TCI-Former: Thermal Conduction-Inspired Transformer for Infrared Small Target Detection

Tianxiang Chen[1,2,3*], **Zhentao Tan**[1,2,3], **Qi Chu**[1,3†], **Yue Wu**[2], **Bin Liu**[1,3], **Nenghai Yu**[1,3]

[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]Alibaba Group
[3]Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences
{txchen,tzt}@mail.ustc.edu.cn, matthew.wy@alibaba-inc.com, {qchu, flowice, ynh}@ustc.edu.cn

Abstract

Infrared small target detection (ISTD) is critical to national security and has been extensively applied in military areas. ISTD aims to segment small target pixels from background. Most ISTD networks focus on designing feature extraction blocks or feature fusion modules, but rarely describe the ISTD process from the feature map evolution perspective. In the ISTD process, the network attention gradually shifts towards target areas. We abstract this process as the directional movement of feature map pixels to target areas through convolution, pooling and interactions with surrounding pixels, which can be analogous to the movement of thermal particles constrained by surrounding variables and particles. In light of this analogy, we propose Thermal Conduction-Inspired Transformer (TCI-Former) based on the theoretical principles of thermal conduction. According to thermal conduction differential equation in heat dynamics, we derive the pixel movement differential equation (PMDE) in the image domain and further develop two modules: Thermal Conduction-Inspired Attention (TCIA) and Thermal Conduction Boundary Module (TCBM). TCIA incorporates finite difference method with PMDE to reach a numerical approximation so that target body features can be extracted. To further remove errors in boundary areas, TCBM is designed and supervised by boundary masks to refine target body features with fine boundary details. Experiments on IRSTD-1k and NUAA-SIRST demonstrate the superiority of our method.

Introduction

Infrared small target detection (ISTD) is challenging because targets are so small that may easily get ignored by generic segmentation networks. Besides, infrared images are of low contrast and low quality, which also bring challenges to this task. Since generic segmentation networks fail to perform well on this task, we hope to explore a new perspective and design a precise and explainable method for ISTD.

ISTD methods are generally categorized into traditional methods and deep-learning-based methods. In early stages, for lack of public ISTD dataset, researchers are limited to traditional methods (Sun, Yang, and An 2020; Marvasti, Mosavi, and Nasiri 2018; Zhang and Peng 2019; Han et al. 2019). However, these methods relying so much on prior knowledge and handcraft features that inevitably suffer very limited performances on images with characteristics inconsistent with the model assumptions.

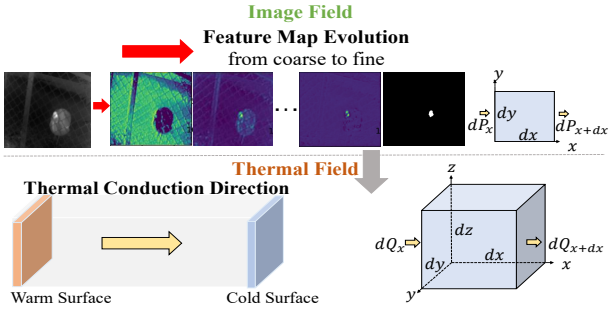

Figure 1: Conversion process between the image field and the thermal field.The feature map evolution process in the image field can be analogous to the thermal conduction process in the thermal field. The upper part depicts the image field and presents the from-coarse-to-fine feature map evolution in the ISTD process. The upper right corner shows the change of pixel value in a 2-D image micro-element. The lower part shows the thermal conduction process of a 3-D micro-element in the thermal field, where thermal energy is conducted spontaneously from high-temperature areas to low-temperature areas.

Recent years have witnessed the research focus of ISTD shifting to deep-learning-based methods. Deep-learning-based methods improve the ISTD performance by a large margin and can be further classidied into CNN-based methods (Chen et al. 2023c; Dai et al. 2021b,a; Zhang et al. 2021; Wang, Zhou, and Wang 2019; Li et al. 2022a; Zhang et al. 2021, 2022d; Zhu et al. 2023; Weng et al. 2023; Du, Wang, and Cao 2023) and hybrid methods (methods combining ViT and CNN) (Wang et al. 2022a; Qi et al. 2022; Chen et al. 2023b; Liu et al. 2021; Zhang et al. 2022a; Chen et al. 2023a). However, despite different module designs, these methods rarely explore a new perspective to look at ISTD, which helps constructing an explainable ISTD network and proposing a potential future research direction. To this end, we propose to understand the feature map evolution process of ISTD from the perspective of thermal conduction.

*Work done during an internship at Alibaba Group.
†Qi Chu is the corresponding author.

In thermodynamics, micro-elements with different heat exhibit different distribution over time in a closed system. Influenced by the heat source and other external factors, heat will spontaneously be conducted from warm areas to cold areas. Similarly, the ISTD process can be regarded as a series of feature maps that change over time constrained by an objective function. The intuitive analogy between ISTD and thermal dynamics is shown in Fig. 1. The upper part shows the feature map evolution process during ISTD, which is a from-coarse-to-fine process gradually highlighting targets using the adjacent pixel information. Specifically, in convolution operations, pixels are determined by multiple adjacent pixels of the previous layer. During ISTD process, the micro-elements with different pixel values in the image move under the constraints of an objective function until some micro-elements with high pixel values gather near small target areas. In this way, the small targets gradually get highlighted. The three consecutive images in the upper right part visualize this process. The lower part describes the spontaneous thermal conduction from high-temperature to low-temperature areas. The bottom right image shows the inflow and outflow of the thermal energy in a 3-D micro-element. The two processes are essentially very similar, so some thermodynamic theories can be transferred to ISTD. The most related work of our paper is (Zhang et al. 2022b), which understands super resolution from thermodynamics perspective, but our task, network modules and the way of analogizing thermal field to image field (the pixel movement of ISTD is directed to target areas, but for super resolution it is unordered) are all different.

