Fluid Dynamics-Inspired Network for Infrared Small Target Detection

Tianxiang Chen, **Qi Chu***, **Bin Liu** and **Nenghai Yu**
School of Cyber Science and Technology, University of Science and Technology of China, China
txchen@mail.ustc.edu.cn, {qchu, flowice, ynh}@ustc.edu.cn

Abstract

Most infrared small target detection (ISTD) networks focus on building effective neural blocks or feature fusion modules but none describes the ISTD process from the image evolution perspective. The directional evolution of image pixels influenced by convolution, pooling and surrounding pixels is analogous to the movement of fluid elements constrained by surrounding variables and particles. Inspired by this, we explore a novel research routine by abstracting the movement of pixels in the ISTD process as the flow of fluid in fluid dynamics (FD). Specifically, a new Fluid Dynamics-Inspired Network (FDI-Net) is devised for ISTD. Based on Taylor Central Difference (TCD) method, the TCD feature extraction block is designed, where convolution and Transformer structures are combined for local and global information. The pixel motion equation during the ISTD process is derived from the Navier–Stokes (N-S) equation, constructing a N-S Refinement Module that refines extracted features with edge details. Thus, the TCD feature extraction block determines the primary movement direction of pixels during detection, while the N-S Refinement Module corrects some skewed directions of the pixel stream to supplement the edge details. Experiments on IRSTD-1k and SIRST demonstrate that our method achieves SOTA performance in terms of evaluation metrics.

1 Introduction

Infrared small target detection (ISTD) has been extensively applied in remote sensing and military tracking system. ISTD is challenging because the targets are always small and ambiguous when surrounded by similar background areas. Besides, infrared images are of low contrast and low SNR. Generic segmentation methods cannot achieve expected performance on this task, so we devote to precise and robust ISTD. ISTD methods can be classified into traditional methods and deep-learning-based methods. In early stages, due to lack of public data set, traditional methods [Sun *et al.*, 2020; Marvasti *et al.*, 2018; Zhang and Peng, 2019; Han *et al.*, 2019; Rivest and Fortin, 1996] take the dominance. However, fully relying on prior knowledge and handcraft features, these methods suffer limited accuracy on images with characteristics inconsistent with model assumptions.

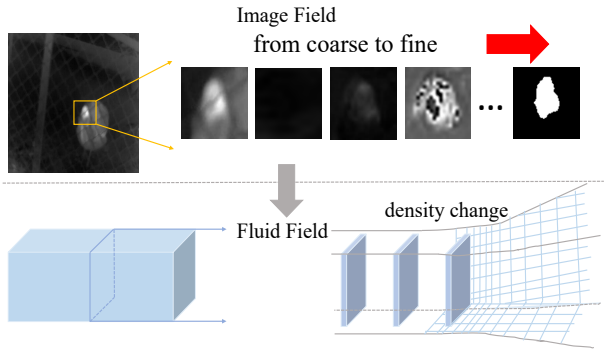

Figure 1: Conversion process between the image field and the fluid field. The upper part represents the feature maps of different layers in the ISTD process, and the yellow arrow indicates the correspondence between the pixels in the original image and the pixels in the output image. The lower part shows the density change of the intersecting surface during fluid movement.

In recent years, deep-learning-based methods have lifted the detection performance by a large margin. These methods can be categorized into CNN-based methods [Dai *et al.*, 2021b; Dai *et al.*, 2021a; Wang *et al.*, 2019; Li *et al.*, 2022a; Zhang *et al.*, 2021; Zhang *et al.*, 2022c; Chen *et al.*, 2023] and transformer-based methods[Wang *et al.*, 2022; Qi *et al.*, 2022; Liu *et al.*, 2021; Zhang *et al.*, 2022a]. Despite different network design concept, these methods mainly focus on devising novel neural blocks or feature fusion modules to adapt to ISTD task, but rarely study the ISTD process from the perspective of image evolution, which is significant for constructing an effective and explainable ISTD network and proposes a potential future research direction.

In fluid dynamics, the fluid elements on the fluid intersecting surface exhibit different density distribution at different time. Similarly, the ISTD process can be perceived as a series of intermediate images that change over time. The intuitive analogy between ISTD and fluid dynamics is presented in Fig. 1. The upper part offers the change of feature

*Corresponding Author.

maps corresponding to the target in the ISTD process, which is to perform from-coarse-to-fine target extraction using the pixel information in the adjacent area. During this process, the edge and shape of targets become more and more clearly visible, and detection results tend to be more precise. The three consecutive images in the upper right part present this process. Specifically, in convolution operations, pixels are jointly determined by multiple adjacent pixels in the previous layer. The lower part describes the density change of a certain intersecting surface in the fluid. The bottom right image shows the density distribution change of the intersecting surface formed by the mutual influence of fluid elements.

