a directional approach to enhance the target area and suppress structural backgrounds. Many methods apply multiscale technology to suppress the background. Nie et al. [38] proposed a multiscale local homogeneity measure to improve the saliency of small targets. Gao et al. [39] utilized multiscale gray and variance difference metrics to enhance the feature representation of small target and mitigate background fluctuations, which improves the detection accuracy. In cluttered background, He et al. [40] enhanced targets by exploiting multiscale differences in intensity distribution changes and gray values.

Inspired by the human visual system, many local contrast measure (LCM)-based methods [41], [42] have been explored. Wei et al. [43] presented multiscale patch-based contrast measure to increase the contrast between the target and background. Huang et al. improved LCM from the aspects of multiscale [44], target shape [45] and the difference between the target and the background [46]. Lu et al. [47] utilized a division scheme of surrounding area to capture the derivative properties of the target. These methods extract the difference between the target and the background from various aspects, but, their performance is limited when the background changes dramatically or the target is hidden in the background.

The optimizing methods based on low-rank matrix recovery theory assume that the raw image is generated by a low-rank subspace, and the small targets are formulated as sparse singularity. The infrared patch-image (IPI) model [48] regarded small target detection as an optimization problem of recovering low-rank and sparse matrices. Dai et al. [49] used the partial sum of singular values instead of the nuclear norm of IPI to constrain the low rank background patch image. For various highly complex background scenes, Wang et al. [50] combined the total variation regularization term and principal component pursuit (TV-PCP) to comprehensively describe background feature. Wang et al. [51] analyzed the multisubspace structure of heterogeneous background data, and proposed a stable multisubspace learning method (SMSL) based on the internal structure of actual images to improve the robustness of the model. Self-regularized weighted sparse (SRWS) [52] model mined the potential information in the background, and transformed the small target segmentation into the optimization problem of extracting clutter from multiple subspaces.

Compared with matrix, tensors have more advantages in handling high-dimensional data [53]. To distinguish real targets from background residuals in heterogeneous scenes, Zhang et al. [54] proposed an edge and corner awareness-based spatial-temporal tensor (ECA-STT) model. Sun et al. [3] extended the properties of multisubspace to infrared patch-tensor (IPT) structure to better characterize the highly heterogeneous IR image background. Kong et al. [55] promoted t-SVD to multimodal t-SVD and enhanced the accuracy of background rank representation in the IPT model. The more accurate the assumptions, the better the performance of these existing IPT-based methods. Therefore, the performance of the model based on optimization relies on the constructed data structure and prior assumptions.

Specific assumptions cannot adapt to the open and diversified background environment. The model based on deep learning has a large capacity and contains a variety of different scenes. In recent years, the release of multiple IR small target datasets has promoted the research of methods based on CNN. Dai et al. [5] proposed the first CNN-based single-frame ISTS model, and designed an asymmetric context modulation (ACM) module to fuse high-level semantics and low-level details. ALCNet [56] transformed the traditional local contrast measurement method into a nonlinear module in the convolution network by combining domain knowledge, which alleviated the problem of the minimum internal characteristics of the pure data-driven methods. ISTDU-Net [57] improved the U-Net [58] segmentation model by increasing the response of small target features and suppressing similar background information, which improves the recognition ability to small targets.

ACM, ALCNet, and ISTDU-Net take ResNet-20 [59] as the backbone network to extract IR small target features, and use the intersection over union (IoU)-based weighted loss function to guide model optimization. Nevertheless, there are obvious differences in size, energy, and layout between IR and visible targets.

On the one hand, the ResNet structure obtains a large receptive field to fuse context information by sacrificing spatial resolution. Different from large-scale RGB targets, the resolution degeneration may cause IR small targets fail to be segmented. On the other hand, the IoU-based loss function minimizes FNs and FPs simultaneously, which increases the difficulty of model optimization and leads to antagonistic decisions. Minimizing FNs aims to predict as many target pixels as possible, while minimizing FPs trends to predict less number of high-confident target pixels.
TABLE I
SEVERAL LATEST ISTS DATASETS

| Datasets      | Background | Year | Samples (train&val) | Samples (test) | Image size | Label type     | Synthetic/real |
|---------------|------------|------|---------------------|----------------|------------|----------------|----------------|
| NUST-ISTS [14]| BC & LSR   | 2019 | 10,000              | 100            | (101~442)×(96~327) | Pixel          | Synthetic      |
| NUAA-ISTS [5 ]| BC & LSR   | 2021 | 341                 | 86             | (135~456)×(96~349) | Pixel/box      | Real           |

Notes: BC and LSR represent background clutters and low spatial resolution, respectively.

Fig. 3. Representative IR images with various backgrounds and targets from the NUST-ISTS [14] and NUAA-ISTS [5] datasets. For better visualization, the framed area is magnified. The top is from NUST-ISTS, and the bottom is from NUAA-ISTS.

Under the framework of generative adversarial networks (GAN) [60], MDvsFA-cGAN [14] used two different segmentation networks as generators, and aims to balance the results of the two generators through adversarial learning between the generators and the discriminator. MDvsFA-cGAN tries to find the Nash equilibrium of the generator and discriminator. In addition, GAN networks are difficult to train and prone to model collapse, which causes the generators to produce samples in the same mode. Differently, our method aims to find the equilibrium state of the two segmentation networks.

The core idea of GAN is game theory [23]. To take advantage of game theory, we design two novel segmentation networks as two players. Under the guidance of the utility function, the two networks play games directly. When two players reach Nash equilibrium in the game, the final results are generated by the segmentation networks.

B. Deep Learning-based Game Theory

Generally, the deep learning-based game is simplified, which consists of three parts [26]: 1) The participants of the game are neural units or neural networks; 2) the choices of each participant; 3) the objective function of each participant. A game in strategic form is given as follows:

\[
(T, (\Theta_k)_{k\in T}, (\Phi_k)_{k\in T})
\]

where \(T\{1, 2, \ldots\}\) is a set of players, \(\Theta_k\) is the set of actions of player \(k \in T\), which is essentially a weight, \(\Phi_k\) is a loss function of the player \(k\). The strategy of player \(k\) is to optimize \(\Phi_k\).

Inspired by the outstanding performance of game theory on principal component analysis [27], we regard the task of ISTS as a competitive game. Different networks as game players focus on different incorrectly segmented pixels. Under the guidance of the utility function, the two players continue to play the game, and finally reach the Nash equilibrium to output the segmentation results.

