

基于 U 型多尺度 Transformer 网络的 红外小目标检测算法

段沛沛¹, 张严², 雒明世¹, 闫效莺¹

(1.西安石油大学 计算机学院, 陕西 西安 710065; 2.陕西科技大学 电子信息与人工智能学院, 陕西 西安 710021)

摘 要:针对红外小目标特征难以提取、易被噪声干扰及复杂背景淹没等问题,提出了一种基于 U 型多尺度 Transformer 网络的检测算法。该算法在 U 型多尺度网络架构下,借助卷积操作提取、强化小目标局部显著性特征,同时又基于 Transformer 机制对图像全局特征进行建模,以获取红外图像背景信息;通过对所生成目标置信图与特征图的自注意力运算,完成了对图像浅层和深层特征的融合,实现了对像素级红外小目标的分割及检测。实验证明,在红外序列图像弱小飞机目标检测跟踪数据集中,即使针对背景复杂且含噪的图像进行检测,所提算法性能仍然优于对比算法,呈现了良好的鲁棒性及稳定、准确的检测效果。在算法阈值选用使 FM 平均值最大的情况下,其检测率为 0.997 2,虚警率为 2.82×10^{-7} ,精确率为 0.912 7,而召回率则为 0.921。

关 键 词:红外小目标检测;图像分割;深度学习;自注意力机制

中图分类号: TN219

文献标志码: A

文章编号: 1000-2758(2025)01-0154-09

红外小目标检测是红外搜索与跟踪系统的关键技术之一,已被广泛用于远程预警、海上监视、战场情报监控等多个领域^[1]。然而,在这些应用场景下,往往面临因目标成像机制、成像距离、噪声干扰等因素造成的检测困难。为此,如何准确、稳定地检测红外小目标依然是一类亟待解决的难题。

几十年来,人们基于单帧图像或序列图像提出了诸多小目标检测方法。相比而言,前者具有复杂度小、实时性高且易于工程实现等优点。为此,本文仅关注单帧检测方法。传统的单帧检测方法多依赖预设特征开展检测,比如:基于形态学滤波的方法^[2-3]、基于局部对比度的方法^[4-5]和基于鲁棒主成分分析的方法^[6]等,但其实预设特征很可能与实际情况不符,此类算法的检测效果自然会受影响。

随着深度学习理论的发展,诸多学者尝试将其引入红外小目标检测,比如:文献[7]设计了一种用于检测目标的 5 层多层感知(MLP)网络;文献[8]

微调了几类目标检测网络(Faster-RCNN 和 Yolo-V3);文献[9]设计了一个端到端的神经网络,其目标检测模块可直接给出目标的类型和位置。此外,还有些学者尝试将基于卷积神经网络的图像语义分割方法^[10-11]用于红外小目标检测,但是受小目标成像机制及特点的影响,这些算法的处理效果仍待提升。实际上,红外图像中背景占据了绝大部分区域。因此,如果能获得图像背景的结构特征,将有助于抑制背景干扰,提高小目标分割或检测的精度。可是,多数基于 CNN 的检测方法常采用不同大小的卷积核捕获相应感受野中的图像特征,要获取更大范围内的特征,需借助池化操作完成,而这会导致图像分辨率下降,甚至造成小目标信息丢失。

近年来,在计算机视觉应用中,有些算法使用 Transformer 模型^[12]作为主干网络,更好地捕获了图像的整体特征^[13-14]。为此,本文借鉴 Transformer 模型及 U-Net 图像分割网络设计思想,提出了一种 U 型多尺度红外小目标分割网络,即 ISTS-UT(infrared small target segmentation with U-shaped Transformer)网络,对图像进行粗粒度分析和高层语义分析,进而提取图像细节特征和全局特征,并通过对小目标分

