# 基于视觉 Transformer 和双解码器的红外小目标检测方法

代少升[1]，刘科生[1]，黄 炼[2]，贺自强[1]，毛兴华[1]，任汶皓[1]
（1. 重庆邮电大学 通信与信息工程学院，重庆 400065；2. 重庆理工大学 电气与电子工程学院，重庆 400054）

**摘要**：当前基于卷积神经网络的红外小目标检测方法在编码器阶段受限于感受野，且解码器在多尺度特征融合中缺乏有效的特征交互。本文提出了一种基于编码器-解码器结构的新方法，针对现有红外小目标检测方法中的问题进行改进。该方法使用视觉 Transformer 作为编码器，能够有效地提取红外小目标图像的多尺度特征。视觉 Transformer 是一种新兴的深度学习架构，其通过自注意力机制捕捉图像中像素之间的全局关系，以处理长程依赖性和上下文信息。此外，本文还设计了一个由交互式解码器和辅助解码器组成的双解码器模块，旨在提高解码器对红外小目标的重构能力。该双解码器模块能够充分利用不同特征之间的互补信息，促进深层特征和浅层特征之间的交互，并通过将两个解码器的结果进行叠加，以更好地重构红外小目标。在广泛使用的公共数据集上的实验结果表明，本文提出的方法在 $F_1$ 和 mIoU 两个评价指标上的性能优于其他对比方法。

**关键词**：红外小目标检测；视觉 Transformer；多尺度特征融合；编解码结构
**中图分类号**：TP391.4 **文献标识码**：A **文章编号**：1001-8891(2024)09-1070-11

# Infrared Small Target Detection Method with Vision Transformer and Dual Decoder

DAI Shaosheng[1]，LIU Kesheng[1]，HUANG Lian[2]，HE Ziqiang[1]，MAO Xinghua[1]，REN Wenhao[1]
(1. *Department of Communication and Information Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*; 2. *Department of Electrical and Electronic Engineering*, *Chongqing University of Technology, Chongqing* 400054, *China*)

**Abstract**: The existing infrared small-target detection method based on convolutional neural networks (CNN) exhibits the problem of a limited receptive field in the encoder stage, and the decoder lacks an effective feature interaction when fusing multiscale features. To address the aforementioned issues, in this study, a new method is proposed based on an encoder–decoder structure. Specifically, a vision transformer is used as an encoder to extract multiscale features from small infrared target images. The vision transformer is an emerging deep-learning architecture that uses a self-attention mechanism to capture the global relationship between all pixels in the input image, thereby effectively processing long-range dependencies and contextual information in the image. Furthermore, a dual-decoder module, comprising an interactive decoder and auxiliary decoder, is proposed to improve the ability of the decoder to reconstruct small infrared targets. The dual-decoder module can make full use of the complementary information between different features, promote interaction between deep and shallow features, and better reconstruct small infrared targets by combining the results of the two decoders. Experimental results on widely used public datasets show that the proposed method outperforms other methods in terms of two evaluation indicators: $F_1$ and mIoU.

**Key words**: infrared small target detection, vision transformer, multiscale feature fusion, encoder and decoder structure

## 0 引言

红外小目标（如空中目标、海上目标以及地面目标等）检测是一项非常重要的任务，它涉及许多视觉

**收稿日期**：2023-05-24；**修订日期**：2023-07-11.
**作者简介**：代少升（1974-），男，博士，教授，主要从事红外成像系统及 SOPC 嵌入式系统的设计与开发。
**通信作者**：黄炼（1992-），男，博士，讲师，主要从事红外小目标、小样本学习等研究。E-mail: hlcysxxy@163.com。

任务，如海上监视[1]、红外跟踪[2]、红外预警[3]和红外成像制导[4]等。相比基于可见光图像的目标识别，红外小目标具有如下特点：首先，红外小目标在图像中所占的像素点非常有限，通常仅表现为一个点，缺乏明确的尺度和形状特征。根据国际光学工程学会（Society of Photo-Optical Instrumentation Engineers，SPIE）对红外小目标的定义：小目标成像的尺寸小于 81 像素即小于 256×256 的 0.15%[5]。其次，红外图像中的背景具有较高的复杂性，包括建筑物、海洋、空中（云层）和陆地等各种环境。在这些背景条件下，红外小目标往往面临对比度较低的挑战，容易与背景融合在一起，难以被准确检测。如图 1 所示，红外小目标图像中目标与背景的相似度高，红外小目标数量少且分布不均匀，容易被忽略或误检。因此，上述的这些问题使得精确检测红外小目标十分困难。

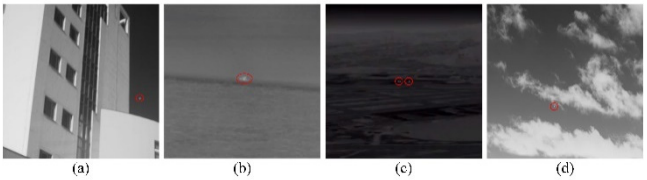



图 1 部分红外小目标图像样本。(a) 建筑物背景下的红外小目标，(b) 海洋背景下的红外小目标，(c) 陆地背景下的红外小目标，(d) 空中(云层)背景下的红外小目标

Fig.1 Some samples of infrared small target images. (a) depicts infrared small target in building background, (b) depicts infrared small target in sea background, (c) depicts infrared small target in land background, and (d) depicts infrared small target in sky background

传统的红外小目标检测方法通常基于各种假设设计手工制作的特征来检测红外小目标。这些方法包括基于形态学滤波的方法、基于局部对比的方法以及基于主成分分析的方法等。其中，基于形态学滤波的方法利用数学形态学中的滤波算子来检测红外小目标，如 TopHat[6]，MaxMedian[7]等。基于局部对比的方法则使用局部对比度作为特征来检测红外小目标。Wei 等[8]提出一种多尺度基于块的对比度量方式，将单个邻域扩展为 8 个不同尺度的邻域。Aghaziyarati 等[9]提出了一种基于平均绝对灰度差的局部对比度度量，以降低漏检率。而基于主成分分析的方法则使用主成分分析来提取红外小目标的特征，以进行检测和分类。Gao 等[10]利用低秩矩阵恢复的思想解决红外小目标检测问题。而为了应对复杂的背景，Wang 等[11]结合变分正则化和主成分追踪（Total Variation Regularization and Principal Component Pursuit, TV-PCP）来描述背景特征。然而，这些传统的方法通常需要手动选择和设计特征，在处理复杂场景和复杂的红外小目标时具有一定的局限性。