C. Infrared Dim Small Target Datasets

Lacking large-scale data sets severely hampered the application of deep learning on small IR targets. The recently emerged NUST-ISTS [14] is the first large-scale dataset for ISTS. NUST-ISTS contains small targets with various real backgrounds, which results in rich data samples. NUAA-ISTS [5] is the first real scene IR small target dataset. NUAA-ISTS provides various annotations including mask and bounding box. The detailed properties of those three datasets are shown in Table I.

Our experiments are mainly implemented on NUST-ISTS and NUAA-ISTS with pixel-level object masks. Some representative IR images of the two datasets are shown in Fig. 3. As can be seen from Fig. 3, the IR targets in the NUST-ISTS and NUAA-ISTS datasets have small sizes and complex backgrounds. As reported in Table I, the training set of NUST-ISTS contains 10,000 images that have 128×128 pixels, and the images in the test set have various sizes, which range from 101×96 to 442×327. NUAA-ISTS contains 427 images. The image resolution of NUAA-ISTS varies from 135×96 to 456×349. From the perspective of data sources, NUST-ISTS mainly uses the combination of real targets and real scenes to generate large-scale datasets. The advantage of the NUAA-ISTS dataset is that the samples cover a variety of natural scenes, and different samples are taken from the real background.
To further analyze the two datasets, we statistically count the number and size of the targets. The distribution of the number of targets in each image in those two datasets is shown in Fig. 4(a). It can be observed that more than 80% images contain only one target. In detail, the images in NUST-ISTS mainly contain single target, while images in NUAA-ISTS usually contain multiple targets. Fig. 4(b) shows the statistical distribution of the target area (number of pixels contained in the target) on each dataset. From the cumulative line chart on the right, we found that more than 90% IR targets contain less than 100 pixels, which occupy less than 1% in the image. We can see that the targets of IR images are small, dim, and scattered. It can be seen from Fig. 4(b) that more than half of the targets contains about 20 pixels.

Usually, small targets (e.g., aircraft, missiles) move rapidly in complex and variable clutter, making IR images have a very low SCR [61]. SCR [48], [62] is used to measure the target intensity and background intensity. In general, the higher SCR of the target, the easier the target is to be segmented. In IR dim small target segmentation, SCR is defined as follows:

$$\text{SCR} = \frac{|\mu_t - \mu_c|}{\sigma_c}$$  \hspace{1cm} (2)

In (2), $\mu_t$ represents the target intensity, which is the mean gray value of the target region. $\mu_c$ and $\sigma_c$ represent the mean and standard deviation of the gray value in the target neighborhood region, respectively. The target neighborhood region is set to be three times the size of the target region in this paper.

As shown in Fig. 4(c), the SCR of about 70% IR small targets is lower than 5. The clutter signals, such as clouds, trees, rocks, ground, etc., account for most energy of the IR images. The target can be regarded as the local extreme point in a specific region, but its energy is very weak in the global background.

Given the above analysis, Fig. 4 indicates that the challenges of ISTS not only lie in the limited information from target, but also lie in the complex and changeful background.

III. PIXELGAME: INFRARED DIM SMALL TARGET SEGMENTATION BASED ON GAME THEORY

As shown in Fig. 5, the proposed pixelGame consists of two subnetworks: 1) FNs-player ($S_1$); and 2) FPs-player ($S_2$), which segment the IR image pixel by pixel under the guidance of respective utility functions. To handle the small targets, we pay attention to deep high-level semantics and shallow high-resolution feature maps at the same time. Therefore, we combine dilated convolution and encode-decode structure to form the backbones of FNs-player and FPs-player.
TABLE II
DETAILED BACKBONE OF THE NETWORKS

| Encoder-decoder       | FDCN_9 | FDCN_13 |
|-----------------------|--------|---------|
| conv-k3-d1-c128       | conv-k3-d1-c64 |
| conv-k3-d2-c128       | conv-k3-d2-c64 |
| conv-k3-d4-c128       | conv-k3-d4-c64 |
| conv-k3-d8-c128       | conv-k3-d8-c64 |
| conv-k3-d16-c128      | conv-k3-d16-c64 |
| conv-k3-d32-c128      | conv-k3-d32-c64 |
| conv-k3-d64-c128      | conv-k3-d64-c64 |
| conv-k3-d128          | conv-k3-d128 |
| conv-k3-d32          | conv-k3-d32 |
| conv-k3-d64          | conv-k3-d64 |
| conv-k3-d128          | conv-k3-d128 |
| conv-k3-d32         | conv-k3-d32 |
| conv-k3-d64         | conv-k3-d64 |

| Head                  | conv-k1-d1-c1 |
|-----------------------|---------------|
| conv-k1-d1-c1          | conv-k1-d1-c1 |

Notes: In the table, "conv-k(m)-d(n)-c(o)" represents a convolutional layer with m×m kernel, dilation factor of n and output channel number of feature maps of o. Head represents the last output layer of the network.

In this section, we introduce how to transform the ISTS task into a game in pixelGame. Specially, we deal with three challenges: 1) How to design suitable subnetworks to control the focus of different players (Section III-A); 2) how to improve the feature representation ability on small targets (Section III-B); 3) how to set a scientific and effective utility function for players to achieve the Nash equilibrium in competitive game (Section III-C).

A. FNs-Player and FPs-Player

Inspired by the game theory in solving antagonistic decisions, two segmentation networks as two players in the game optimize their utility functions, respectively. The subnetworks $S_1$ and $S_2$ segment the IR image $I$ pixel by pixel, as shown in Fig. 5. Formally, it can be represented as

$$
\begin{align*}
S_1(I) & \rightarrow O_1, \\
S_2(I) & \rightarrow O_2
\end{align*}
$$

(3)

where $O_1$ and $O_2$ denote to the segmentation results of two players, respectively.

In the small target segmentation task, the FNs and FPs are difficult to balance delicately. We separate FNs and FPs, and employ two players to divide and conquer. In order to obtain better performance, the two subnetworks use different structures according to tasks. FNs-player and FPs-player use FDCN with different network depth and dilation factor.