收稿日期:2023-09-23

基金项目:陕西省自然科学基金基础研究计划(S2023-JC-YB-2680)资助

作者简介:段沛沛(1980—),讲师

通信作者:段沛沛(1980—) e-mail:hello_pei@126.com

割实现检测目的。

1 本文方法

文中所提 ISTS-UT 网络可分为两大模块:下采样编码模块和上采样解码模块,每个模块均采用 3 层递进结构,该网络的架构如图 1 所示。

基于 ISTS-UT 网络的小目标检测过程分为下采

样编码和上采样解码 2 个处理阶段。在下采样编码阶段,借助卷积操作获取小目标的局部显著性信息,并通过对生成特征图的 Transformer 编码,求取目标与背景的全局相关性信息,实现像素级差异性信息向区域级相关性信息的转变。在上采样解码阶段,则会凭借多轮次的 Transformer 解码及反卷积操作,提升分析精度,实现像素级的小目标分割,达成检测目的。ISTS-UT 检测网络结构如图 1 所示。

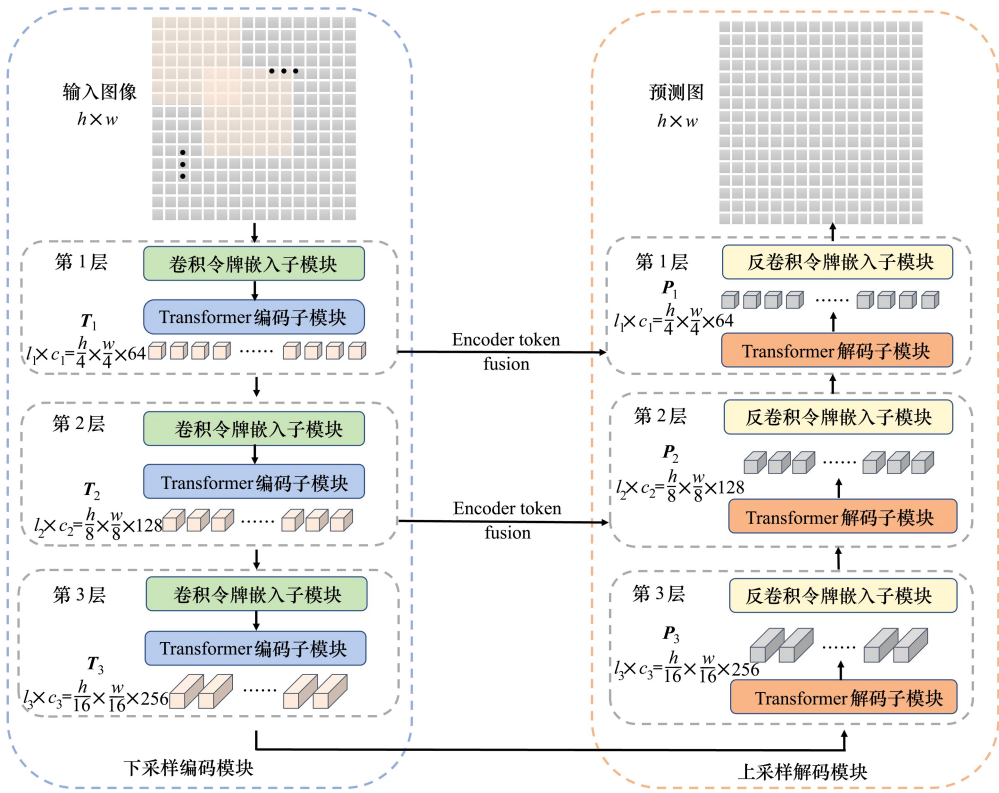


图 1 ISTS-UT 检测网络结构图

1.1 下采样编码模块

图 1 中下采样编码模块采用了 3 层递进结构,以获取反映图像结构信息的特征向量组。每层操作可分为两步,分别由卷积令牌嵌入子模块和 Transformer 编码子模块完成。