相比之下，基于深度学习方法，如基于生成对抗网络（Generative Adversarial Network, GAN）和基于编码器-解码器结构的方法，他们通过数据驱动的方式去学习红外小目标的特征。GAN-Based 方法在生成器和判决器之间采用对抗性学习的方式来平衡误检和漏检，如图 2(a)所示。Wang 等[12]提出了一种基于对抗学习的方法将图像分割视为生成对抗网络的优化问题，其主要思想是使用对抗性学习来平衡误检和漏检。但由于在模型的训练过程中获得最佳的平衡模型比较困难，因此基于生成对抗网络的方法会存在模型崩溃问题。而基于编码器-解码器的方法使用编码器提取红外小目标图像的特征并使用解码器对红外小目标进行重构，如图 2(b)所示。由于其简单的结构和训练过程，基于编码器-解码器结构的方法受到了越来越多的关注。Li 等[13]设计了一个带有级联通道和空间注意模块（Channel and Spatial Attention Module，CSAM）的三向密集嵌套交互模块（Dense Nested Interactive Module，DNIM），以实现渐进式特征交互和自适应特征增强。Wu 等[14]通过将红外小物体检测建模为语义分割问题，提出了一种简单明了的红外小物体检测框架，称为 U-Net 中的 U-Net（UIU-Net）。Dai 等[15-16]引入了局部对比度度量的概念，提出了深度参数较少的非线性特征细化层。

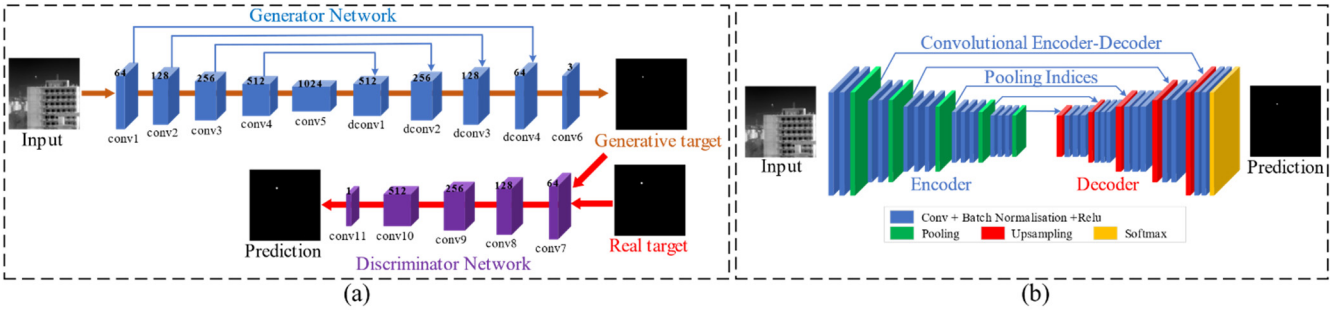



图 2 两种典型方法的网络结构。(a)基于 GAN 的方法，(b)基于编码器-解码器结构的方法

Fig.2 The network architecture of two typical methods. (a) the GAN-Based method and (b) the Encoder-Decoder-Based method

与传统方法相比，深度学习方法可以自动提取特征，克服手动选择和设计特征的限制，能够在处理复杂场景和复杂的红外小目标时取得更好的性能。然而，基于卷积神经网络（Convolutional Neural Network，CNN）的编码器通常使用固定大小的卷积核，其感受野有限，无法完全捕捉目标与背景之间的全局相关性，也就是像素之间的相似性。这对于红外小目标检测是非常不利的，因为在特征提取阶段对全局相关性进行建模可以提高多尺度特征的辨别能力。此外，多尺度特征融合方法有助于增强红外小目标的特征表示。但常用的解码器通常采用固定的解码路径来融合多尺度特征。例如，Li 等[17]直接融合多级特征，通过级联操作实现特征融合。而 Zhang 等[18]则是将多级特征集成到多个分辨率中，并在特定分辨率下用这些特征预测最终结果。然而，上述融合方法没有考虑多尺度特征之间的重要程度。在红外小目标检测任务中，Huang 等[19]也只是简单地将深层特征和浅层特征沿通道维度拼接起来，没有针对不同特征进行通道或空间信息的交互。Dai 等[15]考虑深层特征和浅层特征之间的通道依赖性，设计通道注意力来调整深层特征的通道信息。他们设计了一个非对称上下文模块（Asymmetric Contextual Modulation, ACM）来替换 U-Net[20]的普通跳跃连接。但上述的这些方法均采用固定的解码路径去融合多尺度特征。固定的解码路径限制了不同尺度特征之间的交互，意味着解码器不能充分利用不同特征之间的互补信息，从而导致次优的红外小目标重构性能。

视觉 Transformer[21]采用自注意机制能够用于解决 CNN-Based 编码器存在的感受野受限问题，该机制能够捕获红外小目标图像中不同位置之间的全局关系，即能够建模远程依赖关系。Chen 等[22]提出了 TransUNet，他们认为 Transformer 可以作为医学图像分割任务的强大编码器，通过结合 U-Net[20]去恢复局部空间信息来增强更精细的细节。在红外小目标检测领域，Liu 等[23]首先提出了探索视觉 Transformer 检测红外小目标的工作，并在红外小目标检测中取得成功。他们首先使用 CNN 来提取局部特征。然后，他们采用 Vision Transformer（ViT）从局部特征中获取有关红外小目标定位的高级信息。然而，他们的单层 ViT 结构只适用于最后一个 CNN 层提取的特征。因此，他们的方法不能完全捕捉形状描述的低级信息，容易混淆真实目标和背景。最新方法的 MTU-Net（Multi-level TransUNet）[24]则结合了多层 ViT 模块和 CNN。首先使用 CNN-Based 编码器去提取多尺度特征。然后，通过 MVTM（Multilevel ViT Module）细化特征以捕获多尺度特征的长距离依赖关系。以上的这些方法表明视觉 Transformer 在红外小目标检测任务中具有很好的应用前景。同时，为了提高解码器重构红外小目标的能力，设计能够充分利用多尺度特征之间互补信息的新型解码模块至关重要。

基于上述动机，本文提出了一种基于编码-解码结构的红外小目标检测方法。首先使用 PVT（pyramid vision transformer）[25]作为编码器去提取多尺度特征。其次，设计了一个由交互式解码器和辅助解码器组成的双解码器模块去充分利用不同尺度特征之间的互补信息。交互式解码器通过级联自上而下融合、加权交叉融合和自下而上融合 3 个过程去融合多尺度特征。在自上而下的融合中，通过将浅层特征引入深层特征中以提高红外小目标的空间表征。在加权交叉融合中，通过将可学习的权重分配给不同的特征以突出它们重要程度。而在自下而上的融合中，通过将深层特征中的语义信息引入浅层特征以增强红外小目标的语义表征。此外，辅助解码器直接融合多尺度特征以获得更加丰富的上下文信息去进一步补充更多细节和语义特征信息。总之，双解码器模块可以有效融合多尺度特征，增强解码器重构红外小目标的能力。

## 1 网络结构

### 1.1 网络的整体结构

本文提出的红外小目标检测方法由编码器和解码器两部分组成，如图 3(a)所示。首先使用基于视觉 Transformer 的编码器来提取输入图像的多尺度特征，因为视觉 Transformer 采用纯自注意力（self-attention）机制能够建模像素之间的全局关系，从而有效地解决了传统 CNN-Based 编码器感受野受限的问题，提升了多尺度特征的表示能力。解码器则用于融合多尺度特征并重构红外小目标，以达到更精确的检测效果。