The detailed encoder-decoder structures of FDCN_9 and FDCN_13 are shown in Table II. FDCN_9 and FDCN_13 are the backbone networks of FNs-player and FPs-player, respectively. In all of the models, the convolutional layers except the last one are followed by batch normalization (BN) [63] and leaky rectified linear unit (leakyReLU) [64]. Specifically, the goal of $S_1$ player is to reduce the FNs of targets, optimizing TNs and FNs. We employ the shallow encoder-decoder network to extract local information and segment all the pixels of the suspected target. The FDCN_9 uses 9-layer convolution, and the dilation factor is increasing from 1 to 16.

Compared with FNs-player, FPs-player increases the accuracy of the predicted pixels belonging to the target class, by optimizing TPs and FPs. The pixels predicted by $S_2$ may be as precise as possible. FPs-player needs a larger context and better local receptive field, so FDCN_13 is deeper and its dilation factor is larger. The FDCN_13 contains 13 convolutional layers, and the maximum dilation factor is 64. Finally, the head layer is used to predict the class of each pixel, generating the binary mask of foreground and background.

B. Maximum Information Modulation

The information modulation methods represented by the attention mechanism aim to make the model focus on task-related information. In RGB object detection and segmentation, SENet [65] adaptively enhances task-relevant channels by learning the dependencies between different channels. Nonlocal network [66] is used to capture long-range dependencies and establish the interaction between two pixels with a certain distance on the image. GCNet [31] improves the nonlocal network and SENet, enabling query-independent lightweight modules to effectively extract global context information. Triplet attention [32] encodes interchannel relation and spatial relation, and establishes the dependencies between them to calculate attention weights.

Unlike RGB images, the SCR of IR dim small targets is very low, and the useful target information is usually submerged in irrelevant clutter and noise. Considering the small targets are arduous to segment, we introduce global max pooling (GMP) [67] and cross-channel max pooling (cMaxPool) [32] to enhance the local salient information of these targets. The MIM aims to increase the pertinence and capacity of extracted features. In FDCN_9 and FDCN_13, we add the MIM module in the skip connection.

The differences between MIM and other attention modules are highlighted in Fig. 6. It can be seen from the visualization results that the MIM performs better in capturing IR small targets with low SCR than other attention mechanisms.

The MIM enhances the salient information related to the target, and suppresses a large amount of noise and clutters that are not related to the target. In (4), the MIM is performed on the features $X \in \mathbb{R}^{C \times H \times W}$ of each layer of two encoders

$$
Z = m(X)
$$

(4)

where $m(\cdot)$ represent the MIM, modulation feature $Z \in \mathbb{R}^{C \times H \times W}$, $C$, $H$, and $W$ represents the channels, height, and width of the feature map, respectively.

The network structure of MIM is shown in Fig. 5. First, the pixel-wise correlation in the spatial domain is used to obtain cross-channel attention $V_i \in \mathbb{R}^{C \times 1 \times 1}$. Specifically, in (5), the feature map $X$ is first transformed through point-wise convolution (PWConv) [68] to fuse the relationship between the features of different channels. Equation (6) makes the MIM pay more attention to the correlation between different spatial positions of feature

$$
Y = PWConv(X),
$$

(5)
$FDCN_{\rho} \in \mathbb{V}_C$ (7) to obtain dual-channel attention $M_3$.

As shown in Fig. 5, the two-layer PWConv realizes the full interaction of different channel information through squeezing and excitation, and further enhances the cross-channel global attention $M_2$, in (9)

$$M_2 \in \mathbb{R}^{C \times 1 \times 1}, \text{and the last layer does not include leakyRelu activation.}$$

Then, we use cMaxPool to obtain the most significant information of different channels. $M_3$ in (10) is the max channel pooling feature map, where each point in the feature map is the max of the points at the same position within feature maps

$$M_3 = \sigma (c\text{MaxPool} (X, 0))$$

where $\sigma$ is the Sigmoid function, $M_3 \in \mathbb{R}^{1 \times H \times W}$.

Finally, guided by (9) and (10), MIM enhances small target features in both spatial and channel dimensions. The final feature is obtained by

$$Z = M_2 \cdot X + M_3 \cdot X$$

C. Utility Function

When evaluating segmentation results, the predicted results are often divided into the following four parts:

1) The number of correctly segmented pixels of target (TPs);
2) the number of incorrectly segmented pixels of background (FPs);
3) the number of correctly segmented pixels of background (TNs);
4) the number of incorrectly segmented pixels of target (FNs).

The target pixels are composed of TPs and FNs. The background pixels contain TNs and FPs. The confusion matrix of ISTS is shown in Table III. In the confusion matrix, the columns represent the predicted masks $O$ of the pixelGame, and the row represents the ground truth $G$ of the input images.

In order to achieve high-quality results for the FNs-player and FPs-player games, we design a novel utility function according to each player’s own focus and the overall constraints of the game. This utility function consists of three parts: 1) Player utility; 2) game utility; and 3) small target constraints. Scientific and effective utility function help the model reach the equilibrium state in the competitive game. They are as follows.

1) Player Utility: For player utility, the main goal of FNs-player and FPs-player is to minimize FNs and FPs. Combined with the confusion matrix in Table III, the utility functions $U$ of players are defined as follows:

$$U(O_1, G) = \frac{\text{FNs}}{\text{TNs} + \text{FNs}},$$

$$U(O_2, G) = \frac{\text{FPs}}{\text{TPs} + \text{FPs}}.$$
between them. Therefore, we define the game utility $G$ as follows:

$$G(O_1, O_2 | G) = \| (O_1 - G)(O_2 - G) \|_2.$$  \hfill (14)

Equation (14) makes the incorrectly segmented pixels of FNs-player and FPs-player as different as possible. Game utility further aggravates the game confrontation between them. Explicit antagonistic utility constraints can not only enhance the complementarity of their results, but also help the game optimization to reach the equilibrium state.

3) Small Constraints: To further handle ISTS, we add a small target constraint to narrow the game space, which ensures that the model is optimized in a reasonable space. Specifically, we increase the penalty for oversized targets by calculating the area of the prediction mask, which prompts the network to focus on the prediction of small targets. Formally

$$A(O) = \frac{1}{N} \sum_{i=1}^{N} o_i$$ \hfill (15)

where $o_i$ denotes the $i$th pixel value in the segmentation result $O$. $N$ represents the size of the image. Therefore, the average intensity of all target pixels is used as a small target constraint $A(O)$ considering the size of different masks.