1) 卷积令牌嵌入子模块

卷积令牌嵌入子模块被用于对图像局部区域上下文信息进行建模,以获取图像不同层次的局部特征,处理过程如图 2 所示。该子模块会将原始图像 $M_{i-1}(i=1)$ 或前层所得特征向量序列 $T_{i-1}(i=2,3)$ 进行卷积,对两者的处理大致相同,只是在处理原始图像时无需做维度变换。以第 $i(i=2,3)$ 层处理为

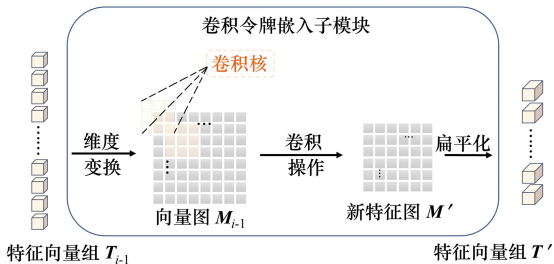


图 2 卷积令牌嵌入子模块工作示意图

例,倘若当前输入为特征向量组 T_{i-1} ,经过维度变换后生成的二维向量图 $M_{i-1} \in \mathbf{R}^{h_{i-1} \times w_{i-1} \times c_{i-1}}$ (其中 h_{i-1} ,

w_{i-1} 和 c_{i-1} 分别为特征图 $\mathbf{M}_{i-1}$ 的高、宽和通道数); 针对 $\mathbf{M}_{i-1}$, 将使用卷积核 K (大小为 $k \times k$, 边缘填充为 p , 步长为 s , 卷积核数目为 n) 与之卷积, 即可得到新特征图 $\mathbf{M}' \in \mathbf{R}^{h_i \times w_i \times c_i}$, 其中

$$h_i = \lfloor \frac{h_{i-1} + 2p - k}{s} + 1 \rfloor \quad (1)$$

$$w_i = \lfloor \frac{w_{i-1} + 2p - k}{s} + 1 \rfloor \quad (2)$$

卷积处理之后, 再将 $\mathbf{M}'$ 转换为一维特征向量组 $\mathbf{T}' \in \mathbf{R}^{h_i w_i \times c_i}$

$$\mathbf{T}' = \text{Flatten}(\text{Conv2d}(\text{Reshape}(\mathbf{T}_{i-1}), k)) \quad (3)$$

在该子模块中, 通过对红外图像的降采样, 获取其浅层细节信息及深层的、更具丰富语义的局部结构信息, 实现对红外图像中小目标的增强。

2) Transformer 编码模块

仍以第 i ($i = 1, 2, 3$) 层的处理为例, 之前所得一维特征向量组 $\mathbf{T}' \in \mathbf{R}^{n_i \times c_i}$ (其中 $n_i = h_i \times w_i$ 为向量数目, c_i 是通道数) 将被送入 Transformer 编码模块进行相关性求解^[12], 具体过程如图 3 所示。

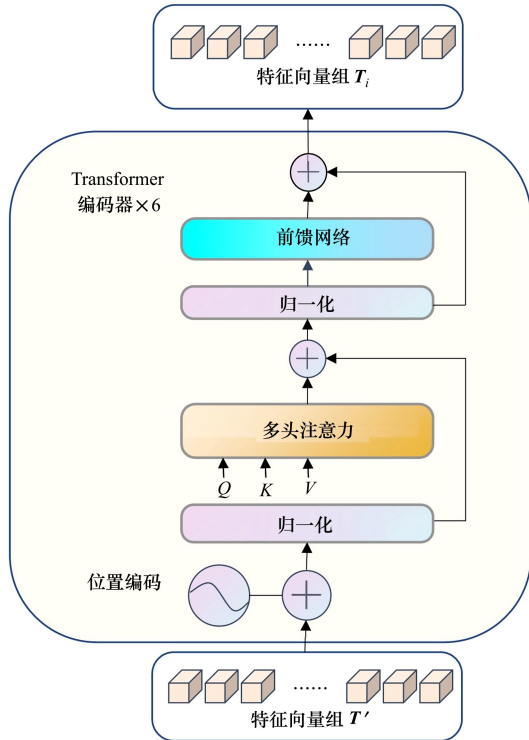


图3 Transformer 编码模块结构示意图

为保留一维特征向量组 $\mathbf{T}'$ 中各向量在原始图像或特征图中的空间信息, 可将其与一维可学习位置向量 $\mathbf{E}_{\text{pos}} \in \mathbf{R}^{n_i \times c_i}$ 叠加, 得到 $\mathbf{Z}_0$ 。