### 1.2 编码器模块

本文选择基于视觉 Transformer 的编码器即 pyramid vision transformer（PVT）作为特征提取的骨干网络。PVT 在视觉 Transformer 中嵌入了金字塔结构并采用纯自注意机制去提取多尺度特征。由图 3(a)所示，编码器模块分为 4 个串联的阶段，每个阶段通过视觉 Transformer 去提取不同尺度的特征。具体地，给定一张单通道的红外小目标图像 $I=\mathbb{R}^{H\times W}$，其中 $H$ 和 $W$ 分别表示输入图像的高度和宽度。PVT 从输入的红外小目标图像中提取多尺度特征 $F$，多尺度特征 $F$ 表示为 $F=\left\{F^i\in\mathbb{R}^{C_i\times H_i\times W_i}\,\middle|\,i=1,2,3,4\right\}$，其中 $F^i$ 表示编码器的第 $i$ 个阶段提取的特征，$C_i$ 表示第 $i$ 个阶段所提取特征的通道数，$H_i$ 和 $W_i$ 分别表示第 $i$ 个阶段所提取特征的高度和宽度。

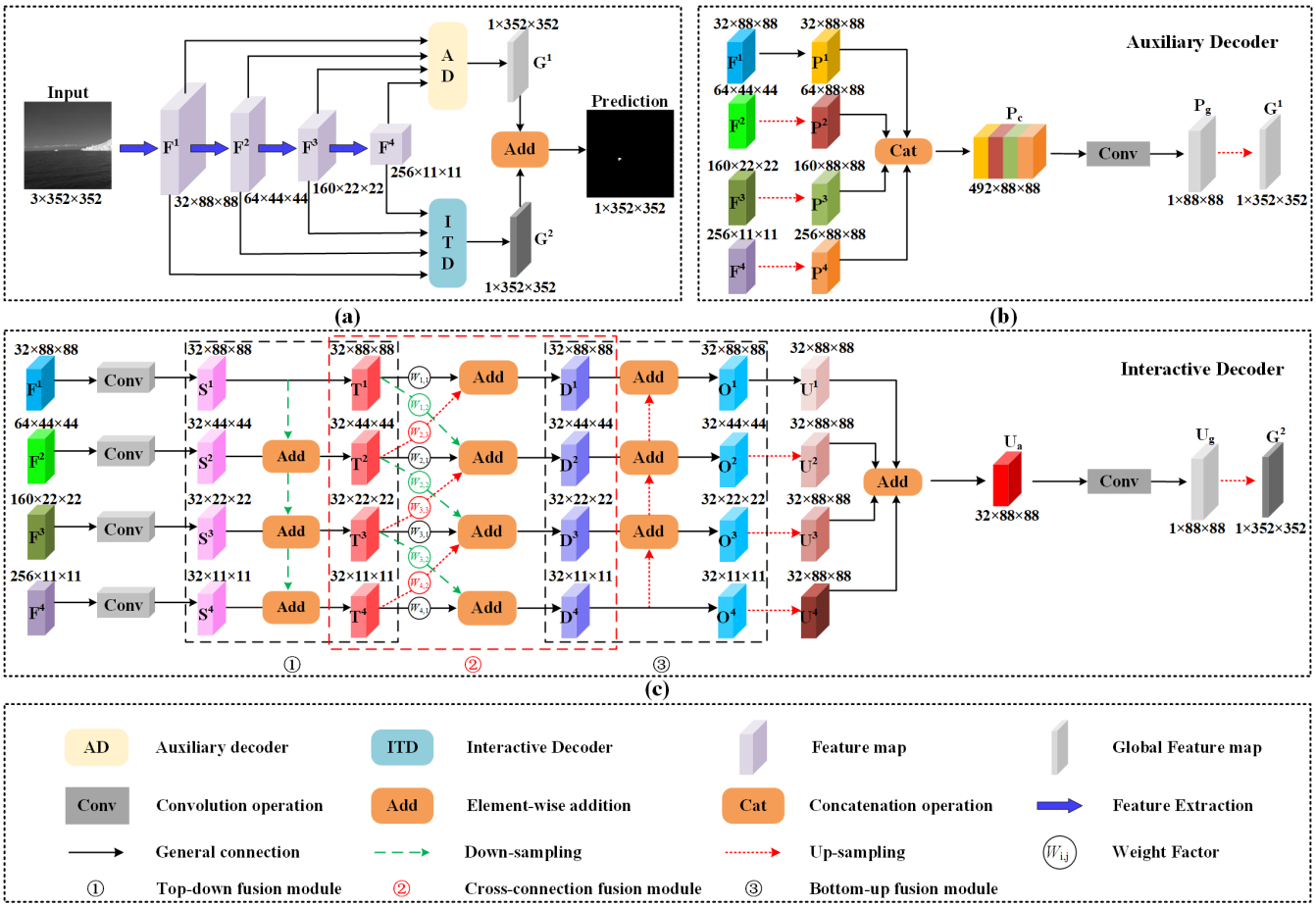



图 3 本文的方法。图(a)本文方法的总体结构，图(b)辅助解码器的实现过程。图(c)交互式解码器的实现过程

Fig.3 The pipeline of our proposed method. The overall architecture in (a), the auxiliary decoder (AD) implementation process in (b), and the interactive decoder (ITD) implementation process in (c)

### 1.3 双解码器模块

双解码器模块是本文提出的一个关键模块，由两个解码器模块组成。交互式解码器由 3 个级联的融合过程组成，旨在有效地融合多尺度特征。这 3 个融合过程分别是自上而下融合、加权交叉融合以及自下而上融合。具体来说，自上而下融合从编码器的浅层特征开始，通过下采样和融合操作逐层增强特征的表达；加权交叉融合则采用一种自适应的方法，通过学习的方式融合多尺度特征；自下而上融合则从编码器的深层特征开始，通过上采样和融合操作逐层重构红外小目标。辅助解码器直接对多尺度特征在通道维度进行拼接，这样可以更充分地利用不同尺度的特征信息，补充更多的细节和语义信息，以获得更好的特征融合效果。通过这样的设计，本文的方法能够更加充分地利用不同尺度的特征，提高了红外小目标检测的性能。