In this paper, we explore a novel research routine by analogizing the pixel movement during ISTD process as thermal conduction in thermodynamics and propose TCI-Former. Based on the thermal conduction differential equation, we derive the pixel movement differential equation (PMDE) in the image domain for ISTD. Our PMDE builds a spatial–temporal constraint to guide the pixel flow direction, so we design our network based on it. On the one hand, we apply the finite difference method to PMDE and propose thermal conduction-inspired attention (TCIA) to help extracting the main body features of targets. On the other hand, only focusing on main body areas of targets inevitably causes errors in segmenting target boundary areas, so we devise thermal conduction boundary module (TCBM) to refine target body features with fine boundary details.

Our contributions can be summarized in three folds:

- We are the first to realize the intrinsic consistency between thermal micro-elements and the image pixels during feature map evolution in ISTD, where the change of heat distribution over time is analogous to the change of pixel values due to pixel movement in consecutive feature map series. We transfer heat conduction theories into the ISTD network design and propose TCI-Former.
- Inspired by the thermal conduction differential equation, we derive our pixel movement differential equation (PMDE) to establish a link between spatial and temporal information of pixel values during ISTD process.
- We incorporate the finite difference method to PMDE and propose thermal conduction-inspired attention (TCIA) to extract target main body features but brings slight errors to target boundary areas. As complement, thermal conduction boundary module (TCBM) is also devised to supplement the target body features with fine boundary details to make up for the errors.
- Our method outperforms others on IRSTD-1k and NUAA-SIRST in terms of evaluation metrics.

Related Work

Infrared Small Target Detection Networks

ISTD networks are generally classified into CNN-based and hybrid types. CNN-based networks mainly extract local features. Dai et al. (Dai et al. 2021a) released the first public ISTD dataset and proposed asymmetric contextual modulation for cross-layer feature fusion. They then proposed AlcNet (Dai et al. 2021b) to preserve local features of small targets. Wang et al. were the first to apply GAN to ISTD and proposed MDvsFA (Wang, Zhou, and Wang 2019), which achieved a trade-off between missed detection and false alarm. DNANet (Li et al. 2022a) devised a dense nested interactive module (DNIM) to progressively interact different level features. ISNet (Zhang et al. 2022d) designed a simple Taylor finite difference-inspired block and a two-orientation attention aggregation module to detect targets.

However, only local features are insufficient to detect all infrared targets because the low contrast background makes many small targets unclear to find. Therefore, researchers turn to hybrid methods (Chen, Wang, and Tan 2022; Wang et al. 2022a; Zhang et al. 2022a) by combining ViT with CNN to complement local features with global dependencies. For example, Chen et al. novelly built a ViT-CNN structure based on fluid dynamics for shape-aware ISTD.

The above ISTD networks focus on building either feature extraction blocks or fusion modules, none of them provide a new understanding of ISTD from the feature map evolution perspective. In this paper, we open a novel research perspective by abstracting the directional movement of pixels with high pixel values to target areas in the ISTD process as heat conduction from warm to cold areas in thermodynamics.

Thermal Conduction Differential Equation

Thermal conduction studies the law of thermal energy transfer due to temperature difference. Wherever there exists a temperature difference, there is a spontaneous conduction of thermal energy from a high-temperature object to a low-temperature object, or from a high-temperature object part to a low-temperature part (Borgnakke and Sonntag 2022).

As the basic law of thermal conduction, thermal conduction differential equation indicates that the heat passing through a given section in unit time is proportional to the rate of temperature change and the area of the section perpendicular to the direction of the section. It is the mathematical expression of the differential form of the temperature distribution in the thermal conduction temperature field. The thermal conduction direction is opposite to the temperature increase direction. The equation is established according to the heat conservation law and Fourier law. The law of heat

conservation can be expressed as $\Delta Q = \Delta E + Q_f$, where Q is the difference between the thermal energy imported and exported from an object. ΔE is the increment of internal energy of the object. Q_f is the heat of formation of the internal heat source in the object. The Fourier law describes the relationship between thermal conductivity and temperature gradient, which is described as $q = -\lambda \frac{\partial T}{\partial n}$, where $\frac{\partial T}{\partial n}$ is the temperature gradient and λ is the thermal conduction coefficient. Rewrite heat conservation equation into the differential form of unit time and space and plug the Fourier Law into the heat conservation equation, we can get the thermal conduction differential equation as follows:

$$\frac{\partial T}{\partial t} = \frac{\lambda}{\rho c}(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}) + \frac{q_v}{\rho c}, \tag{1}$$

where q_v is the heat of formation of the internal heat source in an object in per unit volume and time.