$$\mathbf{Z}_0 = \mathbf{T}' + \mathbf{E}_{\text{pos}} \quad (4)$$

$\mathbf{Z}_0$ 将被送入 L 个 (本设计中 $L = 6$) Transformer 编码器串联组成的 Transformer 编码模块中进行处理, 各 Transformer 编码器^[12] 的操作如 (5) ~ (6) 式所示。

$$\mathbf{Z}'_l = \text{MSA}(\text{Norm}(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1}, \quad l = 1, 2, \dots, L \quad (5)$$

$$\mathbf{Z}_l = \text{MLP}(\text{Norm}(\mathbf{Z}'_l)) + \mathbf{Z}'_l, \quad l = 1, 2, \dots, L \quad (6)$$

所得 $\mathbf{Z}_l \in \mathbf{R}^{h_i w_i \times c_i}$ 为 L 次编码后所得一维向量组, 即下采样编码模块在当前层级 (第 i 层) 的输出 $\mathbf{T}_i$ 。

文中算法设定卷积核尺寸 $k = [(7 \times 7), (3 \times 3), (3 \times 3)]$, 通道数 $c = [64, 128, 256]$, 步长 $s = [4, 2, 2]$, 边缘填充为 $p = [3, 1, 1]$ 。经过 3 层编码处理, 将逐层得到 $\mathbf{T}_1 \in \mathbf{R}^{h_1 w_1 \times c_1}$, $\mathbf{T}_2 \in \mathbf{R}^{h_2 w_2 \times c_2}$ 和 $\mathbf{T}_3 \in \mathbf{R}^{h_3 w_3 \times c_3}$ 3 个向量组, 其中 $h_1 w_1 = \frac{h}{4} \times \frac{w}{4}$, $h_2 w_2 = \frac{h}{8} \times \frac{w}{8}$, $h_3 w_3 = \frac{h}{16} \times \frac{w}{16}$, $[c_1, c_2, c_3] = [64, 128, 256]$, $\mathbf{T}_3$ 为下采样编码模块的最终输出。

1.2 上采样解码模块

ISTS-UT 网络中的上采样解码模块被用于对不同尺度的目标特征向量组 $\mathbf{T}_i$ ($i = 1, 2, 3$) 进行相关性分析, 以获取检测图像从粗粒度到像素级精度的预测图 $P'_{M_{i-1}}$ ($i = 3, 2, 1$)。该模块也为 3 层递进结构, 每层通过两步操作获取一维预测向量组及分辨率更高的预测图。这两步操作分别由 Transformer 解码子模块和反卷积令牌嵌入子模块完成。

1) Transformer 解码子模块

该模块由 L 个 (文中 $L = 6$) 标准的 Transformer 解码器^[12] 串联叠加而成, 可在生成目标预测图的同时, 采用跳跃连接把编码阶段的目标特征信息引入 Transformer 解码过程, 以实现目标特征融合, 解决深层网络目标特征丢失的问题。该子模块的具体构造及工作过程如图 4 所示。

Transformer 解码子模块位于上采样解码模块各层入口, 其输入除第 3 层为前一模块的特征向量组 $\mathbf{T}_3$ 外, 其余 2 层 ($i = 2, 1$) 的输入均为前一层生成的目标预测向量组 $\mathbf{P}'_i$ 和来自下采样编码模块的相同尺度下的目标特征向量组 $\mathbf{T}_i$ 。该模块所涉计算如 (7) ~ (8) 式所示。

$$\mathbf{P}_i = \text{MLP}(\text{MSA}(\text{Norm}(\mathbf{P}'_i + \mathbf{E}_{\text{pos}}), \text{Norm}(\mathbf{T}_i + \mathbf{E}_{\text{pos}}))), \quad i = 2, 1 \quad (7)$$