In this paper, we explore a novel research routine by abstracting the movement of pixels in the ISTD process as the flow of fluid in fluid dynamics (FD) and propose FDI-Net. On the one hand, most ISTD networks use ResNet as backbone, which is based on one-order Euler method. Intuitively, some mathematical theories (e.g., the definite difference method used for solving FD issues) can be borrowed to devise or improve the ISTD network backbone structure. Therefore, we adopt Taylor Central Difference (TCD) method and propose TCD feature extraction block to locate target pixel from irrelevant background. On the other hand, since small targets in infrared images are of ambiguous shapes and a slight mistake will significantly impact evaluation index, precise detection of edge areas matters to the detection quality of small targets. In fluid dynamics, the Navier–Stokes (N-S) equation is a collection of motion equations describing the momentum conservation of a viscous incompressible fluid and generates a spatial–temporal constraint to guide the flow direction. Thus, we introduce this equation to ISTD and devise N-S Refinement Module to refine the coarse features extracted from TCD block with fine edge details.

Our contributions can be summarized in three folds:

- We are the first to draw an analogy between the dynamic process in the fluid field and the image field in ISTD, where the change of fluid element distribution on the fluid intersecting surface is used to describe the change of pixels in consecutive feature map series. This analogy interprets the changes of microscopic pixels in the ISTD process.
- We propose FDI-Net based on the rule of hydrodynamic fluid element motion that consists of a Taylor Central Difference (TCD) convolution-transformer hybrid structure and an N-S refinement module. Using TCD method, the former structure can serve as the backbone structure to more effectively extract global and local feature. In addition, the latter module converts the spatial information in infrared images into time information and complements the extracted features with finer edge details.
- Our method outperforms others on IRSTD-1k and SIRST in terms of evaluation metrics.

2 Related Work

2.1 Infrared Small Target Detection Networks

Generally speaking, ISTD networks can be classified to two categories: CNN-based and transformer-based networks. CNN-based networks mainly focus on local feature extraction. Dai et al. [Dai *et al.*, 2021a] proposed asymmetric contextual modulation (ACM) for cross-layer information exchange to improve ISTD performance. They also designed AlcNet [Dai *et al.*, 2021b], including a local attention module and a cross-layer fusion module to preserve local features of small targets. Wang et al. proposed MDvsFA [Wang *et al.*, 2019], which applied GAN to ISTD, and achieved a trade-off between miss detection and false alarm. AGPCNet [Zhang *et al.*, 2021] divided the infrared image into patches for better extraction of global and local information and uses cross layer feature fusion to recover lost low-level details. DNANet [Li *et al.*, 2022a] progressively interacted high-level and low-level features. ISNet [Zhang *et al.*, 2022c] designed a simple Taylor finite difference-inspired block and a two-orientation attention aggregation module to detect targets. BAUENet [Chen *et al.*, 2023] introduced uncertainty to ISTD by enhancing contexts with target uncertain area maps.

Only extracting local features is not enough because of ubiquitous target ambiguity caused by low contrast and quality of infrared images. Therefore, some transformer-based methods have been put forward to complement local details with global information by adding transformer to CNN structures. IAANet [Wang *et al.*, 2022] just concatenated local patch outputs from a simple CNN with the original transformer, causing limited feature extraction especially in ambiguous scenarios. RKformer [Zhang *et al.*, 2022a] applied Runge-Kutta method to build coupled CNN-Transformer blocks to highlights infrared small targets and suppresses background interference.

The above ISTD networks focus on building either effective neural blocks or feature fusion modules, none of them describe the ISTD mechanism from the image evolution perspective during the from-coarse-to-fine detection process. In this paper, we open a novel research perspective by abstracting the movement of pixels in the ISTD process as the flow of fluid in fluid dynamics.

2.2 Neural Ordinary Differential Equation

Recently, ODE-inspired networks have gained wide attention since it can deliver higher-accuracy feature extraction performances. Weinan [Weinan, 2017] firstly discovered the link between and discrete ODE and ResNet [He *et al.*, 2016]. Some insightful networks [He *et al.*, 2019; Li *et al.*, 2021; Lu *et al.*, 2019; Zhang *et al.*, 2022b] can also be interpreted from the ODE perspective. To the best of our knowledge, [Zhang *et al.*, 2022c; Zhang *et al.*, 2022a] are the only two works that design ISTD network with ODE theories. From a dynamic system perspective, the two works adopt finite difference method and Runge-Kutta method respectively for network design. Inspired by the two works, we create a new research routine by borrowing the useful idea from the field of FD, where the movement of pixels in the ISTD process is analogous to the flow of fluid. Contrary to the above networks, we devise FDI-Net following the explicit central difference method and Navier-Stokes equation in fluid dynamics, which presents a superb capacity of preserving clear details in ISTD.