Pixel Movement Differential Equation (PMDE)

In a unit of time, the thermal change of a micro-element can be expressed as: [the difference between the imported and exported heat] + [the thermal energy generated by the internal heat source] = [the increase in thermodynamic energy]. The difference between the imported and exported heat corresponds to the feature map pixel value difference between inflow and outflow (ΔP_f). The thermal energy generated by the internal heat source corresponds to the change in the pixel's own value (ΔP_s). The total increase in thermodynamic energy corresponds to the overall change in pixel value (ΔP). Accordingly, in the image field we have:

$$[\Delta P_f] + [\Delta P_s] = [\Delta P]. \tag{2}$$

Similar to the derivation of TCDE, the Pixel Movement Differential Equation (PMDE) can be derived as follows.

Pixel Value Difference between Inflow and Outflow

Within dt, we denote the pixel values flowing into the micro-element along the x-axis and y-axis as dP_x and dP_y, respectively. Similarly, there are also pixel values flowing out of the micro-element along both axes, which we describe as dP_{x+dx} and dP_{y+dy}, respectively. Subsequently, according to the relationship between the difference and the derivative, and combine the pixel value difference in the x-direction and the pixel value difference in the y-direction to get the whole value difference

$$\begin{aligned} dP_x &= p_x dydt, dP_y = p_y dxdt, \\ dP_{x+dx} &= p_{x+dx} dydt = (p_x + \frac{\partial p_x}{\partial x} dx) dydt, \\ dP_{y+dy} &= p_{y+dy} dxdt = (p_x + \frac{\partial p_y}{\partial y} dy) dxdt, \\ \Delta P_f &= dP_{x+dx} - dP_x + dP_{y+dy} - dP_y \\ &= -(\frac{\partial p_x}{\partial x} + \frac{\partial p_y}{\partial y}) dxdydt, \end{aligned} \tag{3}$$

where dp_x, dp_y, dp_{x+dx}, dp_{y+dy} are respectively the inflow and outflow pixel value intensity along the x-axis and y-axis, which measure the pixel values flowing in and out within per unit area and per unit time. According to the Fourier law in thermodynamics (Borgnakke and Sonntag 2022), which characterizes the relationship between the heat flow and the micro-element temperature gradient in the heat conduction process, dp_x, dp_y can be calculated as follows:

$$dp_x = -\lambda \frac{\partial P}{\partial x}, dp_y = -\lambda \frac{\partial P}{\partial y}. \tag{4}$$

Change in Pixel's Own Value

For each pixel in the infrared image, its own pixel value changes over time and follows $P_s = p_s dxdydt$. p_s represents pixel intensity, which is the pixel value generated within per unit area and time. P_s is the increase of the image micro-element's pixel value due to its internal points' spontaneous pixel value changes. Here we only consider the effect of the difference between the imported and exported pixel values, so pixel value of each point is fixed and will not change, which means $p_s = 0$.

Overall Change in Pixel Value

According to the correspondence between the variables in image field and heat conduction field, we can get the relationship between the pixel value change rate ($\frac{\partial P}{\partial t}$) and the overall pixel value change ΔP during feature map evolution. The change in the micro-element's pixel value can be expressed as:

$$\Delta P = a \frac{\partial P}{\partial t} dxdydt, \tag{5}$$

where a is a constant. From Eq.(2) to Eq.(5), we can get the relationship between the pixel value change rate and gradient during the ISTD process, which is the final expression of pixel movement differential equation (PMDE):

$$\frac{\partial P}{\partial t} = \alpha(\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2}), \tag{6}$$

where $\alpha = (\lambda / a)$. PMDE builds the link between spatial and temporal information of pixel values in an image. In the next section we will use the equation to devise two modules which respectively focus on target body and boundary parts to reflect the flow of pixels.

Methodology

Overall Architecture

The overview of our TCI-Former is displayed in Fig. 2. TCI-Former has a U-Net-like encoder-decoder structure, where the encoder is composed of several Thermal Conduction-Inspired Transformer (TCIT) blocks stacked sequentially while the decoder is built upon three plain deconvolution layers following the common practice. Skip connections are added between the corresponding encoder and decoder layers for cross-layer feature fusion. A fully convolutional segmentation head is connected after the decoder to offer the final predictions. The added circle in the stage blocks denotes the position coding operation for the input tokens. Specifically, each TCIT block contains a Thermal