Finally, we choose equal weights according to many experimental attempts. The utility $\Phi$ of pixelGame is as follows:

$$\Phi(O_1, O_2 | G) = U(O, G)+G(O_1, O_2 | G)+A(O).$$ \hfill (16)

D. PixelGame Network

FDCN$_9$ and FDCN$_{13}$ are the backbones of pixelGame network. A high-resolution prediction feature map is indispensable for small target segmentation. The dilated convolution captures a larger receptive field without reducing the spatial resolution of the feature. The dilation factors of the decoder are symmetrical to that of the encoder. In the encoder-decoder structure, the feature mapping with the same dilation factor exchanges information across layers through skip connection.

The larger dilation factor results in a larger receptive field. On the one hand, some pixels in the large receptive field are not fully utilized. On the other hand, the long-distance dependence of the pixels captured by the large receptive field is not accurate. Therefore, we use MIM module to improve small object feature representation.

The specific implementation is shown in Algorithm 1.

### Algorithm 1: PixelGame: ISTS as a Nash Equilibrium

**Data:** input image $I$, ground truth $G$, utility function $\Phi$, game player networks $\Theta \{S_1, S_2\}$, maximum information modulation (MIM)

**Output:** equilibrium state of the game $O$

1. $S_1 \leftarrow \text{MIM}(\text{FDCN}_9)$
2. $S_2 \leftarrow \text{MIM}(\text{FDCN}_{13})$
3. **for** epoch **do**
4.   $O_1, O_2 \leftarrow S_1(I), S_2(I)$
5.   $\Phi \leftarrow \text{Eq. (16)}$
6.   $S_1, S_2 \leftarrow \arg \min \Phi$
7. **end**
8. $O \leftarrow \text{mean}(O_1, O_2)$
9. **return** $O$

11. **Function** MIM (FDCN$_{2n+1}$):
12.   $\text{FDCN}_{2n+1} = \{x_1, \ldots, x_n, x_{n+1}, \ldots, z_1, o\}$,
13.   where $x_n$ and $z_n$ are the feature map of each layer of the encoder and decoder, respectively.
14. **for** layer $k \in [2, n]$ **do**
15.   $z \leftarrow m(x_k)$ using Eq. (4)
16.   $z_k^* \leftarrow z_k + z$
17. **end**
18. **return** $S \leftarrow \{x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, z_1^*, z_1, o\}$

D. PixelGame Network

FDCN$_9$ and FDCN$_{13}$ are the backbones of pixelGame network. A high-resolution prediction feature map is indispensable for small target segmentation. The dilated convolution captures a larger receptive field without reducing the spatial resolution of the feature. The dilation factors of the decoder are symmetrical to that of the encoder. In the encoder-decoder structure, the feature mapping with the same dilation factor exchanges information across layers through skip connection.

The larger dilation factor results in a larger receptive field. On the one hand, some pixels in the large receptive field are not fully utilized. On the other hand, the long-distance dependence of the pixels captured by the large receptive field is not accurate. Therefore, we use MIM module to improve small object feature representation.

The specific implementation is shown in Algorithm 1.

### IV. EXPERIMENTAL RESULTS AND ANALYSIS

In order to analyze the potential of pixelGame based on game theory, we compare it with the related SOTA methods on NUST-ISTS and NUAA-ISTS datasets. Moreover, we conducted ablation experiments to verify the effectiveness of different components in pixelGame. In particular, the following questions will be investigated in our experimental evaluation.

Q1: Our key insight is to transform the antagonistic decision of FNs and FPs using the same strategy into a competitive game in which two player participate. Based on the game theory, we study how FPs-player and FPs-player achieve a delicate balance under the guidance of utility function with inherent complementarity and optimization antagonism. (Section IV-B1).

Q2: We further explore the contribution of the composition of utility function to the Nash equilibrium in pixelGame. (Section IV-B2).

Q3: In an antagonistic decision of pixelGame, whether the performance of a multiplayer strategy game is better than that of a single integrated objective through weighted sums. (Section IV-B3).

Q4: We compare our pixelGame with other SOTA methods and show the segmentation results. The effectiveness of MIM module is proved. (Section IV-C).

A. EXPERIMENTAL SETUP

1) Implementation Details: The whole pixelGame network is then trained with the proposed utility function in an end-to-end manner. Specifically, the two player networks are trained separately and alternately. In each epoch, the parameters of the FPs-player network are fixed first, and the FNs-player network is trained. Then fix the parameters of the FNs-player network and train the FPs-player network. Finally, the two networks are trained in a loop until a Nash equilibrium is reached.

Same as the setting in [56], in order to stack images of different sizes into a batch, the size of each image is resized to $512 \times 512$, and randomly cut to $480 \times 480$ during training. In addition, each image is random flipping, cropping, and rotation. The pixelGame is implemented by PyTorch. We use Adam as the optimizer. The minibatch size is set as 8. The learning rate is set as $10^{-5}$ for the FNs-player and FPs-player. The training epoch of NUST-ISTS is
70 and that of NUAA-ISTS is 400. The experiment is conducted on a computer with 3.0 GHz CPU, 128 GB RAM, and four NVIDIA GeForce RTX 3090 GPUs.

2) Evaluation Metrics: We use precision ($P$), recall ($R$), $F_1$ score ($F_1$), and IoU to evaluate the ISTS methods.

The precision measures the proportion of correctly segmented target pixels in all segmented target pixels. The recall measures the proportion of correctly segmented target pixels in all true target pixels. In (17) and (18), $P$ and $R$ are defined as

$$P = \frac{\sum_{i=1}^{N} o_i g_i}{\sum_{i=1}^{N} o_i}, \quad (17)$$

$$R = \frac{\sum_{i=1}^{N} o_i g_i}{\sum_{i=1}^{N} g_i}. \quad (18)$$

Beside, IoU is also used to measure the coincidence between the predicted mask and the ground truth. IoU is defined in (19)

$$\text{IoU} = \frac{\sum_{i=1}^{N} o_i g_i}{\sum_{i=1}^{N} (o_i + g_i - o_i g_i)}. \quad (19)$$

In (17), (18), and (19), $o_i$ denotes the probability that the $i$th pixel in the segmentation result $O$ is the foreground. $g_i$ is the value of the corresponding position of ground truth $G$.

In order to evaluate the advantages and disadvantages of different algorithms, the concept of $F_1$ value is proposed based on (17) and (18). $F_1$ evaluates $P$ and $R$ together. $F_1$ is defined as follows:

$$F_1 = 2 \times \frac{P \times R}{P + R}. \quad (20)$$

B. Ablation Study

In this section, we study questions Q1–Q3 raised above.