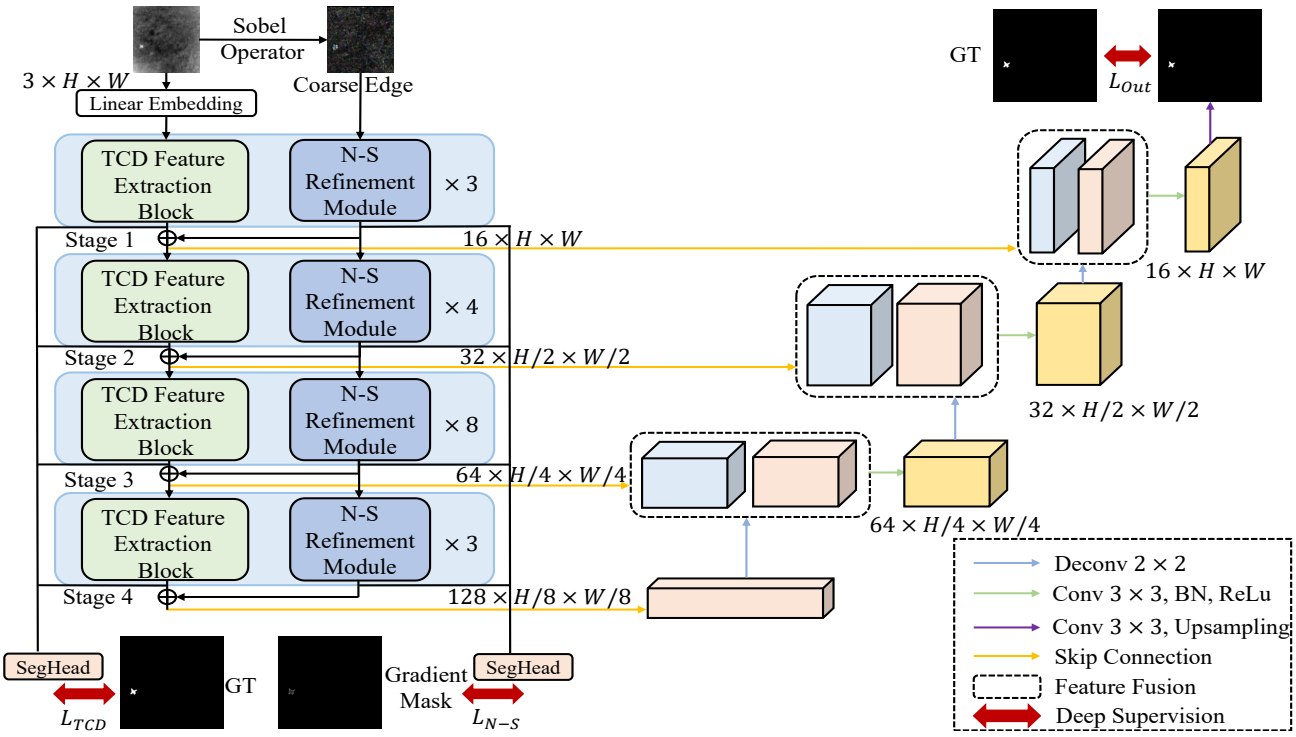

Figure 2: Overall architecture of FDI-Net for ISTD. The core is the encoder, including a TCD Feature Extraction branch (in green color) and a N-S Refinement branch (in blue color). The TCD Feature Extraction block is designed according to Taylor Central Difference method to generally predict target pixels. The N-S Refinement Module completes the conversion of space–time information and refines coarse predictions with edge details.

2.3 Fluid Dynamics and Neural Network

As a branch of mechanics, fluid dynamics studies the interaction and flow laws of the static and moving state of the fluid itself and the relative motion between the fluid and the solid boundary wall under the action of various forces. Typically, fluid micro element is the basic unit in fluid dynamics and its behaviors (fluid motion) are described by the Navier-Stokes equations, which consist of multiple differential equations derived from three basic physical principles: Newton's second law, the law of mass conservation and the law of energy conservation. Common variables like velocity v, pressure p, density ρ and temperature T can be expressed as functions of position (x, y, z) and time t. The Navier-Stokes equations describe the link between the properties of fluid elements over time and space, and thus makes it possible for information to get transformed to different dimensions. Analogously, in the ISTD process pixels in different areas of an infrared image differ in values, whose properties are described by functions of position (x, y) and depth location i of the feature map where the pixel is in the network.

In recent years, researchers have begun to try linking fluid dynamics with neural networks for specific tasks. SPNets framework [Schenck and Fox, 2018] was introduced to integrate fluid dynamics with deep networks for liquid manipulation. Zhang et al.[Zhang *et al.*, 2022b] proposed a second-order finite difference residual network for super resolution. In [Deng *et al.*, 2019], a general super-resolution reconstruction strategy for turbulent velocity fields was proposed using a GAN framework. In [Belbute-Peres *et al.*, 2020], differentiable PDE solvers and GNNs were combined for fluid flow prediction.

Based on the analogy between fluid field and image field, we model the ISTD process to extract more useful feature information and achieve better detection performance. The variables in the Navier-Stokes equation can be determined by the thermodynamic equation $f(\rho, p, t)$ and the material properties of the medium, which can be solved given the initial and boundary conditions. Because this equation has second-order derivative terms, it is difficult to find an exact solution except under some special conditions. Therefore, we turn to numerical solutions and notice that finite difference method is commonly used in solving fluid dynamics problems. Inspired by this, we use high-order central difference method to model the ISTD process.

3 Proposed Method

3.1 Analogy of ISTD and Fluid Dynamics

The paper is inspired by the analogy between the ISTD process and the fluid dynamics process, because the pixel movement during ISTD can be analogous to the movement of fluid flow in fluid dynamics. The analogy can be interpreted from two aspects. From the microscopic view, the fluid element studied in fluid dynamics as the basic unit moves under the joint constraints of temperature, pressure, and surrounding fluid elements. The image pixel changes in a targeted man-

ner and gets influenced by basic operations such as convolution, pooling and some surrounding pixels. From the macroscopic perspective, the distribution of an intersecting surface in fluid changes regularly over time. In the ISTD process, after multi-step feature extraction and image reconstruction, the 2-D feature maps change towards the final prediction where only target pixels are highlighted.