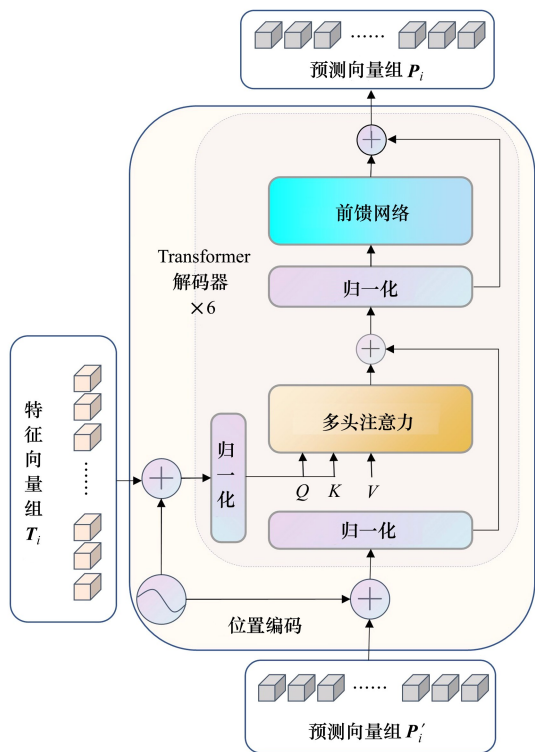


图 4 Transformer 解码子模块结构示意图

在第 3 层中,(7)式将简化为

$$P_3 = \text{MLP}(\text{MSA}(\text{Norm}(T_3 + E_{\text{pos}}))) \quad (8)$$

2) 反卷积令牌嵌入子模块

使用深层网络的确有助于获取图像更深层次的语义信息,但对于获取图像更多的细节信息却并无助益。为此,算法采用反卷积令牌嵌入子模块对由 P_i 转换所得二维预测图 P_{Mi} 进行处理,实现对目标的细粒度密集预测,具体处理过程如图 5 所示。

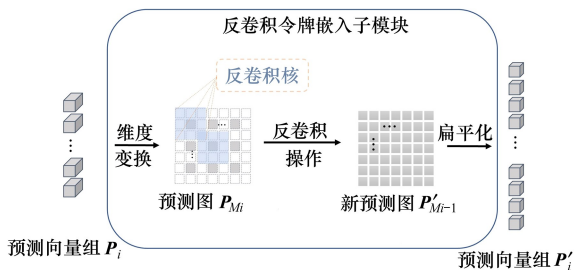


图 5 反卷积令牌嵌入子模块工作示意图

上采样解码模块也为 3 层结构,每层都涉及反卷积操作,其中卷积核尺寸 $k = [(3 \times 3), (3 \times 3), (7 \times 7)]$,通道数 $c = [256, 128, 64]$,步长 $s = [2, 2, 4]$,边缘填充 $p = [1, 1, 3]$ 。经过上采样解码模块各

层处理后,将得到分辨率各不相同的新预测图 P'_{M2} , P'_{M1} , P'_{M0} ,其分辨率分别为 $h_2 w_2 = \frac{h}{8} \times \frac{w}{8}$, $h_1 w_1 = \frac{h}{4} \times \frac{w}{4}$ 和 $h_0 w_0 = h \times w$, P'_{M0} 为与输入图像分辨率相同的目标检测结果。

2 实验及分析

2.1 数据集及实验平台简介

1) 数据集

本文仿真所用数据集为一类红外序列图像弱小飞机目标检测跟踪数据集^[15],该数据集由国防科技大学 ATR 实验室发布,包含 22 个数据子集,30 组轨迹,16 177 幅图像和 16 944 个目标,其内每帧图像的分辨率为 256×256 。