1) Players Game and Nash Equilibrium (Q1): In Fig. 7(a) and (b), there is a clear trend of gradual steady state. As the game between players goes deep, the performance of the players is improved gradually, and finally reaches the Nash equilibrium.

From a single index, FNs-player has the highest recall and FPs-player has the highest precision. But from the final $F_1$ score, we can see that the pixelGame has achieved better performance than two players. $F_1$ score in Fig. 7(a) and (b) illustrates the feasibility and superiority of our separate optimization method.

For different players, FNs-player tends to predict pixels of all potential targets, so it can be seen from recall in Fig. 7(a) and (b) that more than 98% of the true target pixels are correctly predicted by FNs-player. Differently, FPs-player puts quality before quantity, which has high precision and low recall scores. Therefore, the precision of FPs-player on two datasets is more than 85%. Two players with opposing strategies promote pixelGame to achieve a better balance in the overall game framework.

Furthermore, from $F_1$ score in the Fig. 7, the antagonistic learning of players at the pixel level makes the two players have better complementarity at the image level. The fusion result of the two players is considerably better than that of a single player. As show in Fig. 7(a), the fusion results of FNs-player and FPs-player are improved by about 0.14 and 0.37 on the NUST-ISTS dataset, respectively. The Fig. 7(b) presents the fusion results of two players are improved by about 0.05 and 0.46 on the NUAA-ISTS dataset, respectively. The performance improvement proves the effectiveness of multiplayer adversarial learning based on game theory. Under the guidance of the proposed utility function, the model achieves higher accuracy and a more delicate balance between FNs and FPs. From Fig. 7, we can conclude that each player is integral and complementary, which together contribute to the final result.

Nash equilibrium [23], [25] represents a static state, where no player is willing to change strategy any more. During the pixelGame, Fig. 9(a) and (b) illustrate the evolution of player utilities for FNs-player and FPs-player on the NUST-ISTS and NUAA-ISTS datasets, respectively. As shown in Fig. 9(a) and (b), the utilities of the two players in the game tend to be steady after decline, and finally reach a stable state. The invariance of the utility functions of two players means that the players do not want to change their strategies, which means the pixelGame reaches a Nash equilibrium.

2) Impact of Utility Function (Q2): To verify the contribution of each utility, we analyze the results on the NUST-ISTS and NUAA-ISTS datasets. As shown in Table IV, the competitive
Fig. 8. Contribution of game utility to model performance. \( U \), \( A \), and \( G \) represent player utility, small target constraint, and game utility, respectively. (a) NUST-ISTS. (b) NUAA-ISTS.

Table IV

\[ \begin{array}{c|c|c|c|c|c} \hline \text{U} & \text{G} & \text{A} & \text{NUST-ISTS} & \text{NUAA-ISTS} \\ \hline \text{✓} & \text{✓} & \text{✓} & 0.5618 & 0.5092 & 0.7418 & 0.6933 \\ \end{array} \]

Notes: In the table, \( U \), \( G \), and \( A \) represent player utility, game utility, and small target constraint, respectively. Bold font highlights the best results in each column.

The contribution of the utility function on NUST-ISTS

The contribution of the utility function on NUAA-ISTS

Fig. 9. Convergence of the utility function in pixelGame. (a) NUST-ISTS. (b) NUAA-ISTS.

Table V

\[ \begin{array}{c|c|c|c|c|c} \hline \text{Objectives} & \text{NUST-ISTS} & \text{NUAA-ISTS} \\ \hline \text{FNs-player} & \text{FPs-player} & \text{FNs-player} & \text{FPs-player} \\ \hline \text{IoU loss} & 0.770/0.66 & 0.76/0.65 \\ \text{Dice loss} & 0.55/0.46 & 0.48/0.40 \\ \text{SS loss} & 0.32/0.22 & 0.32/0.21 \\ \text{Ours} & 0.27/0.17 & 0.35/0.29 \\ \end{array} \]

Notes: Bold font highlights the best results in each column.

game guided by player utility has achieved high results on \( F_1 \) and IoU. FNs and FPs, as two kinds of incorrectly segmented pixels, are the two opposing sides in the overall goal of the game. As shown in Table IV, most of the incorrectly segmented pixels are correctly segmented. In addition, it can be observed from Table IV that game utility makes the results of the two more complementary, and increases by about 5% and 10% on NUST-ISTS and NUAA-ISTS, respectively. The game utility makes the direct competition between the two players more purposeful. The area constraint of the small target is conducive to the faster convergence of the player network, and it improves the performance of the model to a certain extent.

We further analyze the impact of game utility \( G \) on FNs-player and FPs-player. The game utility urges the two players to pay attention to different predicted pixels. As shown in Fig. 8(a) and (b), the game utility results in a better Nash equilibrium, i.e., higher \( F_1 \) and IoU scores.

3) Advantage of Separate Game Objectives Compared to Combined Objectives (Q3): Next, we compare the performance of different classical combined objective functions, including Dice loss [17], IoU loss [18], and sensitivity specificity loss (SS loss) [21], etc. The parameter controlling the balance between sensitivity and specificity in SS loss is set to 0.5.

Table V shows that the model performance is poor under the guidance of the combined loss functions. In ISTS, IR targets are dim, small, and sparse within images. These challenges make commonly used combined loss functions unable to effectively focus a small number of target pixels, and cannot distinguish background and noise.

IoU and Dice similarity coefficient look very similar in terms of equations, and both are the most commonly used evaluation metrics in object segmentation. As reported in Table V, using IoU loss and Dice loss in NUST-ISTS and NUAA-ISTS datasets, FNs-player and FPs-player only reach about 0.50 and 0.76 on \( F_1 \), and the IoU of these two subnetworks are about 0.45 and 0.65, respectively. The optimization objectives of both mainly focus on TPs, treating two classes of incorrectly segmented pixels equally. However, due to the IR targets are small, the model leads to oversegmentation easily. The pixels of oversegmentation are mainly noise and background near the target, which are mainly reflected in FPs. The model cannot effectively balance a small
number of but very critical pixels such as the edges of dim small targets (FPs and FNs), resulting in stagnant performance.

Although FNs and FPs are considered in SS loss explicitly, their performance is not ideal. From the perspective of features, features of FP and FNs are similar, but they are different or even opposite in optimization strategies. For ISTS, the SS loss function uses the same strategy to optimize the countermeasure decision, so that the model cannot converge to the global optimal solution.