The particle movement processes in the fluid field and the image field are described respectively as follows. In fluid dynamics, interaction forces exist between particles, so each particle moves under the combined constraints of internal and external forces. For each particle, forces exist in pairs, such as friction, pressure and viscosity, so each particle is influenced by its surrounding particles. The ISTD process is similar, with each pixel depending on other pixels in its vicinity. Therefore, we can borrow the basic theory of fluid dynamics to devise new ISTD network structure. Present ISTD algorithms use 2-D convolution kernels for filtering. The value of each pixel is updated based on the products of its neighboring pixels and the convolution kernel weights. Different image filtering effects, for example edge extraction, can be achieved by applying corresponding convolution kernels. The pixel movement under various convolution operations and loss functions finally leads to the final detected result with clear edges. The ISTD process between the original image and the final result can be interpreted as interpolating the target pixels (i.e., movement of pixels) using surrounding pixels while generating clear edges. The inherent structure of edges serves as a constraint on the pixel movement process.

The consistency of the particle motion law in image and fluid fields inspires us to use the fluid dynamics theory for novel ISTD network construction. Considering that the problems in fluid dynamics are continuity problems, the exact solution to related equations is difficult to find, so we turn to numerical solver and choose Taylor Central Difference (TCD) method with third-order accuracy to build TCD feature extraction block to preliminarily locate target areas from ambiguity. Besides, the N-S equation in fluid dynamics builds the relationship between fluid element properties over time and space. We can apply the N-S equation to the ISTD process to convert the space information of image pixels into time dimension, and this can bring edge refinement to the TCD body part extraction result and achieve better detection performance.

3.2 Overall Architecture

We propose our FDI-Net based on Taylor Central Difference method and the Navier-Stokes equation for ISTD. As shown in Fig. 2, the encoder contains a backbone network composed of a cascaded TCD feature extraction branch (Section 3.3) and a parallel N-S Refinement module (Section 3.4). In the backbone, the TCD feature extraction branch generally determines the primary location of target pixels, which helps distinguishing ambiguous targets from irrelevant background interference. Besides, we devise a parallel N-S Refinement module to refine coarse predictions from TCD blocks with fine edge details. The two branch features at each stage are added before sent to decoder for upsampling by skip connections. Sobel operator is applied to get the initial coarse edge map of input image before fed into the N-S branch and the gradient value of the masks are used as labels to deep supervise this branch to guarantee target edges are explored. Specific losses (Section 3.6) are adopted to supervise the generation of TCD and N-S side outputs and the final prediction.

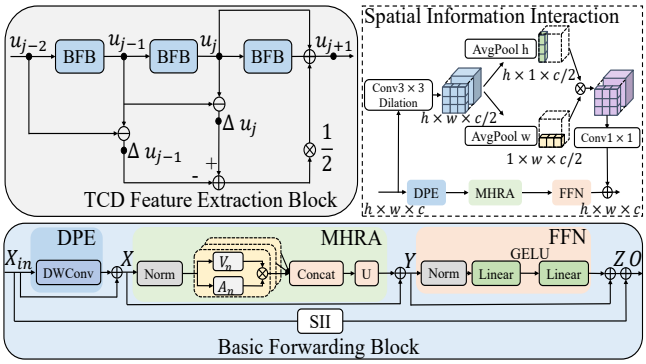

Figure 3: The structure of the TCD Feature Extraction block, which embodies the TCD solver of ODE by learning the feature change using cascaded convolutional layers. See Section 3.2 for details.

3.3 TCD Feature Extraction Block

Due to the low quality and low contrast of infrared images, in many cases targets are in ambiguity and not easy to get distinguished from surroundings, leading to many miss detection and false alarms. To tackle this issue, we propose a specific residual block to explore effective target features from irrelevant background interference. To this end, the residual block is implemented as an ODE solver based on an efficient numerical scheme, Taylor Central Difference (TCD), i.e., we devise a basic TCD feature extraction block to serve as part of the backbone network. We choose TCD due to its better accuracy in solving ODE than Euler method [Anderson and Wendt, 1995].

To further describe target pixels from ambiguity, spatial information is required to eliminate uncertainty. Hence, we also design the Spatial Information Integration (SII) across the basic Uniformer [Li *et al.*, 2022b] block to constitute basic forwarding block (BFB) and embody it into our TCD feature extraction block to further promote information exchange while encoding more precise spatial information. Specifically, BFB consists of four key modules: Dynamic Position Embedding (DPE), Multi-Head Relation Aggregator (MHRA), Feed-Forward Network (FFN) and Spatial Information Integration (SII):

$$\begin{aligned} X &= DPE(X_{in}) + X_{in} \\ Y &= MHRA(Norm(X) + X \\ Z &= FFN(Norm(Y) + Y \\ O &= Z + SII(X_{in}) \end{aligned} \tag{1}$$

where X_{in} is the input token tensor. SII mines the relationship between pixels, not just patch tokens, by introducing attention in two spatial dimensions, making our TCD feature extraction block more suitable for ISTD task. The structures of SII and TCD feature extraction block are illustrated in Fig. 4.