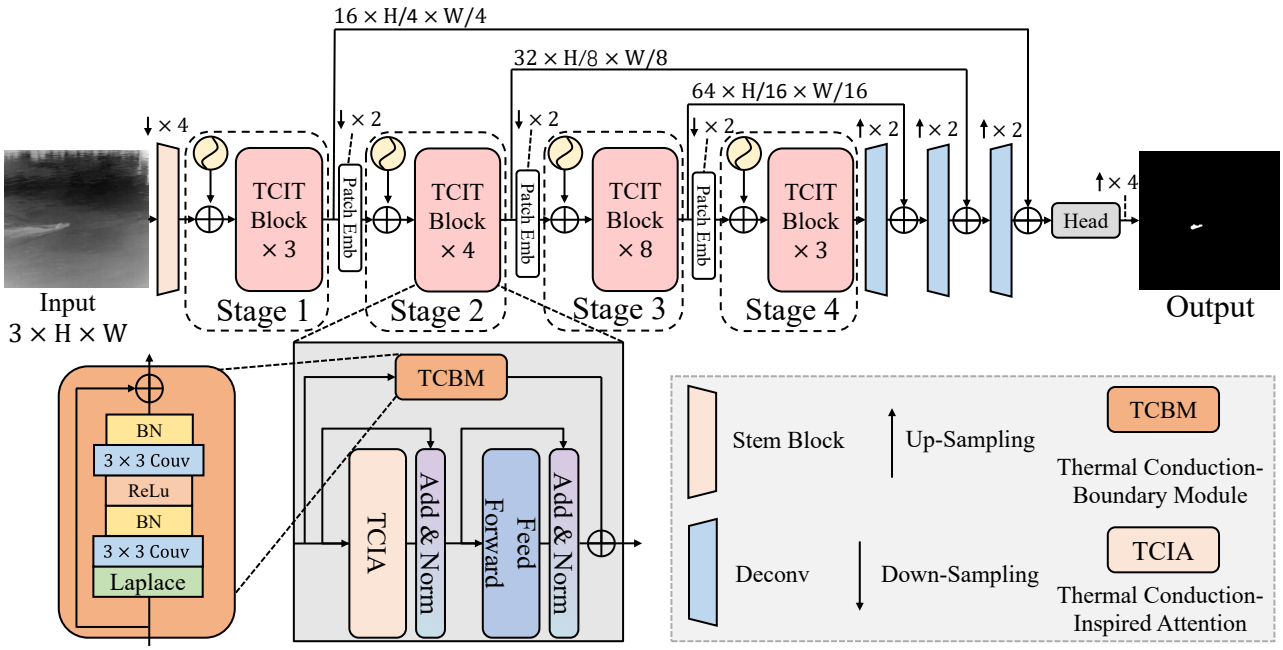

Figure 2: Overall architecture of our TCI-Former with an encoder-decoder structure. The encoder is composed of several TCIT blocks. Each TCIT block contains two key components: Thermal Conduction-Inspired Attention (TCIA) and Thermal Conduction Boundary Module (TCBM), which are both devised based on our derived pixel movement differential equation (PMDE). PMDE is inspired by the thermal conduction differential equation (TCDE) in heat dynamics.

Conduction-Inspired Attention (TCIA) and a Thermal Conduction Boundary Module (TCBM). The TCIT has a parallel structure of global attention and convolution to assemble their merits of modelling local and global information simultaneously. The global attention structure of TCIT block is TCIA, which concentrates on target body information from horizontal and vertical directions in the same way as thermal conduction. The convolutional structure of TCIT is TCBM, which refines target body features with boundary details.

Thermal Conduction-Inspired Attention

The finite difference method is a numerical ODE solver. We apply the method to our PMDE to extract the target main body feature, which can be regarded as an approximation of the whole target feature. Thus, we propose TCIA to explore the rule of target body feature extraction during feature map evolution. Here we use the second-order finite difference equation, which is expressed as:

$$\begin{aligned}\frac{\partial^2 P^t_{i,j}}{\partial x^2} &= \frac{P^t_{i+1,j} - 2P^t_{i,j} + P^t_{i-1,j}}{(\Delta x)^2},\\ \frac{\partial^2 P^t_{i,j}}{\partial y^2} &= \frac{P^t_{i,j+1} - 2P^t_{i,j} + P^t_{i,j-1}}{(\Delta y)^2},\end{aligned} \tag{7}$$

where $P^t_{i,j}$ is the pixel value in position (i, j) in the t-th feature map layer. Applying Eq.(7) to Eq.(6) we have

$$\begin{aligned}P^{t+1}_{i,j} - P^t_{i,j} = \alpha\Big(&\frac{P^t_{i+1,j} - 2P^t_{i,j} + P^t_{i-1,j}}{(\Delta x)^2} + \\ &\frac{P^t_{i,j+1} - 2P^t_{i,j} + P^t_{i,j-1}}{(\Delta y)^2}\Big).\end{aligned} \tag{8}$$

Defining $\Delta x = \Delta y$, we can get the final expression of the target main body part feature extraction rule as follows:

$$P^{t+1}_{i,j} = \gamma(P^t_{i+1,j} + P^t_{i-1,j} + P^t_{i,j+1} + P^t_{i,j-1} - 4P^t_{i,j}) + P^t_{i,j}, \tag{9}$$

where γ denotes $\frac{\alpha}{\Delta x \Delta y}$. Eq.(9) describes that the pixel value at a certain position in a certain feature map layer is determined by its surrounding pixels in x and y axis of its former layer feature map.

Based on Eq.(9), we devise TCIA to extract the main body features of small targets during feature map evolution. Fig. 3 shows the structure of TCIA. The input of TCIA is P^t and the output is $\gamma(P^t_{i+1,j} + P^t_{i-1,j} + P^t_{i,j+1} + P^t_{i,j-1} - 4P^t_{i,j})$, which is obtained through horizontal and vertical conduction attentions Δy and Δx to aggregate surrounding pixel information of the former layer before element-wise addition with P^t. The channel of $P^t \in \mathbb{R}^{C\times H\times W}$ is divided into four groups before shifting each channel group to different directions by $+1$ or -1. In this way, the receptive field of P^t after spatial shift is rhombic, which corresponds to $P^t_{i+1,j} + P^t_{i-1,j} + P^t_{i,j+1} + P^t_{i,j-1}$. We can get $P^t_{i+1,j} + P^t_{i-1,j} + P^t_{i,j+1} + P^t_{i,j-1} - 4P^t_{i,j}$ through residual operation and then linearly project this term into Q, K, V. The horizontal conduction Δy is implemented by taking average of query feature map on the horizontal direction. In the same way, the vertical conduction Δx squeezes query feature map on the vertical direction. The same operations are also conducted upon K and V, so we can get $Q_h, K_h \in \mathbb{R}^{H\times C_{qk}}$, $V_h \in \mathbb{R}^{H\times C_v}$ and $Q_v, K_v \in \mathbb{R}^{W\times C_{qk}}$, $V_v \in \mathbb{R}^{W\times C_v}$. Each of the two conduction attentions re-