#### 1.3.1 辅助解码器

如图 3(b)所示，辅助解码器模块中，本文采用上采样操作将编码器提取的不同尺度特征统一为相同的分辨率，即 88×88 像素。这一操作使得不同尺度的特征能够在空间上进行对齐，方便进行融合和后续的处理。经过上采样操作后，得到了同一分辨率下新的多尺度特征 $P$。新的多尺度特征 $P$ 可以表示为 $P=\left\{P^i\in\mathbb{R}^{C_i\times 88\times 88}\,\middle|\,i=1,2,3,4\right\}$。然后，在通道维度上将它们拼接在一起，形成了一个组合特征 $P_c$。接下来，使用 1×1 卷积对组合特征进行降维，得到一个全局特征 $P_g\in\mathbb{R}^{1\times 88\times 88}$。最后，将全局特征 $P_g$ 上采样以匹配输入图像的分辨率，并得到预测结果 $G^1$。具体的细节如公式(1)所示。

$$G^1=\mathrm{Up}(\mathrm{Conv}(\mathrm{Cat}(\mathrm{Up}(F^1),\mathrm{Up}(F^2),\mathrm{Up}(F^3),\mathrm{Up}(F^4))))\in\mathbb{R}^{1\times H\times W} \tag{1}$$

式中：Up(·)表示上采样操作；Cat(·)表示在通道维度上进行拼接；Conv(·)表示 1×1 卷积操作。

#### 1.3.2 交互式解码器

交互式解码器的结构如图 3(c)所示。针对编码器提取的多尺度特征 $F$，为了方便进行后续的特征融合，采用了 1×1 卷积操作对每个特征进行维度变换，使

得它们具有相同的维度。通过卷积操作后得到了新的多尺度特征 $S$。新的多尺度特征可以表示为 $S=\left\{S^i\in\mathbb{R}^{32\times H_i\times W_i}\,\middle|\,i=1,2,3,4\right\}$。交互式解码器是由 3 个不同的融合过程级联组合而成的关键模块。下面将对这个模块进行详细介绍。

首先，自上而下融合是指从浅层特征向深层特征进行逐级融合的过程。浅层特征通常包含了图像的细节信息，而深层特征具有更抽象和语义化的信息。自上而下融合过程能够逐渐将浅层特征的细节信息融合到深层特征中，使得特征具备更全面的表征能力。具体做法如下：首先，通过下采样操作，将浅层特征的分辨率调整为与深层特征相同。然后，将下采样后的浅层特征与深层特征进行逐元素相加，实现特征的叠加。融合的方式如公式(2)所示：

$$T^i=\mathrm{Down}(S^{i-1})+S^i \tag{2}$$

式中：Down(·)表示下采样操作；$S^i$ 表示多尺度特征中的第 $i$ 个特征。需要注意的是，对于最上层的特征不进行下采样操作。即 $S^1$ 和 $T^1$ 表示同一个特征。

其次，加权交叉融合是一种动态权重融合的方式。不同尺度的特征对于红外小目标的检测有不同的贡献度，因此需要根据具体情况来调整它们的融合程度。加权交叉融合通过学习动态权重，对不同尺度特征进行加权融合，使得每个特征都能够发挥其最大的作用，从而提高整体的检测性能。具体来说，加权交叉融合过程涉及以下几个具体步骤：首先，对于相邻的特征，进行上采样或下采样操作，使它们具有相同的分辨率。这是为了确保不同尺度的特征能够对齐，方便后续的融合操作。接下来，为每个特征分配一个可学习的权重因子 $w$，用于控制不同特征之间的融合强度。这些权重因子在模型训练过程中会自适应地更新，以最大限度地利用不同特征之间的互补信息，并突出每个特征的重要性。最后，通过加权求和的方式将特征进行融合，得到融合后的特征表示。通过以上步骤，加权交叉融合过程能够根据每个特征的重要性和互补性，自适应地融合不同尺度的特征。公式(3)描述了加权交叉融合的过程。

$$D=\left\{D^i=\sum_{j=1}^{N}w_{i,j}T^i\,\middle|\,i=1,2,3,4;j=1,2,3,4\right\} \tag{3}$$

式中：$N$ 的取值与 $i$ 有关，当 $i$=1,4 时，$N$ 的取值为 2；而当 $i$=2,3 时，$N$ 的值为 3。$w_{i,j}$ 赋给每个特征的权重值，$i$ 表示第 $i$ 个特征，$j$ 表示第 $j$ 个权重值。

最后，自下而上融合指从深层特征向浅层特征进行逐级融合的过程。通过将深层特征向上传递并与浅层特征进行融合，可以将语义信息引入到浅层特征中，提高特征的表征能力。具体做法如下：首先，通过上采样操作，将深层特征的分辨率调整到与浅层特征相同，以确保它们在空间上对齐。然后，将上采样后的深层特征和对应的浅层特征进行逐元素相加，实现特征的叠加。融合的方式如公式(4)所示：

$$O^i=\mathrm{Up}(D^{i+1})+D^i \tag{4}$$

式中：Up(·)表示上采样操作；$D^i$ 表示多尺度特征中的第 $i$ 个特征。需要注意的是，对于最下层的特征不进行上采样操作。即 $O^4$ 和 $D^4$ 表示同一个特征。

通过此模块后得到了融合后的多尺度特征 $O=\left\{O^i\in\mathbb{R}^{32\times H_i\times W_i}\,\middle|\,i=1,2,3,4\right\}$，其具有相同的维度，即具有相同的通道数。为了获得最终的预测结果，首先，通过上采样后得到了具有一致分辨率的多尺度特征 $U=\left\{U^i\in\mathbb{R}^{32\times 88\times 88}\,\middle|\,i=1,2,3,4\right\}$。然后，对多尺度特征 $U$ 采用逐元素相加的方式获得了叠加之后的特征 $U_\mathrm{a}$。与辅助解码器类似，使用 1×1 卷积对特征 $U_\mathrm{a}$ 进行降维，得到了一个全局特征 $U_\mathrm{g}$。最后，将全局特征 $U_\mathrm{g}$ 上采样以匹配输入图像的分辨率，并得到预测结果 $G^2$。具体的细节如公式(5)所示：

$$G^2=\mathrm{Up}(\mathrm{Conv}(\mathrm{Add}(\mathrm{Up}(O^1),\mathrm{Up}(O^2),\mathrm{Up}(O^3),\mathrm{Up}(O^4))))\in\mathbb{R}^{1\times H\times W} \tag{5}$$

式中：Up(·)表示上采样操作；Add(·)表示在逐像素相加；Conv(·)表示 1×1 卷积操作。

#### 1.3.3 双解码器的输出

根据图 3(a)所示，双解码器模块的最终输出结果由两个部分组成。首先是辅助解码器得到的预测结果 $G^1$，其次是交互式解码器通过将不同尺度的特征进行交互融合后得到的预测结果 $G^2$。为了充分利用辅助解码器和交互式解码器各自的优势，我们采用了逐元素相加的方式将这两个预测结果进行叠加，得到了最终的预测结果 $G$。这种叠加方式能够综合利用两个解码器的预测结果，从而更好的重构红外小目标。