为了对本文所提网络进行训练和测试,将选用子集 5~22,共包含 14 433 个目标的 14 600 幅红外小目标图像组成实验用数据集,并从中随机抽取 70% 作为训练集,剩余 30% 作为测试集。

在该网络使用 Adam 优化器进行训练时,初始学习率设为 10^{-4} ,共训练 200 轮次,每批训练样本数设为 16,所有训练图像大小设为 256×256 。此外,该数据集内的图像仅标注了目标位置,并未给出尺寸信息,因此算法训练及测试环节将根据目标实际情况,设定其像素为 3×3 或 5×5 ,并选取图中以目标为中心的图块作为感兴趣的局部区域,所划定的局部区域分别为 15×15 或 17×17 。

2) 实验平台

本文提出的红外小目标检测算法采用 Python 编程实现,程序调试及训练、测试实验开展所用软硬件环境为: CUDA 版本为 11.2, Pytorch 版本为 1.17; CPU 为 Intel (R) Core (TM) i5-12600U @ 3.70 GHz, 32 GB RAM,显卡使用 RTX 3060。

2.2 评价指标

一般而言,目标检测的精度及难易程度与红外图像中目标和背景的差异程度密切相关。为了客观地评价检测性能,可先依据局部信杂比(local signal-to-clutter ratio, LSCR)对图像复杂度进行简要评估,其计算公式如(9)式所示。

$$R_{\text{LSC}} = \frac{|\mu_{\text{target}} - \mu_{\text{patch}}|}{\sigma_{\text{patch}}} \quad (9)$$

式中: μ_{target} 表示图像中目标区域的亮度平均值;

μ_{patch} 表示所提取图像块的亮度平均值; σ_{patch} 表示图像块中像素点亮度的标准差。

实际上,采用 LSCR 只能在一定程度上反映检测难易程度,为了做更进一步评估,还可以使用平均方差加权信息熵 (average variance-weighted information entropy, AVWIE)。要计算平均方差加权信息熵,需将图像划分成 N 块并计算每个图像块的方差加权信息熵

$$H_l(I) = -\frac{1}{256} \sum_{s=0}^{255} (s - \bar{s})^2 p_s \lg p_s \tag{10}$$

式中: s 表示灰度值; p_s 表示灰度值为 s 的像素点分布比例; $\bar{s}$ 表示图块的像素平均灰度值。

图像的平均加权信息熵即分割后各图像块方差加权信息熵的平均值。

$$H(I) = \frac{1}{N} \sum_{k=0}^N H_l(I_k) \tag{11}$$

式中: I_k 表示第 k 个图像块, N 表示图像块的数目。

本次仿真所用数据集中图像分辨率为 256×256 ,故选择 $N = 16$,用于计算实验用图像的平均加权信息熵。

此外,文中所提算法是基于分割思想实现检测的,因此还可采用通用的图像分割评价指标评价算法,如精确率 (Precision)、召回率 (Recall) 和 F-度量值 (FM),相应公式分别为

$$P_{\text{re}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{12}$$

$$R_{\text{ec}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{13}$$

$$F_{\text{M}} = \frac{2 \times P_{\text{re}} \times R_{\text{ec}}}{P_{\text{re}} + R_{\text{ec}}} \tag{14}$$

式中: N_{TP} 表示归类为目标且判定正确的像素点数目; N_{FP} 表示归类为目标但实际为背景的分类错误

的像素点数目; N_{FN} 表示实际为目标但是漏检的像素点数目。

事实上,精确率或召回率可以在一定程度上反映算法的检测性能,但即便精确率或召回率两者数值均较大,也并不意味着检测方法真出色。F-度量值常被认为是衡量分割整体质量的可靠指标,因此文中也将其作为检测算法性能的指标之一。

最后,为了更全面地反映算法检测性能,文中将分析评价目标检测性能的 ROC 曲线,其中检测率 (P_{d}) 和虚警率 (F_{a}) 的求解如 (15) ~ (16) 式所示。

$$P_{\text{d}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{15}$$

$$F_{\text{a}} = \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TN}}} \tag{16}$$