Form	Object of study	Method
Integral*	Finite control volume	Eulerian
Integral	Finite control volume	Lagrangian
PDE*	Fluid element	Eulerian
PDE	Fluid element	Lagrangian

Table 1: Four forms of continuity equation in fluid dynamics, where "Integral" and "PDE" mean Integral equation and Partial differential equation, and the form with or without * denote the equation is in conservative form or not.

Specifically, we discretize the ODE using the Taylor central finite difference equations with third-order accuracy, i.e.,

$$(\frac{\partial u}{\partial x})_j = \frac{2u_{j+1} - 3u_j + 2u_{j-1} - u_{j-2}}{2\Delta x} \tag{2}$$

where u is a target depending on the input variable x. The above formula can be rewritten as:

$$(\frac{\partial u}{\partial x})_j \Delta x = u_{j+1} - \frac{3}{2}u_j + u_{j-1} - \frac{1}{2}u_{j-2} \tag{3}$$

The left side of the equation can be defined as the change of u from step $j-2$ to step $j+1$, which can be approximated with a specific block, denoted as $\delta f(x)$. Therefore, the above equation can be rewritten as:

$$u_{j+1} = u_j + \frac{1}{2}\Delta u_j - \frac{1}{2}\Delta u_{j-1} + \delta f(x) \tag{4}$$

In this paper, we design the TCD feature extraction block to implement Equation (4), where $\delta f(x)$ is implemented as our proposed basic forwarding block.

Perceiving the pixels of the infrared image as a collection of points following a certain distribution, the proposed TCD structure can determine the primary movement direction of pixels, leading the diffused points to the main body areas of small targets. However, during this process, the pixels cannot avoid a certain degree of dispersion, causing uncertainty in edge areas in the final prediction. To address this problem, additional constraints should be applied to the ISTD process so that edge information can be enriched. Thereby, we consider devising an edge module from the perspective of dynamic pixel-value distribution to refine the residual extraction process with finer edge details.

3.4 N-S Refinement Module

The continuity equation, which describes the transport behavior of a conserved quantity, is included in N-S equations in fluid dynamics. Continuity equation can be divided into four forms as shown in Table 1 depending on different object of study and computation method. Since each of these forms can be mutually converted, we only consider the PDE in conservative form

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \vec{v}) = 0 \tag{5}$$

where ρ denotes the fluid density, $\vec{v}$ is the velocity vector, and $\vec{v} = (u\vec{i}, w\vec{j})$. According to the multiplication law of the Hamiltonian operator, equation (5) can be rewritten as

$$\frac{\partial \rho}{\partial t} + \rho \nabla \cdot \vec{v} + \vec{v} \nabla \rho = 0 \tag{6}$$

During ISTD process, the extracted feature map will change as the network goes deep, and this process can be viewed as the change on an intersecting surface of 3-D fluid. Therefore, we can transfer the motion equation in the fluid field to the 2-D image domain. Equation (5) can be rewritten as follows:

$$\frac{\partial \rho}{\partial t} + \rho(\frac{\partial u}{\partial x} + \frac{\partial w}{\partial y}) + u\frac{\partial \rho}{\partial x} + w\frac{\partial \rho}{\partial y} = 0 \tag{7}$$

Where the divergence change of $\vec{v}$ can be understood as the change of pixel number per unit image area. In the encoder stage, there is no fusion of images, which means that pixels do not flow in or out. Therefore, we assume that $\frac{\partial u}{\partial x} = 0$ and $\frac{\partial w}{\partial y} = 0$, meaning that the partial derivatives of two velocity components u and w along their corresponding directions are 0. Then, equation (7) is converted to

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + w\frac{\partial \rho}{\partial y} = 0 \tag{8}$$

In the image field, ρ stands for the pixel value, t stands for the layer depth of the feature map where the pixel is, and $\frac{\partial \rho}{\partial x}$ and $\frac{\partial \rho}{\partial y}$ are the gradients of the pixel value in different directions. Specifically, the gradients of edge areas are larger, while for smooth areas the gradients are smaller. The objective variable in equation (8) is $\frac{\partial \rho}{\partial t}$, we convert this variable to the difference form as follows

$$\rho_i = \rho_{i-1} - u\frac{\partial \rho_i}{\partial x} - w\frac{\partial \rho_i}{\partial y} \tag{9}$$

Equation (9) builds the relationship between the change of pixel value in time and spatial domains during feature extraction. We choose a residual block layer to calculate the partial derivative term and rewrite equation (9) as

$$F_i = -u\frac{\partial \rho_i}{\partial x_i} - w\frac{\partial \rho_i}{\partial y_i} = -\vec{v}\nabla\rho_i \tag{10}$$

On the right side of equation (10), $\vec{v}$ stands for the forward extraction direction, $\nabla\rho_i$ is the partial derivatives of pixel values in the x-and y-directions, which is the gradient term. As we know, the gradients of edge areas are larger, while for smooth areas the gradients are smaller. Thus, we place basic residual block to N-S Refinement Module to extract edge feature, the direction of which serves as the $\vec{v}$ term; and we use the Sobel operator to explore the gradients of each infrared image, which serve as the $\nabla\rho_i$ term. Combining the residual block and Sobel operator together constitutes our N-S Refinement Module. Inspired by the N-S control equation, the N-S Refinement Module converts the space information (along the x, y directions) to time information (along the i direction), thereby refining coarse TCD block predictions with edge details.