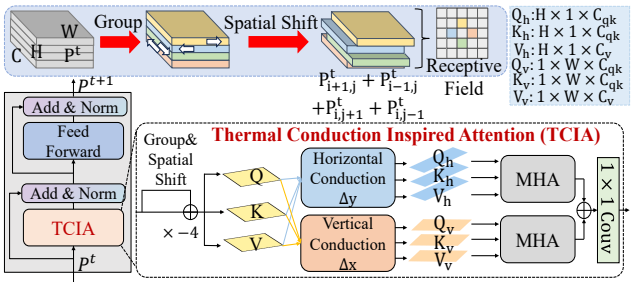

Figure 3: Overall architecture of our proposed Thermal Conduction-Inspired Attention (TCIA), which is devised based on finite difference method and PMDE derived from the TCDE in heat dynamics.

serves the global information to a single axis, so that each position on the feature map propagates information only on two squeezed x-axis and y-axis features. Then the Q, K, V vectors are fed into multi-head attentions and then added together for horizontal and vertical feature aggregation to realize the $\gamma(P^t_{i+1,j}+P^t_{i-1,j}+P^t_{i,j+1}+P^t_{i,j-1}-4P^t_{i,j})$ term. For the last term $P^t_{i,j}$ in Eq.(9), it is added through the residual and layer norm operation in the transformer block. In this way, the TCIA based on Eq.(9) is realized.

Thermal Conduction Boundary Module

TCIA helps extracting target body features, but the features extracted by TCIA branch alone are not fine enough in near boundary regions because the finite difference method used in TCIA is a numerical method, which inevitably brings small errors. A certain degree of dispersion exists due to numerical uncertainty during pixel value movement. To solve this, we need to refine the coarse target body features with fine boundary details to make up for the uncertain errors. We notice that our PMDE itself has already contained boundary information (second-order derivative terms), so to extract target boundary features we design Thermal Conduction Boundary Module (TCBM) based on PMDE. The differential form of Eq.(6) can be described as:

$$P^{t+\Delta t} - P^t = \Delta t \alpha (\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2}.) \tag{10}$$

During ISTD, the extracted feature maps are arranged in a chronological order. The PMDE establishes the relationship between the change of pixel value in temporal domain ($P^{t+\Delta t} - P^t$) and 2-D spatial domain ($\frac{\partial^2 P^t}{\partial x^2}$, $\frac{\partial^2 P^t}{\partial y^2}$) during feature extraction. Defining the time step Δt as 1, we can explore the boundary feature evolution rule between two consecutive feature maps. The specific expression of PMDE can be rewritten as:

$$P^{t+1} - P^t = h\alpha(\frac{\partial^2 P^t}{\partial x^2} + \frac{\partial^2 P^t}{\partial y^2}), \tag{11}$$

where t means the t-th residual calculation. h is the step size between the t-th and t+1-th residual calculation. The TCBM applies spatial information to make up for the lack of boundary refinements during feature extraction in the encoder. The right side of Eq.(11) is the second derivative of P^t in the x- and y-directions, respectively. Thus, with this item, we obtain the spatial information which can be used as the residual supplementary for time information, that is, the information in the forward extraction process. $\frac{\partial^2 P^t}{\partial x^2}$ and $\frac{\partial^2 P^t}{\partial y^2}$ have larger value at boundary areas, therefore TCBM is sensitive to target boundaries and can play a complementary role to the target body features. Our TCBM incorporates a Laplace operator into a residual block, where the Laplace operator is used to realize the $\frac{\partial^2 P^t}{\partial x^2}$ and $\frac{\partial^2 P^t}{\partial y^2}$ terms.

Loss Function

Dice loss (Sudre et al. 2017) measures the difference between a mask prediction and the ground truth. It can also relieve sample imbalance problem and is defined as:

$$L_{dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|}, \tag{12}$$

where X denotes the mask prediction and Y is the ground truth. Our final loss function L_{Final} includes L_{Seg} as the main loss function and Target Boundary loss (L_{TB}) and Interior Body loss (L_{IB}) as two auxiliary loss functions. L_{Final} is calculated as:

$$L_{Final} = L^{hyb}_{Seg} + L^{hyb}_{TB} + L^{hyb}_{IB}. \tag{13}$$

L_{IB} and L_{Seg} share the same Y as the ground truth mask, while the X of L_{IB} is the segmentation head output from the TCIA encoder branch, and the X of L_{Seg} is the final prediction output. The X of L_{TB} is the segmentation head output from the TCBM encoder branch, and the Y of L_{TB} is the boundary mask label.

Experiments

Experimental Settings

Datasets. We choose NUAA-SIRST (Dai et al. 2021a) and IRSTD-1k (Zhang et al. 2022d) as our experimental datasets. NUAA-SIRST contains 427 infrared images of various sizes while IRSTD-1k consists of 1,000 real infrared images of 512 × 512 in size. IRSTD-1k is a more difficult ISTD dataset with richer scenarios. For each dataset, we use 80% of images as training set and 20% as test set.