### 1.4 损失函数

二分类任务中通常使用二进制交叉熵损失函数（The Binary Cross-Entropy loss, BCE）作为模型的损失函数。如公式(7)所示：

$$L_{\mathrm{BCE}}=-y\log\hat{y}-(1-y)\log(1-\hat{y}) \tag{7}$$

式中：$y$ 表示真实的标签值；$\hat{y}$ 表示预测值。通常情况下，二进制交叉熵损失函数也被广泛应用于红外小目标检测任务中，但是在红外小目标检测任务中，会出现前景像素和背景像素极度不平衡的情况。因为在

红外小目标图像中，前景像素（即小目标）的数量远远少于背景像素（即非小目标）。如果直接采用平等对待前景和背景像素的二进制交叉熵损失函数，会导致模型过度关注背景像素，而无法有效地检测前景像素。

为了缓解不平衡的问题，本文尝试使用 Focal loss[26]损失函数。Focal loss 最初是为解决目标检测中的类别不平衡问题而设计的，通过降低易分类样本的权重，让模型更加关注难分类样本，从而提高目标检测的性能。然而，红外小目标检测与传统目标检测存在较大的差异，因此，Focal loss 是否适用于红外小目标检测任务需要实验验证。Focal loss 的具体表述如公式(8)所示：

$$L_{\text{Focal}} = -\alpha\left(1-\hat{y}\right)^{\gamma} y\log\hat{y} - \left(1-\alpha\right)\hat{y}^{\gamma}\left(1-y\right)\log\left(1-\hat{y}\right) \tag{8}$$

式中：$\alpha$表示平衡因子；$\gamma$表示调制因子；$y$ 表示真实的标签值；$\hat{y}$ 表示预测值。

实验结果表明，与 BCE 相比，Focal loss 能够略微提升检测性能。但是，为了更好地缓解前景像素和背景像素不平衡问题，本文考虑对 Focal loss 进行改进以使其更好地适用于红外小目标检测任务。具体的做法是，在 Focal loss 中添加了一个权重因子$\theta$。$\theta$增加了难分类样本的损失值，以保证模型在训练过程中能够更加关注这些难分类样本。实验表明，改进的损失函数可以获得更好的检测性能。其公式如(9)所示：

$$\begin{aligned} L_{\text{IFocal}} = &-\alpha\left(1-\hat{y}\right)^{\gamma} y\log\hat{y} - \\ &\left[\left(1-\alpha\right)\hat{y}^{\gamma}+\theta\right]\left(1-y\right)\log\left(1-\hat{y}\right) \end{aligned} \tag{9}$$

## 2 实验结果与分析

### 2.1 数据集介绍

1）ISTS-DATA[12]：ISTS-DATA 是一个专门用于卷积神经网络训练的红外小目标数据集。它也是第一个针对红外小目标检测而设计的数据集。数据集中的训练集由 10000 张图像组成，这些图像包含了各种自然场景和合成场景下的红外小目标，背景环境复杂，能够充分考察算法的泛化能力和鲁棒性。此外，该数据集还包含 100 张测试图像，用于测试算法的准确性和稳定性。

2）NUAA-SIRST[15]：由 427 张红外小目标图像组成，其中包含 480 个目标实例。大多数图像只包含一个目标，但也有少数图像包含多个目标。在这个数据集中，很多目标都非常暗淡，且隐藏在杂乱无章的复杂背景中，这为小目标检测任务增加了难度。本文选择 NUAA-SIRST 数据集作为测试集来验证方法的泛化能力。

3）IRSTD-1K[27]：一个包含 1000 张红外小目标图像的数据集，这些图像由红外相机拍摄，涵盖了不同种类的小目标，例如无人机、生物、船舶以及车辆等。此外，数据集中的场景也非常多样化，包括大海、河流、林木、山区、城市和云等多种背景。同时，数据集中存在着噪音和杂波等因素，对小目标的检测带来了挑战。同样地，选择 IRSTD-1K 数据集作为测试集进行测试，以进一步验证方法的泛化能力。

### 2.2 训练环境和参数设置

本文中的模型训练采用 PyTorch 框架，实验所用的计算机 CPU 为 i5-12400，主频 2.50 GHz，GPU 为 Nvidia GTX 1080Ti。本文使用改进的 Focal loss 进行训练并使用 AdamW 优化器进行优化。初始学习率为 1e−4，batch size 为 4，训练的轮数（epoch）设置为 50。本文在模型训练阶段使用 ISTS-DATA 数据集，并在测试阶段使用了上述提到的 3 个数据集。为了确保输入网络的数据具有一致的尺寸，对数据集进行了预处理。采用了 PyTorch 框架中的库函数，如调整图像尺寸（Resize）、将图像归一化（Normalize）以及将图像转化为张量（ToTensor）等，进行图像尺寸的调整和预处理。经过预处理后，所有图像被调整为统一的尺寸，即 352×352 像素的分辨率。这样的统一尺寸有助于确保网络能够处理相同尺寸的输入，并提供一致的特征表示，从而更好地进行红外小目标检测。

### 2.3 评价指标

本文使用常用的评价指标来评估各种算法的检测性能。即 Precision, Recall, $F_1$ 和 mIoU。它们的定义如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{11}$$

$$F_1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{12}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}} \tag{13}$$

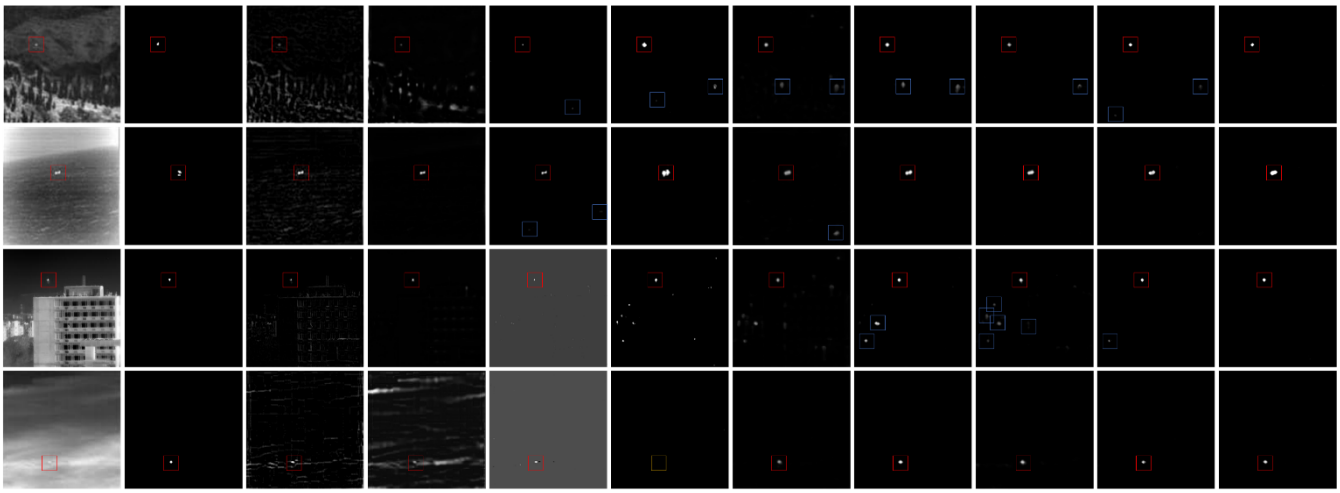
式中：TP 表示被模型正确预测为目标类像素的数量；FP 表示被模型预测为目标类的背景像素数量；FN 表示被模型预测为背景类的目标像素数量。$F_1$ 综合考虑了 Precision 和 Recall，是一个被广泛使用的评价指标。本文选择 $F_1$ 作为主要的性能评价指标。同时，为了更全面地评估算法检测结果，还选择了每个类别 IoU 的平均值，即 mIoU，作为另一个重要的评价指标。