3.5 Overall Structure of FDI-Net

Combining the N-S Refinement branch with the TCD Feature Extraction structure, we can obtain the final expression of the overall feature extraction of the encoder as

$$u_{j+1} = u_j + \frac{1}{2}\Delta u_j - \frac{1}{2}\Delta u_{j-1} + \delta f(x) + NS(u_j) \tag{11}$$

where $NS(u_j)$ denotes the N-S Refinement function defined in equation (10). The final prediction is obtained after the decoder. Therefore, the final training objective is

$$\dot{\theta} = \underset{\theta}{argmin} \sum_{i=1}^{N} ||F_{FDI}(I) - GT||_1 \tag{12}$$

where we set I as the input and GT as the ground truth, and θ refers to all the learnable parameters in the proposed FDI-Net.

3.6 Loss Function

ISTD is regarded as a pixel-level classification problem, with each pixel in the image classified as target or background. Generally, the problem is solved by using binary cross-entropy loss. However, the pixels of targets only account for a small proportion, which means that the pixel distributions of target and background are very uneven. In this case, if the loss function treats target pixels and background pixels equally, the model constrained by the loss function is optimized toward background categories to fall into the local optima. Furthermore, the pixel accuracy of target is reduced under the constraint of the loss function. To address sample imbalance, we tailor a hybrid loss L_{hyb} for ISTD based on binary cross-entropy loss and Dice loss [Sudre *et al.*, 2017].

$$L_{hyb} = L_{BCE} + \lambda L_{Dice} \tag{13}$$

where λ is a coefficient to balance between L_{BCE} and L_{Dice} and is set to 1. From equation (13), the similarity between the predicted result and the label only determines L_{Dice}, which is independent of the number of training samples and can mitigate the sample imbalance problem. However, Dice loss has strong gradient changes, which causes unstable training. Considering the above, we design a new loss function L_{hyb}, which not only maintains gradient stability but also prevents falling into the local optimum.

Our total loss function L_{Total} includes L_{Out} as the main loss function and TCD interior area loss (L_{TCD}) and N-S edge area loss (L_{N-S}) as two auxiliary loss functions. All of these losses are in our hybrid loss form.

$$L_{Total} = L_{Out} + L_{TCD} + L_{N-S} \tag{14}$$

4 Experiment

4.1 Datasets, Metrics and Implementation Details

We choose IRSTD-1k and SIRST as benchmarks. IRSTD-1k consists of 1,000 real infrared images of 512×512 in size, containing various kinds of small targets and scenes. SIRST contains 427 infrared images.

For evaluation metrics, since there has not been any single set of evaluation metrics get widely accepted for ISTD task, we choose the most commonly used IoU and $nIoU$ (Normalized Intersection over Union) as our evaluation metrics, which are defined as:

$$\begin{aligned} IoU &= \frac{A_i}{A_u} \\ nIoU &= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{TP[i]}{T[i] + P[i] - TP[i]}\right) \end{aligned} \tag{15}$$

Method	SIRST		IRSTD-1k	
	IoU	nIoU	IoU	nIoU
Tophat	7.14	5.20	10.06	7.44
PSTNN	22.40	22.35	24.57	17.93
MSLSTIPT	10.30	9.58	11.43	5.93
MDvsFA	60.30	58.26	49.50	47.41
ACM	72.33	71.43	60.97	58.02
AlcNet	74.31	73.12	62.05	59.58
DNANet	74.88	73.81	63.00	53.2
IAANet	75.31	74.65	59.93	57.53
UIUNet	78.25	75.15	64.97	63.78
RKformer	77.24	74.89	64.12	64.18
ISNet	80.02	78.12	68.77	64.84
FDI-Net (Ours)	**81.86**	**81.08**	**72.01**	**71.99**

Table 2: Quantitative results of different methods on SIRST and IRSTD-1k in terms of IoU(%), nIoU(%).

where A_i and A_u are the areas of intersection region and union region. N is the total number of samples, $TP[.]$ is the number of true positive pixels, $T[.]$ and $P[.]$ denotes the number of ground truth and predicted positive pixels. The algorithm is implemented in Pytorch, with Adaptive Gradient (AdaGrad) as optimizer, initial learning rate set to 0.05 and weight decay coefficient to 0.0004. A Titan Xp GPU is used for training, with batch size set to 4. Training on SIRST and IRSTD-1k takes 600 epochs and 400 epochs respectively.

4.2 Quantitative Results

We select some most well-performed ISTD methods for comparison. Among them, Tophat [Rivest and Fortin, 1996], PSTNN [Zhang and Peng, 2019] and MSLSTIPT [Sun *et al.*, 2020] are traditional methods; MDvsFA [Wang *et al.*, 2019], ACM [Dai *et al.*, 2021a], AlcNet [Dai *et al.*, 2021b], DNANet [Li *et al.*, 2022a], UIUNet [Wu *et al.*, 2022] and ISNet [Zhang *et al.*, 2022c] are CNN-based methods; IAANet [Wang *et al.*, 2022] and RKformer [Zhang *et al.*, 2022a] are transformer-based methods. The results of DNANet, IAANet, UIUNet are implemented by ourselves with source codes, the rest are referred from [Zhang *et al.*, 2022c]. As shown in Table 1, the proposed FDI-Net achieves the SOTA performance in terms of all the evaluation metrics compared with all other methods on both benchmarks in IoU and $nIoU$. Traditional methods perform poorly for highly depending on hand-crafted features. Nevertheless, most CNN-based and transformer-based methods lay insufficient emphasis on target edges, causing lower IoU and $nIoU$. Our FDI-Net delivers promising results owing to the designed TCD block to effectively extract useful features from disturbance and the N-S refinement module to further refine extracted features with edge details.