Evaluation Metrics. We compare our TCI-Former with other SOTA methods in terms of both pixel-level and object-level evaluation metrics. The pixel-level metrics include Intersection over Union (IoU) and Normalized Intersection over Union ($nIoU$), while the object-level metrics include Probability of Detection (P_d) and False-Alarm Rate (F_a).

IoU measures the accuracy of detecting the accuracy of detecting the corresponding object in a given dataset. $nIoU$ is the normalization of IoU, which can make a better balance between structural similarity and pixel accuracy of infrared small targets. IoU and $nIoU$ are defined as:

$$IoU = \frac{A_i}{A_u}, nIoU = \frac{1}{N}\sum_{i=1}^{N}(\frac{TP[i]}{T[i] + P[i] - TP[i]}), \tag{14}$$

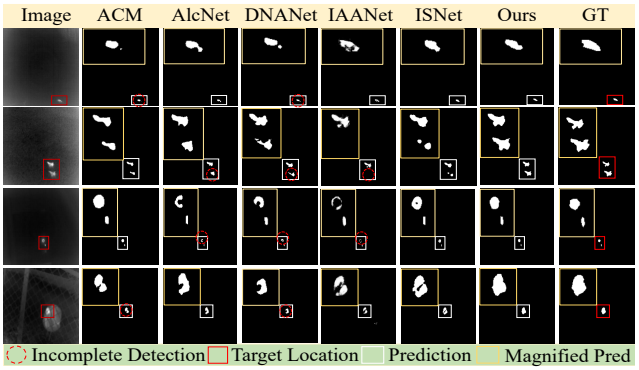

Figure 4: Result visualization of different ISTD methods.

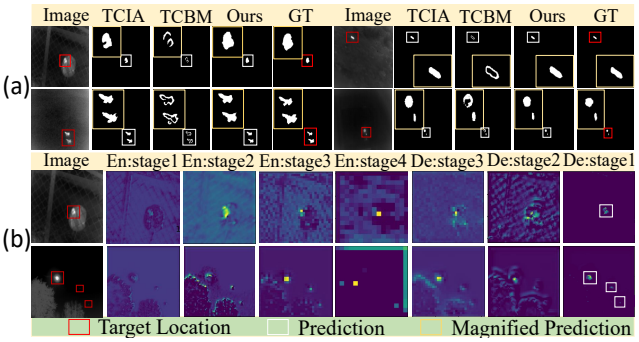

Figure 5: Visualization of (a) TCIA and TCBM branch outputs and (b) intermediate stage feature map evolution.

where A_i and A_u are the areas of intersection region and union region between the prediction and ground truth, respectively. N is the total number of samples, $TP[.]$ is the number of true positive pixels, $T[.]$ and $P[.]$ is the number of ground truth and predicted positive pixels.

P_d calculates the ratio of the number of correctly predicted targets N_{pred} to all targets N_{all}. F_a refers to the ratio of falsely predicted target pixels N_{false} to all the pixels in the infrared image N_{all}. P_d and F_a are calculated as follows:

$$P_d = \frac{N_{pred}}{N_{all}}, F_a = \frac{N_{false}}{N_{all}}. \tag{15}$$

Optimization. The algorithm is implemented in Pytorch, with Adaptive Gradient (AdaGrad) as the optimizer with the initial learning rate set to 0.05 and weight decay coefficient set to 0.0004. A Titan XP GPU is used for training, with batch size set to 4. Training on SIRST and IRSTD-1k takes 800 epochs and 600 epochs respectively.

Comparison with SOTA Methods

Quantitative Comparisons. We select some SOTA ISTD methods for comparison. As shown in Table 1, our TCI-Former performs the best in terms of pixel-level and object-level metrics on both datasets.

For the pixel-level metrics ($IoU, nIoU$), the deep-learning methods generally surpass the traditional methods because deep-learning methods do not rely heavily on prior knowledge and handcraft features as traditional methods do. However, deep-learning methods lay insufficient emphasis on target edges, causing limited IoU and $nIoU$. Our TCI-Former achieves the best performance on both IoU and $nIoU$, meaning that our method achieves the best shape-aware segmentation performance thanks to our TCBM.

For the object-level metrics (P_d, F_a), how to reach a trade-off between P_d and F_a is challenging because the two metrics are mutually exclusive. Traditional methods fail to balance the two metrics but deep-learning methods make it. Our TCI-Former achieves the best object-level metrics results except that our F_a is second only to RKformer (Zhang et al. 2022a) in NUAA-SIRST. However, our F_a significantly outperforms it in IRSTD-1k, which is a more difficult ISTD dataset with richer scenarios. The results demonstrate that our method can learn better representations to find the small targets covered by low contrast and noisy background owing to our TCIA, which mimics thermal conduction to extract target main body features.

Visual Comparisons. Visual results with closed-up views of different methods is shown in Fig. 4. As shown in Fig. 4, most CNN-based methods suffer incomplete detection for lack of extracting global contexts. Hybrid method generally outperforms CNN-based methods with fewer severely incomplete detection cases, but still cannot predict accurate target shapes. Compared with other methods, our method significantly curtails bad cases and achieves better boundary-aware segmentation of small targets. This is because our network can not only extract target body features like thermal conduction, but also refine body features with fine boundary information.