### 2.4 对比实验和结果分析

为了证明本文提出的方法在检测精度和检测效率等方面的综合性能，本章节选取了多种基于传统方法和基于深度学习的方法进行对比，包括 Top-Hat（Top-HatTransform）[6]、LEF（Local Energy Factor）[28]、IPI（Infrared patch-image model）[29]、MDvsFA-cGAN（Miss Detection vs. False Alarm-cGAN）[12]、ALCNet（Attentional Local Contrast Network）[16]、LSPM（Local Similarity Pyramid Modules）[19]、UIU-Net（U-Net in U-Net）[14]、DNANet（Dense Nested Attention Network）[13]等。为了进行公平比较，本文在相同的数据集上对基于深度学习的方法在相同的条件下进行训练并在 3 个不同的测试数据集上进行测试。

#### 2.4.1 ISTS-DATA 数据集上的比较

首先，本文在 ISTS-DATA 数据集上对上述方法进行了全面的比较，评估它们在 4 个评价指标上的性能表现。同时，为了更全面地评估方法的表现，本文还进行了定性的实验比较，并将结果可视化展示在图 4 中。从表 1 中可以看出，本文的方法在 ISTS-DATA 数据集获得了最好的 $F_1$（0.7032）和 mIoU（0.5384），与传统的方法相比具有显著的优势。此外，相对于基于深度学习的方法，本文的方法也表现优异。$F_1$ 综合考虑了 Precision 和 Recall 去评价算法的性能。因此，单一的 Precision 和 Recall 并不能准确地评价方法的好坏。比如，IPI 得到了最高的 Precision（0.7537）但是却牺牲了 Recall，最终使得 $F_1$ 的值很低。而本文的方法能够在 Precision 和 Recall 之间达到很好的平衡。此外，从表 1 中可以得到，传统红外小目标检测方法的性能明显低于基于深度学习的方法。



(a) Infrared image (b) Ground truth (c) Top-Hat (d) LEF (e) IPI (f) MDvsFA-cGAN (g) ALCNet (h) LSPM (i) UIU-Net (j) DNANet (k) Ours

图 4 不同方法的可视化结果。红色、黄色和蓝色的框分别代表正确检测到的目标、未检测到的目标和错误检测到的目标

Fig.4 Visualization results of different methods. Boxes in red, yellow, and blue represent correctly detected targets, miss detected targets, and falsely detected targets, respectively

表 1 不同算法在 ISTS-DATA 数据集上的实验结果

Table 1 Experimental results of different algorithms on the ISTS-DATA dataset

| Methods | Precision | Recall | $F_1$ | mIoU |
|---|---|---|---|---|
| Top-Hat | 0.5106 | 0.2202 | 0.3077 | 0.1536 |
| LEF | 0.5071 | 0.2745 | 0.3562 | 0.1675 |
| IPI | 0.7537 | 0.3452 | 0.4735 | 0.2036 |
| MDvsFA-cGAN | 0.6335 | 0.6562 | 0.6447 | 0.4686 |
| ALCNet | 0.6658 | 0.6641 | 0.6649 | 0.4995 |
| LSPM | 0.6559 | 0.6762 | 0.6659 | 0.5078 |
| DNANet | 0.6233 | 0.6876 | 0.6539 | 0.4857 |
| UIU-Net | 0.5969 | 0.6972 | 0.6432 | 0.4740 |
| Ours | 0.6858 | 0.7216 | 0.7032 | 0.5384 |

在 ISTS-DATA 数据集上定性的比较结果如图 4 所示。我们选择了 4 张具有代表性的红外小目标图像，涵盖了不同背景环境和目标类型。这些图像包括陆地、海洋、空中以及建筑物等背景下的红外小目标。第一列表示原始图像，第二列表示标签值，其余各列分别表示各种方法的预测结果。从图 4 所示的结果中可以看出，虽然包括 TopHat 和 LEF 在内的传统方法可以准确地检测出不同背景中的红外小目标，但背景的形状仍然清晰地展现出来了。这表明传统方法无法有效地将目标与背景分开。同样地，基于深度学习的方法，如 MDvsFA-cGAN、ALCNet、UIU-Net 和 DNANet，也会遇到同样的问题。而且 MDvsFA-cGAN 方法还存在漏检问题，它无法检测到被云层遮挡的红

外小目标。本文的方法在不同的场景中均取得了令人满意的结果。基于视觉 Transformer 的编码器在提取红外小目标图像的多尺度特征时可以对图像中所有像素之间的关系进行长距离建模，从而增强红外小目标的特征表示。此外，本文的方法采用了双解码器模块，可以充分利用不同尺度特征之间的互补信息，并考虑不同特征之间的交互，从而在重构红外小目标方面表现出更好的性能。

**2.4.2** NUAA-SIRST 和 IRSTD-1k 上的比较

为了验证本文提出方法的泛化能力，选择了 3 种目前基于深度学习方法中比较先进的方法（LSPM、DNANet 和 UIU-Net）进行比较。本文的方法和以上 3 种方法的比较均在 ISTS-DATA 数据集上训练得到的最优模型上进行测试。实验结果如表 2 所示。

表 2 不同算法在 NUAA-SIRST 和 IRSTD-1k 数据集上的实验结果

Table 2 Experimental results of different algorithms on NUAA-SIRST and IRSTD-1k datasets

| Methods | NUAA-SIRST | | IRSTD-1k | |
|---|---|---|---|---|
| | $F_1$ | mIoU | $F_1$ | mIoU |
| LSPM | 0.7313 | 0.5764 | 0.5516 | 0.3809 |
| DNANet | 0.7065 | 0.5462 | 0.5207 | 0.3502 |
| UIU-Net | 0.6645 | 0.4976 | 0.4998 | 0.3331 |
| Ours | 0.7609 | 0.6202 | 0.6238 | 0.4517 |

通过表 2 的结果可以发现，本文提出的方法在 NUAA-SIRST 和 IRSTD-1k 这两个数据集上都表现出了最佳的检测性能。在 NUAA-SIRST 数据集上，与其他 3 种方法相比，本文的方法获得了最高的 $F_1$（0.7609）和 mIoU（0.6202）。这些结果说明本文所提出的方法在不同的数据集上都能有较好的检测性能。