Comparing the results of combined objectives and separate optimization in Table V, it can be seen that FNs-player achieves leading performance by using combined objectives. Considering the beneficial balance of FNs-player to FPs-player, the performance of pixelGame is further improved. This also proves the prominent performance of our FNs-player and FPs-player network structure for ISTS.

### C. Comparison to SOTA Approaches

Finally, we solve the problem Q4 by comparing our pixelGame with several CNN-based methods and other traditional mathematical modeling methods. These methods can be categorized into the following four groups:

1) Traditional spatial domain-based methods (Top-Hat [7], the modified new white top-hat transformation (MNWTH) [8], Max-Median [33], the novel weighted image entropy (NWIE) [34], mixture of gaussians (MoG) with markov random field (MRF) [35], density peaks infrared small target detection (DPIR) [36], absolute directional mean difference (ADMD) [37]);

2) human visual system-based methods (LCM [41], improved local contrast measure (ILCM) [42], multiscale patch-based contrast measure (MPCM) [43], relative local contrast measure (RLCM) [44], tri-layer local contrast measure (TLLCM) [45], weighted strengthened local contrast measure (WSLCM) [46], multidirectional derivative-based weighted contrast measure (MDWCM) [47]);

3) optimization-based methods (IPI [48], SMSL [51], non-negative infrared patch-image model based on partial sum minimization of singular values (NIPPS) [49], TV-PCP [50], SRWS [52], weighted Schatten p-norm minimization (WSNM) and spatial-temporal infrared patch-tensor structure [53], ECA-STT [54], nonconvex tensor fibered rank approximation (NTFRA) [55]);

4) CNN-based algorithm (MDvsFA-cGAN [14], ACM [5], ALCNet [56]).

For a long time, the lack of an open benchmark has been one of the bottlenecks hindering the development of ISTS, which allows various algorithms to be compared fairly. We have summarized the existing target detection and segmentation methods of IR small targets, which is helpful to promote ISTS work. The comparison between the pixelGame and the SOTA methods is reported in Table VI. In experiments, we evaluate the performance of the algorithm at the pixel level using precision, recall, \( F_1 \) and IoU.

**Compared with traditional algorithms, pixelGame is comfortably ahead in both single score (precision and recall) and comprehensive score \( (F_1 \) and IoU). In CNN-based models, pixelGame better suppresses FNs and FPs, and achieves a more delicate balance between precision and recall. In Table VI, on**

| Method               | NUST-ISTS Precision | Recall | \( F_1 \) | IoU  | NUA-ISTS Precision | Recall | \( F_1 \) | IoU  |
|----------------------|---------------------|--------|----------|------|---------------------|--------|----------|------|
| Top-Hat [7] (OE 1996) | 0.09                | 0.21   | 0.12     | 0.08 | 0.14                | 0.11   | 0.12     | 0.13 |
| Max Median [33] (ISR 1999) | 0.05               | 0.14   | 0.05     | 0.03 | 0.04                | 0.18   | 0.04     | 0.03 |
| MNWTH [8] (PR 2010) | 0.23                | 0.61   | 0.27     | 0.18 | 0.18                | 0.27   | 0.22     | 0.27 |
| IPI [43] (TP 2013) | 0.51                | 0.49   | 0.50     | 0.38 | 0.31                | 0.48   | 0.34     | 0.28 |
| LCM [41] (TGRS 2013) | 0.15                | 0.36   | 0.21     | 0.13 | 0.15                | 0.29   | 0.22     | 0.14 |
| ILCM [42] (GRSL 2014) | 0.14                | 0.22   | 0.20     | 0.13 | 0.14                | 0.25   | 0.21     | 0.13 |
| MPCM [43] (PR 2016) | 0.28                | 0.45   | 0.34     | 0.24 | 0.29                | 0.37   | 0.34     | 0.27 |
| NWIE [34] (TAES 2016) | 0.23               | 0.21   | 0.21     | 0.18 | 0.31                | 0.21   | 0.23     | 0.14 |
| NIPPS [49] (INPHY 2017) | 0.12               | 0.29   | 0.15     | 0.10 | 0.29                | 0.31   | 0.32     | 0.21 |
| TV-PCP [50] (ICV 2017) | 0.38               | 0.25   | 0.34     | 0.22 | 0.32                | 0.47   | 0.38     | 0.40 |
| SMSL [51] (TGRS 2017) | 0.54                | 0.20   | 0.26     | 0.17 | 0.40                | 0.45   | 0.41     | 0.40 |
| RLCM [44] (GRSL 2018) | 0.39                | 0.45   | 0.41     | 0.31 | 0.46                | 0.44   | 0.44     | 0.31 |
| MoG-MRF [35] (PR 2018) | 0.22               | 0.45   | 0.28     | 0.20 | 0.25                | 0.34   | 0.31     | 0.23 |
| DPIR [36] (GRSL 2019) | 0.52                | 0.21   | 0.24     | 0.17 | 0.45                | 0.32   | 0.37     | 0.27 |
| TLLCM [45] (GRSL 2019) | 0.52               | 0.42   | 0.41     | 0.35 | 0.60                | 0.62   | 0.54     | 0.44 |
| WSLCM [46] (GRSL 2020) | 0.38               | 0.30   | 0.33     | 0.22 | 0.58                | 0.42   | 0.50     | 0.38 |
| WSNMSTSTP [53] (INPHY 2019) | 0.28             | 0.15   | 0.15     | 0.13 | 0.26                | 0.25   | 0.21     | 0.25 |
| MDvsFA-cGAN [14] (ICCV 2019) | 0.63          | 0.65   | 0.61     | 0.47 | 0.72                | 0.77   | 0.71     | 0.60 |
| MDWCM [47] (GRSL 2020) | 0.38               | 0.42   | 0.40     | 0.43 | 0.32                | 0.43   | 0.37     | 0.41 |
| ECA-STT [54] (TGRS 2020) | 0.53               | 0.46   | 0.49     | 0.38 | 0.79                | 0.58   | 0.64     | 0.53 |
| ADM [57] (SP 2020) | 0.57               | 0.22   | 0.27     | 0.19 | 0.69                | 0.38   | 0.42     | 0.30 |
| ACM [5] (WACP 2021) | 0.55               | 0.74   | 0.60     | 0.44 | 0.83                | 0.84   | 0.83     | 0.71 |
| SRWS [52] (NC 2021) | 0.56               | 0.34   | 0.44     | 0.33 | 0.74                | 0.47   | 0.53     | 0.39 |
| NTFRA [55] (TGRS 2021) | 0.47               | 0.43   | 0.46     | 0.38 | 0.67                | 0.55   | 0.57     | 0.44 |
| ALCNet [56] (TGRS 2021) | 0.36               | 0.73   | 0.60     | 0.45 | 0.88                | 0.54   | 0.85     | 0.73 |
| **pixelGame (w/o MIM)** | **0.60**          | **0.72** | **0.61** | **0.48** | **0.86**          | **0.85** | **0.85** | **0.74** |
| **pixelGame** | **0.66**          | **0.74** | **0.64** | **0.51** | **0.88**          | **0.86** | **0.86** | **0.75** |