To demonstrate the target body location effect of TCIA and boundary refinement effect of TCBM, we visualize the segmentation head outputs of TCIA branch and TCBM branch in Fig. 5 (a). To present the from-coarse-to-fine feature map evolution process, we visualize intermediate feature maps of all stages in encoder (En: stage1,2,3,4) and decoder (De: stage3,2,1) in Fig. 5 (b). We can find that the small target areas gradually get highlighted like heat conducted from warm to cold areas from the decoder stage 3,2,1 feature maps, which complies with our analogy.

Ablation Study

Impact of Each Module. The ablation study of TCIA and TCBM is shown in Table 2. The baseline uses basic pyramid ViT (Wang et al. 2022b) as encoder. Table 2 demonstrates the positive effects of both designs and combining them together brings the best results, implying that they are complementary to each other. The reason is that ViT block equipped with TCIA can extract main target body features from surrounding areas in orthogonal directions, while TCBM in parallel refines the coarse body features with boundary details to improve detection performance.

Impact of TCIA. To ablate TCIA, we compare our TCIA with multi-head self-attention (MHSA) (Wang et al. 2022b), cross-shaped window self-attention (CSWSA) (Dong et al. 2022) and the multi-head relation attention (MHRA) (Li et al. 2022b). As shown in Table 3, our TCIA outperforms

Method	Type	NUAA-SIRST				IRSTD-1k			
		IoU ↑	nIoU ↑	Pd ↑	Fa ↓	IoU ↑	nIoU ↑	Pd ↑	Fa ↓
PSTNN (Zhang and Peng 2019)	Trad	22.40	22.35	77.95	29.11	24.57	17.93	71.99	35.26
MSLSTIPT (Sun, Yang, and An 2020)	Trad	10.30	9.58	82.13	1131	11.43	5.93	79.03	1524
MDvsFA (Wang, Zhou, and Wang 2019)	CNN	60.30	58.26	89.35	56.35	49.50	47.41	82.11	80.33
ACM (Dai et al. 2021a)	CNN	72.33	71.43	96.33	9.325	60.97	58.02	90.58	21.78
AlcNet (Dai et al. 2021b)	CNN	74.31	73.12	97.34	20.21	62.05	59.58	92.19	31.56
DNANet (Li et al. 2022a)	CNN	75.27	73.68	98.17	13.62	<u>69.01</u>	<u>66.22</u>	91.92	17.57
Dim2Clear (Zhang et al. 2023)	CNN	77.20	75.20	99.10	6.72	66.3	64.2	93.7	20.9
FC3-Net (Zhang et al. 2022c)	CNN	74.22	72.64	99.12	6.569	64.98	63.59	92.93	15.73
IAANet (Wang et al. 2022a)	Hybrid	75.31	74.65	98.22	35.65	59.82	58.24	88.62	24.79
RKformer (Zhang et al. 2022a)	Hybrid	77.24	74.89	99.11	**1.580**	64.12	64.18	93.27	18.65
ISNet (Zhang et al. 2022d)	CNN	<u>80.02</u>	<u>78.12</u>	<u>99.18</u>	4.924	68.77	64.84	<u>95.56</u>	<u>15.39</u>
TCI-Former	Hybrid	**80.79**	**79.85**	**99.23**	<u>4.189</u>	**70.14**	**67.69**	**96.31**	**14.81**

Table 1: Quantitative results of different methods on NUAA-SIRST and IRSTD-1k. The figures in bold and underline mark the highest and the second highest ones in each column.

Method	IoU ↑	nIoU ↑	Pd ↑	Fa ↓
Baseline	62.82	60.59	92.97	26.37
+TCIA	67.26	65.03	94.55	19.83
+TCIA+TCBM	**70.14**	**67.69**	**96.31**	**14.81**

Table 2: Ablation study of each module on IRSTD-1k.

Method	IoU ↑	nIoU ↑	Pd ↑	Fa ↓
MHSA	66.73	64.69	94.05	19.22
CSWSA	68.23	66.05	95.36	17.41
MHRA	68.86	66.87	95.74	16.76
TCIA	**70.14**	**67.69**	**96.31**	**14.81**

Table 3: Ablation study of TCIA on IRSTD-1k.

Method	IoU ↑	nIoU ↑	Pd ↑	Fa ↓
ResBlock	68.93	66.62	95.90	16.35
Roberts+ResBlock	69.51	67.38	96.02	15.70
TCBM	**70.14**	**67.69**	**96.31**	**14.81**

Table 4: Ablation study of TCBM on IRSTD-1k.

Method	Param	FLOPs	Inf
ACM (Dai et al. 2021a)	0.52	2.02	0.01
DNANet (Li et al. 2022a)	4.7	56.34	0.15
IAANet (Wang et al. 2022a)	14.05	18.13	0.29
RKformer (Zhang et al. 2022a)	29.00	24.73	0.08
TCI-Former	3.66	5.87	0.04

Table 5: Comparison of the model parameters (M), FLOPs (G) and inference time (s) of different methods.

others in all metrics, showing better small target location ability. The reason is that in TCIA the spatial shift operation enables the encoder block to be more aware of boundaries, which helps extracting more complete target body features. The superiority of TCIA demonstrates our analogy between ISTD process and thermal conduction process is effective.

Impact of TCBM. In Table 4 we compare TCBM (Laplace+Resblock) with basic Resblock and basic Resblock with Roberts operator to examine the boundary feature extraction effect of different designs. Our TCBM delivers the best result, because (1) edge operators help basic Resblock to extract edges and (2) the edges extracted by Roberts operator is thick and less accurate.