**2.4.3** 测试结果分析

根据测试结果，我们可以观察到各个方法在 3 个数据集上的表现存在较大的差异。这种差异可能由以下原因所引起：

首先，考虑图像中背景与目标的强度。通过从 3 个测试数据集中随机选择的红外小目标图像如图 5 所示，可以看到不同数据集之间的背景特点差异。在 NUAA-SIRST 数据集中，目标与背景之间的差异较为明显，背景对目标的干扰相对较小。然而，在 IRSTD-1K 和 ISTS-DAT 数据集中，图像的背景更加复杂，而且小目标的亮度较低，这使得目标与背景之间的区分较为困难，导致模型在这个两个数据集上的表现较差。

其次，考虑数据集的背景类型和目标数量。根据表 3 中的数据，我们可以看到 ISTS-DATA 和 NUAA-SIRST 测试数据集中的红外小目标图像主要背景类型相似，并且主要包含单目标图像。然而，在 IRSTD-1K 测试数据集中，存在更多的多目标图像。相比其他两个数据集，模型可能无法完全准确地检测出图像中的所有红外小目标，从而影响了检测效果。此外，在此数据集中出现了大量以林木为背景的图像，如图 5 所示，背景环境明显比其他两个数据集更复杂。这一差异也导致了模型在此数据集上的检测效果较差。

(a) ISTS-DATA (b) NUAA-SIRST (c) IRSTD-1K

图 5 测试数据集中红外小目标图像对比

Fig.5 Comparison of infrared small target images in the test datasets

表 3 数据集的分析与比较

Table 3 Analysis and comparison of Datasets

| Dataset | Quantity/pieces | Background type | Small target types | Single object count/Multiple object count |
|---|---|---|---|---|
| ISTS-DATA | 100 | Land, Clouds, Buildings, Ocean, et al. | Land, Aerial, and Marine Targets | 75/25 |
| NUAA-SIRST | 427 | Clouds, Buildings, et al. | Primarily Aerial Targets | 365/62 |
| IRSTD-1k | 1000 | Clouds, Trees, et al. | Mainly aerial and land targets. | 655/345 |

综上所述，图像中背景与目标的强度、背景类型以及目标数量等是导致模型在不同数据集上表现差异的主要因素。这些因素的差异性可能导致模型在某些数据集上无法准确区分目标和背景，从而影响了检测性能。

### 2.5 模型复杂度比较

当输入图像的分辨率为 352×352 像素时，求得模型的参数量（Params）和浮点运算量（floating-point operations per second, FLOPs）。这两个指标用于评价模型的复杂度。根据表 4 的数据，与几种典型的深度学习方法相比，本文的方法具有最低的 FLOPs。这意味着本文的方法具有更快的推理速度。与使用卷积运算进行特征提取的网络不同，本文的方法采用基于视觉 Transformer 的网络来提取多尺度特征。在特征提取阶段，没有使用卷积运算，而是采用了自注意力机制来捕捉图像中的特征关系，从而显著减少了参数的数量。通过使用基于视觉 Transformer 的网络，能够在保持良好检测性能的同时降低了模型的复杂度。

表 4 深度学习方法参数量和浮点运算量比较

Table 4 Comparison of FLOPs and Params of deep learning methods

| Methods | FLOPs | Params |
|---|---|---|
| MDvsFA-cGAN | 988.44G | 15.23M |
| ALCNet | 14.52G | 8.56M |
| LSPM | 233.31G | 31.14M |
| DNANet | 53.99G | 4.70M |
| UIU-Net | 206.08G | 50.54M |
| Ours | 8.84G | 7.18M |

### 2.6 消融实验

在本节中，我们首先进行了损失函数的消融实验，以评估其对模型性能的贡献。接下来，我们验证了本文方法中每个模块的有效性。

#### 2.6.1 损失函数比较

通过对比实验来验证改进 Focal loss 的有效性。具体而言，在训练模型时分别采用了 BCE、Focal loss 和基于 Focal loss 改进的损失函数。实验结果如表 5 所示，相比于 BCE 损失，Focal loss 能够略微提高检测性能，将 $F_1$ 由 0.6675 提升至 0.6758。但是，本文中使用的改进 Focal loss 损失函数在评价指标 $F_1$ 上表现最好，与其他两个损失函数相比，该损失函数能更好地适用于红外小目标检测任务。在 Focal loss 中添加的权重因子能够使模型更关注难分类的样本，有助于提升检测性能。

表 5 不同损失函数下的 $F_1$

Table 5 The value of $F_1$ under different loss functions

| Loss Function | $F_1$ |
|---|---|
| BCE loss | 0.6675 |
| Focal loss | 0.6758 |
| Improved focal loss (Ours) | 0.7032 |

#### 2.6.2 编码器比较

为了验证 PVT 作为编码器去提取多尺度特征的性能，本文进行了对比实验。在以往的红外小目标检测研究中，通常使用 VGG16 或 ResNet50 等传统卷积神经网络作为编码器来提取多尺度特征。因此，通过将 VGG16 和 ResNet50 分别替换为 PVT，并在相同的实验设置下进行对比，我们能够评估 PVT 作为编码器在红外小目标检测中的性能表现。这样的对比实验能够提供有关不同编码器对于红外小目标检测的影响的信息，进一步揭示 PVT 在该任务中的优势和潜力。

实验结果如表 6 所示。通过表 6 的结果可以发现，使用 PVT 作为编码器去提取多尺度特征相比于使用 VGG16 和 ResNet50，在检测性能上具有明显的优势。特别是，本文所提出的方法在 ISTS-DATA 数据集上取得了较好的检测性能，$F_1$ 为 0.7032，比 VGG16 和 ResNet50 均提升了 0.03 左右。

表 6 不同编码器的性能比较

Table 6 Comparison of different encoders

| Encoder | $F_1$ |
|---|---|
| PVT (ours) | 0.7032 |
| VGG16 | 0.6714 |
| ResNet50 | 0.6774 |

#### 2.6.3 解码器比较

本文的对比实验基准包括了 PVT、交互式解码器（ITD）和辅助解码器（AD）。为了评估交互式解码器的贡献，我们进行了两个实验。首先，在第一个实验中，我们采用了基于特征金字塔结构的解码器（Feature Pyramid Network, FPN）来替换 ITD，并通过实验测试。实验结果如表 7 所示。

表 7 不同编码器的性能比较

Table 7 Comparison of different decoders

| Methods | $F_1$ |
|---|---|
| PVT + ITD + AD (ours) | 0.7032 |
| PVT + FPN + AD | 0.6706 |
| PVT + AD | 0.6565 |
| PVT + ITD | 0.6831 |
| PVT + FPN | 0.6652 |