式中, N_{TN} 表示非目标物体被正确地识别为非目标的数量。

2.3 实验分析

为了验证所提方法的性能,在此将基于 ISTS-UT 网络的检测算法与其他 7 种检测算法进行比较。这 7 种方法可以分为 4 类:①基于形态学滤波的算法 (Top-Hat);②基于对比度增强原理的算法 (MPCM、MDBWCM);③基于特征增强的算法 (IPT);④基于 CNN 的目标分割方法 (MDvsFA-cGAN、LSPM、DNA)。

对比分析将在 2.1 节中提到的数据集基础上展开,所用测试数据集包含 4 380 幅图片,4 330 个目标。此外,由于所有对比算法都会先求预测图,而后通过设置阈值实现对小目标及背景的分割,在此将各算法中的阈值设为能使 F-度量值的平均值达到最大的数值。基于前述条件,所得的 8 种不同检测算法的定量评估结果见表 1。

表 1 不同红外小目标检测算法性能比较

算法	门限	检测率	虚警率	精确率	召回率	F-度量值
Top-Hat	0.65	0.726 0	8.73×10^{-5}	0.373 0	0.100 0	0.138 0
MPCM	0.82	0.580 0	4.62×10^{-5}	0.367 0	0.098 0	0.135 0
MDBWCM	0.85	0.581 0	3.31×10^{-5}	0.273 0	0.494 0	0.316 0
IPT	0.58	0.778 0	1.13×10^{-4}	0.426 0	0.129 0	0.174 0
MDvsFA-GAN	0.27	0.871 4	3.62×10^{-6}	0.776 7	0.734 0	0.729 6
DNA	0.55	0.927 0	2.47×10^{-6}	0.777 7	0.790 1	0.759 4
LSPM	0.44	0.989 1	1.53×10^{-7}	0.937 9	0.874 1	0.894 2
ISTS-UT	0.62	0.997 2	2.82×10^{-7}	0.912 7	0.921 0	0.909 4

显然,本文算法及表 1 中基于 CNN 的小目标检测方法的多项评估指标明显高于其他类型的算法。进一步比较发现,本文基于 ISTS-UT 网络实现的检测算法既具有很低的虚警率(2.82×10^{-7}),又具有最高检测率 0.997 2,同时其 F-度量值为所有算法中的最大值 0.909 4。就 F-度量值而言,其测度综合考虑了精确率和召回率的影响,以 LSPM 方法为例,虽然其精确率可以高达 0.937 9,但却是以牺牲召回率(0.874 1)为代价,因此此类方法的 F-度量值为 0.894 2。相比而言,文中 ISTS-UT 算法在精确率和召回率之间做了更好的平衡。

为了呈现不同算法性能受阈值影响的情况,图 6 附上了不同算法 F-度量值随阈值变化的可视化结果。

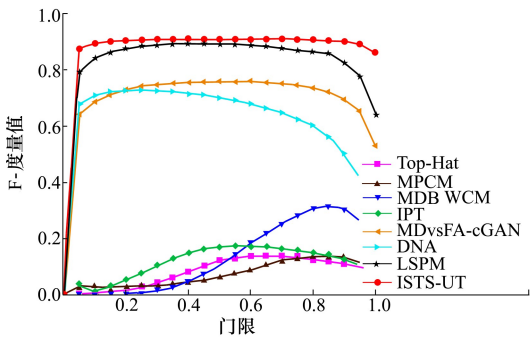


图 6 不同算法 F-度量值对比图

如图 6 所示,在分割阈值为 0.1 时,文中所提 ISTS-UT 算法的 F-度量值就已明显高于其他算法,为 0.875;而当阈值达到 0.25 左右时,其 F-度量值基本趋于稳定,之后仅微弱增加;在阈值为 0.62 时,F-度量值达到最高值 0.91。这说明,相较对比算法,文中算法的 F-度量值受阈值影响最小,算法更具鲁棒性。