Model Complexity Analysis

We also compare the model complexity of different methods in terms of parameter number (M), FLOPs (G) and inference time (s), as shown in Table 5. Compared with other methods, our method doesn't have many parameters and has acceptable FLOPs and inference time. This is because we squeeze the dimensions of q, k, v before attention operations in our TCIA, which reduces model parameters and improves efficiency. Our model reach a general balance among different model complexity indicators.

Conclusion

Motivated by the analogy of pixel movement during ISTD process and thermal conduction in thermodynamics, we propose TCI-Former for ISTD. We first derive PMDE for the image domain from thermodynamic equation. We then apply finite difference method to PMDE and devise TCIA and embed it into encoder block to extract target main body features by simulating the thermal conduction process. We also propose TCBM based on PMDE to parallelly refine the target body features with fine boundary details. Experiments on NUAA-SIRST and IRSTD-1k prove the superiority of TCI-Former, which explores a new research routine.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62121002, No. 62272430, No. U20B2047) and the Fundamental Research Funds for the Central Universities.

References

Borgnakke, C.; and Sonntag, R. E. 2022. *Fundamentals of thermodynamics*. John Wiley & Sons.

Chen, G.; Wang, W.; and Tan, S. 2022. IRSTFormer: A Hierarchical Vision Transformer for Infrared Small Target Detection. *Remote Sensing*, 14(14): 3258.

Chen, T.; Chu, Q.; Liu, B.; and Yu, N. 2023a. Fluid dynamics-inspired network for infrared small target detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 590–598.

Chen, T.; Chu, Q.; Tan, Z.; Liu, B.; and Yu, N. 2023b. ABM-Net: Coupling Transformer with CNN Based on Adams-Bashforth-Moulton Method for Infrared Small Target Detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 1901–1906. IEEE.

Chen, T.; Chu, Q.; Tan, Z.; Liu, B.; and Yu, N. 2023c. 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)*, 1–5. IEEE.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021a. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 950–959.

Dai, Y.; Wu, Y.; Zhou, F.; and Barnard, K. 2021b. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11): 9813–9824.

Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; and Guo, B. 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12124–12134.

Du, S.; Wang, K.; and Cao, Z. 2023. BPR-Net: Balancing Precision and Recall for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Han, J.; Liu, S.; Qin, G.; Zhao, Q.; Zhang, H.; and Li, N. 2019. A local contrast method combined with adaptive background estimation for infrared small target detection. *IEEE Geoscience and Remote Sensing Letters*, 16(9): 1442–1446.

Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; and Guo, Y. 2022a. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*.

Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; and Qiao, Y. 2022b. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*.

Liu, F.; Gao, C.; Chen, F.; Meng, D.; Zuo, W.; and Gao, X. 2021. Infrared Small-Dim Target Detection with Transformer under Complex Backgrounds. *arXiv preprint arXiv:2109.14379*.

Marvasti, F. S.; Mosavi, M. R.; and Nasiri, M. 2018. Flying small target detection in IR images based on adaptive toggle operator. *IET Computer Vision*, 12(4): 527–534.

Qi, M.; Liu, L.; Zhuang, S.; Liu, Y.; Li, K.; Yang, Y.; and Li, X. 2022. 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.

Sudre, C. H.; Li, W.; Vercauteren, T.; Ourselin, S.; and Jorge Cardoso, M. 2017. 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*, 240–248. Springer.

Sun, Y.; Yang, J.; and An, W. 2020. 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.

Wang, H.; Zhou, L.; and Wang, L. 2019. 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*, 8509–8518.

Wang, K.; Du, S.; Liu, C.; and Cao, Z. 2022a. Interior Attention-Aware Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022b. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 415–424.

Weng, Z.; Li, P.; Zhuang, X.; Yan, X.; Gong, L.; Xie, H.; and Wei, M. 2023. ifUNet++: Iterative Feedback UNet++ for Infrared Small Target Detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

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

Zhang, M.; Bai, H.; Zhang, J.; Zhang, R.; Wang, C.; Guo, J.; and Gao, X. 2022a. RKformer: Runge-Kutta Transformer with Random-Connection Attention for Infrared Small Target Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1730–1738.

Zhang, M.; Wu, Q.; Guo, J.; Li, Y.; and Gao, X. 2022b. Heat transfer-inspired network for image super-resolution reconstruction. *IEEE Transactions on neural networks and learning systems*.

Zhang, M.; Yue, K.; Zhang, J.; Li, Y.; and Gao, X. 2022c. Exploring Feature Compensation and Cross-level Correlation for Infrared Small Target Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1857–1865.

Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; and Guo, J. 2022d. ISNet: Shape Matters for Infrared Small Target Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 877–886.

Zhang, M.; Zhang, R.; Zhang, J.; Guo, J.; Li, Y.; and Gao, X. 2023. Dim2Clear network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Zhang, T.; Cao, S.; Pu, T.; and Peng, Z. 2021. AGPCNet: Attention-Guided Pyramid Context Networks for Infrared Small Target Detection. *arXiv preprint arXiv:2111.03580*.

Zhu, J.; Chen, S.; Li, L.; and Ji, L. 2023. Sanet: Spatial Attention Network with Global Average Contrast Learning for Infrared Small Target Detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.