根据表 7 的结果可知,ITD 的检测性能优于 FPN。相比于 FPN，ITD 能够更好地利用不同特征之间的互补信息，并促进浅层特征和深层特征之间的充分交互。通过促进特征之间的信息交互,ITD 能够弥补 FPN 在特征传递和融合方面的不足。这样，ITD 可以更好地捕捉目标的细节和上下文信息，从而提高了检测的准确性和鲁棒性。另一个实验是去掉交互式解码器，仅使用辅助解码器进行实验。从表 7 的结果可以明显看出，模型的检测性能显著下降。这进一步证明了交互式解码器的有效性和重要性。仅使用辅助解码器，模型无法充分利用特征之间的交互信息，导致特征的表征能力受限。此外，由表 7 中的实验结果可知，即使未使用辅助解码器，仅使用 ITD 或 FPN 也能够实现较好的检测性能。然而，仅使用 AD 的效果并不理想，因为 AD 直接融合了来自编码器的多尺度特征，而没有考虑特征之间的交互作用。相比之下，ITD 和 FPN 都考虑了特征之间的信息交互，从而提升了特征的表征能力。然而，无论是 ITD 还是 FPN，在有无 AD 的情况下，检测性能存在差别。这说明 AD 对提高红外小目标的检测性能方面有一定的作用。AD 直接在通道维度上拼接多尺度特征，捕捉了这些特征中的细节信息和语义信息，与 ITD 或 FPN 的结果进行叠加，从而提升了特征的表征能力。双解模块的方式有助于解码器更准确地重构红外小目标，提高了检测性能。

## 3 结束语

本文提出了一种新颖的方法，利用基于视觉 Transformer 网络作为编码器和双解码器模块来实现红外小目标的检测。首先，本文使用基于视觉 Transformer 的网络作为编码器，用于提取多尺度特征。与传统的卷积操作不同，这种基于自注意力机制的编码器能够更好地捕捉图像中的全局依赖关系，从而提高特征的表征能力。其次，设计的双解码器模块可以更好地利用多尺度特征之间的互补信息，促进不同尺度特征之间的交互，以更好地重构红外小目标。本文进行了大量实验来证明此方法的有效性。在公共数据集上的表现优于目前最先进的方法，并且在不同的数据集上具有较好的泛化性能。

## 参考文献：

[1] 刘洋, 战荫伟. 基于深度学习的小目标检测算法综述[J]. **计算机工程与应用**, 2021(2): 37-48.
LIU Yang, ZHAN Yinwei. Survey of small object detection algorithms based on deep learning[J]. *Computer Engineering and Applications*, 2021(2): 37-48.

[2] HUANG Y, LI X, LU R, et al. Infrared maritime target tracking via correlation filter with adaptive context-awareness and spatial regularization[J]. *Infrared Physics & Technology*, 2021, **118**: 103907.

[3] 韩金辉, 魏艳涛, 彭真明, 等. 红外弱小目标检测方法综述[J]. **红外与激光工程**, 2022, **51**(4): 20210393. doi: 10.3788/IRLA20210393.
HAN Jinhui, WEI Yantao, PENG Zhenming, et al. Infrared dim and small target detection: a review[J]. *Infrared and Laser Engineering*, 2022, **51**(4): 20210393. doi: 10.3788/IRLA20210393.

[4] SUN Y, YANG J, AN W. Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, **59**(5): 3737-3752.

[5] ZHANG W, CONG M, WANG L. Algorithms for optical weak small targets detection and tracking: review[C]//*IEEE Proceedings of the* 2003 *International Conference on Neural Networks and Signal Processing*, 2003(1): 643-647.

[6] ZENG M, LI J, PENG Z. The design of top-hat morphological filter and application to infrared target detection[J]. *Infrared Physics & Technology*, 2006: **48**(1): 67-76.

[7] Deshpande S D, ER M H, Venkateswarlu R, et al. Max-mean and max-median filters for detection of small targets[C]//*Proceedings of the SPIE*, 1999, **3809**: 74-83.

[8] WEI Y, YOU X, LI H. Multiscale patch-based contrast measure for small infrared target detection[J]. *Pattern Recognition*, 2016, **58**: 216-226.

[9] Aghaziyarati S, Moradi S, Talebi H. Small infrared target detection using absolute average difference weighted by cumulative directional derivatives[J]. *Infrared Physics & Technology*, 2019, **101**: 78-87.

[10] GAO C, WANG L, XIAO Y, et al. Infrared small-dim target detection based on Markov random field guided noise modeling[J]. *Pattern Recognition*, 2018, **76**: 463-475.

[11] WANG X, PENG Z, KONG D, et al. Infrared dim target detection based on total variation regularization and principal component pursuit[J]. *Image and Vision Computing*, 2017, **63**: 1-9.

[12] WANG H, ZHOU L, WANG L. Miss detection vs. false alarm: adversarial learning for small object segmentation in Infrared images[C]// *IEEE/CVF International Conference on Computer Vision* (ICCV), 2019: 8508-8517.

[13] LI B, XIAO C, WANG L, et al. Dense nested attention network for infrared small target detection[J]. *IEEE Transactions on Image Processing*, 2023, **32**: 1745-1758.

[14] WU X, HONG D, CHANUSSOT J. UIU-Net: U-Net in U-Net for infrared small object detection[J]. *IEEE Transactions on Image Processing*, 2023, **32**: 364-376.

[15] DAI Y, WU Y, ZHOU F, et al. Asymmetric contextual modulation for infrared small target detection[C]// *IEEE Winter Conference on Applications of Computer Vision* (WACV), 2021: 949-958.

[16] DAI Y, WU Y, ZHOU F, et al. Attentional local contrast networks for infrared small target detection[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, **59**(11): 9813-9824.

[17] LI G, YU Y. Deep contrast learning for salient object detection[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 478-487.

[18] ZHANG P, WANG D, LU H, et al. Amulet: Aggregating multi-level convolutional features for salient object detection[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2017: 202-211.

[19] HUANG L, DAI S, HUANG T, et al. Infrared small target segmentation with multiscale feature representation[J]. *Infrared Physics & Technology*, 2021, **116**: 103755.

[20] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//*Medical Image Computing and Computer-Assisted Intervention–MICCAI*, 2015: 234-241.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv:1706.03762, 2017.

[22] CHEN J, LU Y, YU Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[23] LIU F, GAO C, CHEN F, et al. Infrared small-dim target detection with transformer under complex backgrounds[J]. arXiv preprint arXiv:2109.14379, 2021.

[24] WU T, LI B, LUO Y, et al. MTU-Net: Multi-level TransUNet for space-based infrared tiny ship detection[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, **61**: 3235002.

[25] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 568-578.

[26] LIN T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2980-2988.

[27] ZHANG M, ZHANG R, YANG Y, et al. ISNET: Shape matters for infrared small target detection[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 877-886.

[28] XIA C, LI X, ZHAO L, et al. Infrared small target detection based on multiscale local contrast measure using local energy factor[J]. *IEEE Geoscience and Remote Sensing Letters*, 2020, **17**(1): 157-161.

[29] GAO C, MENG D, YANG Y, et al. Infrared patch-image model for small target detection in a single image[J]. *IEEE Transactions on Image Processing*, 2013, **22**(12): 4996-5009.