*Notes: The best result in each column is in Red, the second is in Blue, and the third is in Green. The pixel Game (W/o MIM) is the baseline model.*
both the NUST-ISTS and NUAA-ISTS datasets, our method achieves the highest $F_1$ and IoU on both datasets, which are 0.64, 0.86 and 0.51, 0.75, respectively. For the foreground-background imbalance problem, our pixelGame achieves the best balance between FNs and FPs. From the comparison of precision and recall, which are 0.66, 0.74 and 0.88, 0.86, we can see that the precision and recall of our model are closer and more balanced.

Moreover, it can be seen from Table VI that deep learning models perform better than the traditional algorithms. Specifically, the CNN-based algorithms mostly design the loss function of the model at the pixel level, such as MDvsFA-cGAN [14] (based on miss detection and false alarm), ACM [5], and ALCNet [56] (based on nIoU). This demonstrates that the pixel-wise loss function has significant advantages in dense segmentation tasks. The traditional algorithms, such as Top-Hat [7] and LCM [41], mostly suppress the background, and enhance the difference between the target and the background at the image level. Similar to the algorithm of recovering the target from the feature space, such as IPI [48] and IPT [52]
In the CNN-based methods, most of the existing methods treat different kinds of error pixels as the same to optimize. For example, ACM [5] and ALCNet [56] use IoU loss as the optimization function. The integrated loss methods are not suitable for conquering extreme foreground-background imbalance problem of ISTS. MDvsFA-cGAN [14] makes the two sub-networks focus on FPs and FNs by loss reweighting, respectively. However, in essence, MDvsFA-cGAN optimizes FNs and FPs at the same time, which is not pure antagonistic learning between FNs and FPs. Differently, our pixelGame optimizes FPs and FNs separately. The division of strategy is beneficial for pixelGame to flexibly choose different optimization strategies, and it ensures that the networks achieve Nash equilibrium.

To demonstrate the effectiveness of MIM, we visualize the feature maps enhanced by MIM module. It can be observed from Fig. 6 that MIM performs better in capturing IR small targets with low SCR than other attention methods. The main content in IR image is the building, while the region of target is very small compared with the main building. SENet, GCNet, and Triplet attention tend to focus on the main building and neglect the small target. Compared to other attention methods, MIM prominently enhances the target information and suppresses the background. Benefiting from the different dilation factors, MIM effectively focuses on the salient region including small targets. Fig. 6 suggests that the enhanced features by MIM are powerful.

In addition, we design ablation experiments by removing MIM (denoted as w/o MIM). We apply simple addition when removing each module. From the last two rows of Table VI, compared to pixelGame (w/o MIM), pixelGame improves the segmentation accuracy by about 3% and 1%, respectively, on the two datasets. As can be seen from Table VII, MIM achieves the best performance on both datasets. The F1 score of MIM is about 2% higher on average than other methods. Tables VI and VII present the effectiveness of MIM for local significant context information enhancement in ISTS tasks. The efficient of MIM is fully proved.

We further study the spatiotemporal complexity of CNN-based methods. The results are shown in Table VIII. ACM and ALCNet use ResNet-20 [59] as the backbone architecture. Our pixelGame employ dilated convolutional networks. However, compared with other CNN-based methods, the computation load of our pixelGame is relatively high because of the increased amount of parameters. It can be seen from Table VIII that pixelGame trades off speed for the accuracy of the model.

Some segmentation results are shown in Fig. 10. It can be illustrated from the Fig. 10 that the pixelGame segments dim small targets with low SCR accurately and robustly. Due to the low spatial resolution, the targets of NUST-ISTS dataset are small and dim. Some bright background noise is incorrectly predicted as target pixels. The background of NUAA-ISTS dataset mostly comes from real scenes, the contrast between the target and the background is relatively larger, and the segmentation results are more refined.

By and large, the method based on deep learning has obvious advantages in feature representation compared with the traditional methods. Combining the advantages of deep learning and attention model, our method based on game theory achieves a better balance in precision and recall, and achieves the best performance in F1 and IoU. For small targets in IR images with low SCR, our proposed MIM effectively improves the quality of extracted features.

V. CONCLUSION

In ISTS, the IR targets are small in size, weak in energy, and sparse in layout. To solve those problems, we formulate the ISTS into playing a competitive game problem. We transform ISTS into a competitive game problem. Under the guidance of utility function, each player optimizes different pixels. In the continuous game confrontation, the network continues to learn and finally reaches Nash equilibrium. The proposed MIM has good performance in sensing small target signals. Compared with traditional methods and deep learning methods based on combined loss function, our pixelGame achieves a better balance in precision and recall, and the highest F1 and IoU.

For future work, we will try to further improve the efficiency of game optimization, and reduce the inference time. In addition, the introduction of prior knowledge can effectively make up the lacking of IR small target information.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their constructive comments, which are very helpful to improve the quality of this article.

REFERENCES

[1] G. Chen et al., “A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 52, no. 2, pp. 936–953, Feb. 2022.

[2] X. Sun, L. Guo, W. Zhang, Z. Wang, and Q. Yu, “Small aerial target detection for airborne infrared detection systems using LightGBM and trajectory constraints,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 9959–9973, 2021.
[3] Y. Sun, J. Yang, and W. An, “Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 5, pp. 3737–3752, May 2021.

[4] C. Liu, F. Xie, X. Dong, H. Gao, and H. Zhang, “Small target detection from infrared remote sensing images using local adaptive thresholding,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 1941–1952, 2022.