4.3 Visual Results

Some visual results obtained by different methods and intermediate branch outputs on two benchmarks are shown in Fig. 4. As can be seen, even in low contrast and ambiguous scenarios, our FDI-Net can not only accurately locate the targets but also segment out complete and precise target shapes.

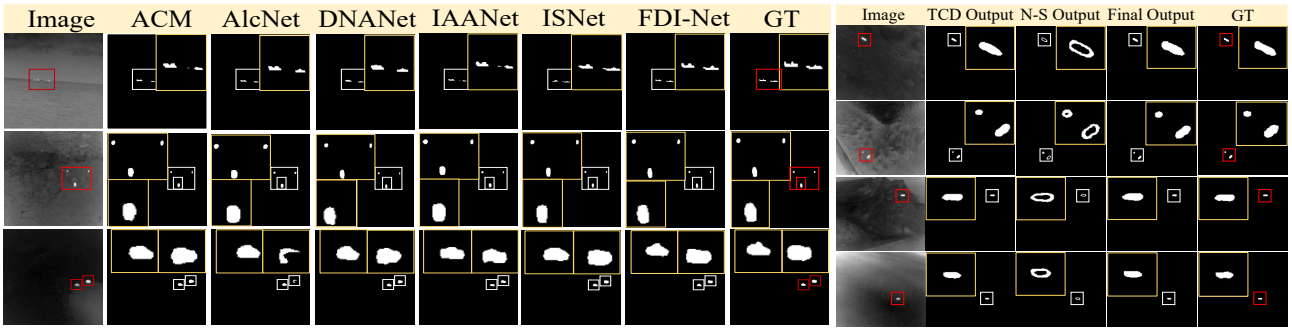

Figure 4: Result visualization of (a) different ISTD methods (b) intermediate branch outputs of FDI-Net. Closed-up views are shown in yellow boxes. Red boxes and white boxes locate GT targets and detected targets respectively.

Method	SIRST		IRSTD-1k	
	IoU	nIoU	IoU	nIoU
UNet	68.85	68.12	57.31	56.26
UNet+TCD	76.53	75.49	65.87	65.71
UNet+NS	75.26	73.98	64.18	63.98
UNet+TCD+NS	**81.86**	**81.08**	**72.01**	**71.99**

Table 3: Ablation study of TCD Feature Extraction Block and N-S Refinement Module in IoU(%), nIoU(%).

Method	SIRST		IRSTD-1k	
	IoU	nIoU	IoU	nIoU
ResBlock	74.67	74.02	61.44	59.86
UniFormer Block	77.83	76.95	64.77	63.62
RK2+BFB	78.66	77,85	69.42	68.56
TCD Block w/o SII	78.24	77.39	69.18	67.98
TCD Block	**81.86**	**81.08**	**72.01**	**71.99**

Table 4: Ablation study of TCD Feature Extraction Block and different block designs in IoU(%), nIoU(%).

Method	SIRST		IRSTD-1k	
	IoU	nIoU	IoU	nIoU
Canny+ResBlock	78.36	77.78	69.41	68.78
Sobel+Conv/RL/Conv	76.84	76.23	69.02	68.23
Sobel+ResBlock (Ours)	**81.86**	**81.08**	**72.01**	**71.99**

Table 5: Ablation study of N-S Refinement Module in IoU(%), nIoU(%).

4.4 Ablation Study

To investigate the effectiveness of each component in our FDI-Net, we perform several ablation studies on SIRST and IRSTD-1k dataset. The ablation study results of TCD Feature Extraction Block (TCD) and N-S Refinement Modules (NS) are shown in Table 3. As can be seen, each of them enhance the performance of the UNet baseline and combining both of them brings the best results, implying that they are complementary to each other. This is because the TCD branch plays a diffusion role to guide the primary movement of pixels for locating target main areas; the N-S module constrains the diffusion process to curtail dispersion so that target edge information can be enriched.

The ablation study of TCD Feature Extraction Block is shown in Table 4, where our TCD Block outperforms other block designs by a large margin, showing its superiority in distinguishing target pixels from ambiguity. RK2+BFB means combining second-order Runge-Kutta method with BFB to substitute TCD Block. The ablation also demonstrates the effectiveness of TCD method and our SII tailored for ISTD.

The ablation study of N-S Refinement Module is shown in Table 5. During data pre-processing, the Sobel operator is used to extract the coarse edge of the target from the input image to serve as the gradient term in equation (10). We try replacing ResBlock with simple Conv/ReLu/Conv block and the Sobel operator with other the Canny operator. Results show that the combination of the Sobel operator and ResBlock delivers the best result.