[5] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, “Asymmetric contextual modulation for infrared small target detection,” in Proc. IEEE/CVF Winter Conf. Appl. Comput., pp. 949–959, 2020.

[6] M. Everingham and J. Winn, “The PASCAL visual objects classes challenge 2012 (VOC2012) development kit,” Pattern Anal. Stat. Modelling Comput. Learn., Tech. Rep., vol. 2007, pp. 1–45, 2012.

[7] J.-F. Rivest and R. Fortin, “Detection of dim targets in digital infrared imagery by morphological image processing,” Opt. Eng., vol. 35, no. 7, pp. 1886–1893, 1996.

[8] X. Bai and F. Zhou, “Analysis of new top-hat transformation and the application for infrared small dim target detection,” Pattern Recognit., vol. 43, no. 6, pp. 2145–2156, 2010.

[9] Z. Yan, Y. Xin, L. Liu, R. Su, and D. Chen, “Robust infrared superpixel image separation model for small target detection,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 10256–10268, 2021.

[10] C. Liu, Y. Liu, and Y. Su, “Modified graph Laplacian model with local contrast and consistency constraint for small target detection,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 5807–5822, 2020.

[11] D. Guo, L. Zhu, Y. Lu, H. Yu, and S. Wang, “Small object sensitive segmentation of urban street scene with spatial adjacency between object classes,” IEEE Trans. Image Process., vol. 28, no. 6, pp. 2643–2653, Jun. 2019.

[12] Z. Fan, D. Bi, L. Xiong, S. Ma, L. He, and W. Ding, “Dim infrared image enhancement based on convolutional neural network,” Neurocomputing, vol. 272, pp. 396–404, 2018.

[13] R. Hanaguchi, A. Fujita, K. Nemoto, T. Inaizumi, and S. Hikosaka, “Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery,” in Proc. IEEE Winter Conf. Comput. Vis., 2018, pp. 1442–1450.

[14] H. Wang, L. Zhou, and L. Wang, “Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 8509–8518.

[15] Q. Liu et al., “Multi-task driven feature models for thermal infrared tracking,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 11604–11611.

[16] J. Ma et al., “Loss odyssey in medical image segmentation,” Med. Image Anal., vol. 71, 2021, Art. no. 102035.

[17] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “intU-net: A self-configuring method for deep learning-based biomedical image segmentation,” Front. Neuroimage, vol. 18, no. 2, pp. 203–211, 2021.

[18] M. A. Rahman and Y. Wang, “Optimizing intersection-over-union in deep neural networks for image segmentation,” in Proc. Int. Symp. Vis. Comput., 2016, pp. 234–244.

[19] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3D fully convolutional deep networks,” in Proc. Int. Workshop Mach. Learn. Med. Imag., 2017, pp. 379–387.

[20] S. R. Hashemi, S. M. Salehi, D. Erdogmus, S. P. Prabh, S. K. Warfield, and A. Gholipour, “Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection,” IEEE Access, vol. 7, pp. 1721–1735, 2019.

[21] B. Trosch, Y. Yoo, L. T. Yang, D. K. Li, A. Traboulsee, and R. Tam, “Deep convolutional encoder networks for multiple sclerosis lesion segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2015, pp. 3–11.

[22] S. Yang, J. Kweon, and Y.-H. Kim, “Major vessel segmentation on X-ray coronary angiography using deep networks with a novel penalty loss function,” in Proc. 2nd Int. Conf. Med. Imag. Deep. Learn., 2019, pp. 1–5.

[23] M. Kallel, R. Aboulaich, A. Habbal, and M. Moakher, “A Nash-game approach to joint image restoration and segmentation,” Appl. Math. Modelling, vol. 38, no. 11/12, pp. 3038–3053, 2014.

[24] T. Wang, Q. Sun, Z. Ji, Q. Chen, and P. Fu, “Multi-layer graph constraints for interactive image segmentation via game theory,” Pattern Recognit., vol. 55, pp. 28–44, 2016.

[25] Y.-G. Hsieh, K. Antonakopoulos, and P. Mertikopoulos, “Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium,” in Proc. Conf. Learn. Theory, 2021, pp. 2388–2422.
X. Wang, Z. Peng, D. Kong, P. Zhang, and Y. He, “Infrared dim target detection based on total variation regularization and principal component pursuit,” *Image Vis. Comput.*, vol. 63, pp. 1–9, 2017.

X. Wang, Z. Peng, D. Kong, and Y. He, “Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5481–5493, Oct. 2017.

T. Zhang, Z. Peng, H. Wu, Y. He, C. Li, and C. Yang, “Infrared small target detection via self-regularized weighted sparse model,” *Neurocomputing*, vol. 420, pp. 124–148, 2021.

Y. Sun, J. Yang, M. Li, and W. An, “Infrared small target detection via spatial-temporal infrared patch-tensor model and weighted schatten $p$-norm minimization,” *Infrared Phys. Technol.*, vol. 102, 2019, Art. no. 103050.

P. Zhang, L. Zhang, X. Wang, F. Shen, T. Pu, and C. Fei, “Edge and corner awareness-based spatial-temporal tensor model for infrared small-target detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10708–10724, Dec. 2021.

X. Kong, C. Yang, S. Cao, C. Li, and Z. Peng, “Infrared small target detection via nonconvex tensor fibered rank approximation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5000321.

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

Q. Hou, L. Zhang, F. Tan, Y. Xi, H. Zheng, and N. Li, “ISTDU-Net: Infrared small-target detection U-Net,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7506205.

O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

I. Goodfellow et al., “Generative adversarial nets,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

F. Liu, C. Gao, F. Chen, D. Meng, W. Zuo, and X. Gao, “Infrared small-dim target detection with transformer under complex backgrounds,” 2021, arXiv:2109.14379.

H. Deng, X. Sun, M. Liu, C. Ye, and X. Zhou, “Entropy-based window selection for detecting dim and small infrared targets,” *Pattern Recognit.*, vol. 61, pp. 66–77, 2017.

S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

X. Zhang, Y. Zou, and W. Shi, “Dilated convolution neural network with LeakyReLU for environmental sound classification,” in *Proc. 22nd Int. Conf. Digit. Signal Process.*, 2017, pp. 1–5.

J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free?: Weakly-supervised learning with convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 685–694.

M. Lin, Q. Chen, and S. Yan, “Network in network,” 2013, arXiv:1312.4400.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.