5 Conclusion

Motivated by the analogy of pixel movement during ISTD process and flow of fluid in fluid dynamics, we propose FDI-Net for ISTD. Specifically, we design a TCD Feature Extraction Block based on central difference method and a N-S Refinement Module based on Navier-Stokes equations in fluid dynamics. The TCD block guides the primary movement of pixels like diffusion to locate target main areas while the N-S module constrains the diffusion process to curtail dispersion to refine main area features of targets with edge details. Experiments on SIRST and IRSTD-1k validate the effectiveness of our network.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62272430, No. U20B2047) and the Fundamental Research Funds for the Central Universities.

References

[Anderson and Wendt, 1995] John David Anderson and John Wendt. *Computational fluid dynamics*, volume 206. Springer, 1995.

[Belbute-Peres *et al.*, 2020] Filipe De Avila Belbute-Peres, Thomas Economon, and Zico Kolter. Combining differentiable pde solvers and graph neural networks for fluid flow prediction. In *international conference on machine learning*, pages 2402–2411. PMLR, 2020.

[Chen *et al.*, 2023] Tianxiang Chen, Qi Chu, Zhentao Tan, Bin Liu, and Nenghai Yu. Bauenet: Boundary-aware uncertainty enhanced network for infrared small target detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Dai *et al.*, 2021a] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 950–959, 2021.

[Dai *et al.*, 2021b] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11):9813–9824, 2021.

[Deng *et al.*, 2019] Zhiwen Deng, Chuangxin He, Yingzheng Liu, and Kyung Chun Kim. Super-resolution reconstruction of turbulent velocity fields using a generative adversarial network-based artificial intelligence framework. *Physics of Fluids*, 31(12):125111, 2019.

[Han *et al.*, 2019] Jinhui Han, Sibang Liu, Gang Qin, Qian Zhao, Honghui Zhang, and Nana Li. A local contrast method combined with adaptive background estimation for infrared small target detection. *IEEE Geoscience and Remote Sensing Letters*, 16(9):1442–1446, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2019] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019.

[Li *et al.*, 2021] Bei Li, Quan Du, Tao Zhou, Shuhan Zhou, Xin Zeng, Tong Xiao, and Jingbo Zhu. Ode transformer: An ordinary differential equation-inspired model for neural machine translation. *arXiv preprint arXiv:2104.02308*, 2021.

[Li *et al.*, 2022a] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 2022.

[Li *et al.*, 2022b] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022.

[Liu *et al.*, 2021] Fangcen Liu, Chenqiang Gao, Fang Chen, Deyu Meng, Wangmeng Zuo, and Xinbo Gao. Infrared small-dim target detection with transformer under complex backgrounds. *arXiv preprint arXiv:2109.14379*, 2021.

[Lu *et al.*, 2019] Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

[Marvasti *et al.*, 2018] Fereshteh Seyed Marvasti, Mohammad Reza Mosavi, and Mahdi Nasiri. Flying small target detection in ir images based on adaptive toggle operator. *IET Computer Vision*, 12(4):527–534, 2018.

[Qi *et al.*, 2022] Meibin Qi, Liu Liu, Shuo Zhuang, Yimin Liu, Kunyuan Li, Yanfang Yang, and Xiaohong Li. Ftc-net: Fusion of transformer and cnn features for infrared small target detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:8613–8623, 2022.

[Rivest and Fortin, 1996] Jean-Francois Rivest and Roger Fortin. Detection of dim targets in digital infrared imagery by morphological image processing. *Optical Engineering*, 35(7):1886–1893, 1996.

[Schenck and Fox, 2018] Connor Schenck and Dieter Fox. Spnets: Differentiable fluid dynamics for deep neural networks. In *Conference on Robot Learning*, pages 317–335. PMLR, 2018.

[Sudre *et al.*, 2017] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[Sun *et al.*, 2020] Yang Sun, Jungang Yang, and Wei An. Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):3737–3752, 2020.

[Wang *et al.*, 2019] Huan Wang, Luping Zhou, and Lei Wang. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8509–8518, 2019.

[Wang *et al.*, 2022] Kewei Wang, Shuaiyuan Du, Chengxin Liu, and Zhiguo Cao. Interior attention-aware network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[Weinan, 2017] E Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

[Wu *et al.*, 2022] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Uiu-net: U-net in u-net for infrared small object detection. *IEEE Transactions on Image Processing*, 2022.

[Zhang and Peng, 2019] Landan Zhang and Zhenming Peng. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sensing*, 11(4):382, 2019.

[Zhang *et al.*, 2021] Tianfang Zhang, Siying Cao, Tian Pu, and Zhenming Peng. Agpcnet: Attention-guided pyramid context networks for infrared small target detection. *arXiv preprint arXiv:2111.03580*, 2021.

[Zhang *et al.*, 2022a] Mingjin Zhang, Haichen Bai, Jing Zhang, Rui Zhang, Chaoyue Wang, Jie Guo, and Xinbo Gao. Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1730–1738, 2022.

[Zhang *et al.*, 2022b] Mingjin Zhang, Qianqian Wu, Jing Zhang, Xinbo Gao, Jie Guo, and Dacheng Tao. Fluid micelle network for image super-resolution reconstruction. *IEEE Transactions on Cybernetics*, 2022.

[Zhang *et al.*, 2022c] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. Isnet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 877–886, 2022.