除了从图像分割角度进行评估外,还可围绕检测性能进行对比,图 7 展示了不同 ISTD 算法的 ROC 曲线。

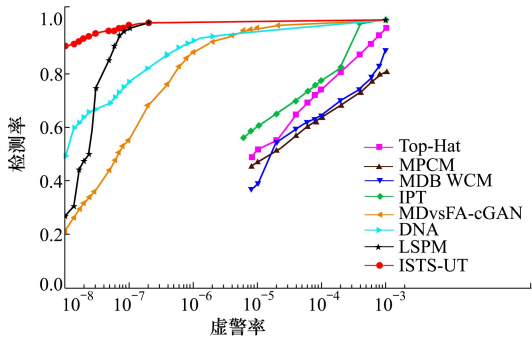


图 7 不同红外小目标检测算法的 ROC 曲线对比图

观察图 7 发现,文中算法的 ROC 曲线最接近图片的左上角,曲线下的面积最大,曲线波动相对较小,呈现了最佳的检测性能及稳定的检测效果。观察 P_d 及 F_d 参数可知,文中算法在取得高检测率的同时,还具有较低的虚警率。事实上,文中算法在检测目标的同时,还对图像背景和噪声进行了有效抑制,而这也说明,采用基于注意力的 Transformer 网络的确能通过捕获更大范围下的图像特征,实现对背景和红外小目标的有效区分。正是因为文中算法是基于整幅图像的特征来检测小目标的,所以它对干扰及复杂背景的抑制能力,明显优于传统的基于局部对比度检测小目标的方法。此外,相比其他几种基于 CNN 的目标检测方法,文中的 ISTS-UT 检测算法因不涉及扩大感受野而引起的小目标特征丢失,避免了由此导致的检测失败。总之,基于 CNN 的检测算法性能明显优于传统检测算法,而文中算法性能优于其他同类对比算法。

为了进一步呈现文中算法与其他算法间检测效果的不同,还从 5 个具有代表性的测试数据集中各选 1 幅图像,如图 8 所示,并依次展开定性分析。由于目标检测的精度及难易程度与红外图像中目标和背景的差异程度密切相关,在此先就图 8 中 5 幅图像相关参数进行测算,具体结果见表 2。

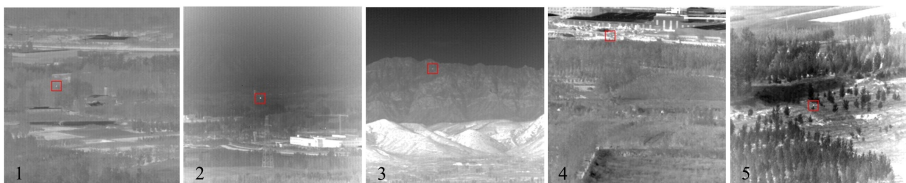


图 8 红外小目标图像(红框标注了目标的位置)

表 2 不同图像参数对比

图像号	平均局部 信杂比(LSCR)	平均方差加权 信息熵(AVWIE)
1	1.751 0	4.484
2	2.677 0	8.283
3	1.638 0	12.258
4	0.112 6	18.009
5	1.603 0	28.706

通常情况下,图像 LSCR 越小表示目标和背景的差异越小,目标检测难度越大。但是,其实采用 LSCR 只能在一定程度上反映检测难易程度,因为该参数的测定只关注了目标在局部区域的对比度信息,并未考虑整幅图像中复杂背景对目标检测的影响。比如,表 2 中图像 3 和 5 的 LSCR 分别为 1.638 和 1.603,虽然二者数值相近,但是由于图像 3 背景为相对平缓的山川,而图像 5 背景为复杂的森林、草地并含有大量噪声点,所以 2 幅图的检测难度其实差异很大。为此,本次实验使用平均方差加权信息熵来描述 1 幅图像局部与整体像素点的灰度变化情况,进一步评估图像的复杂程度及检测难度。总之,依据表 2 的测算结果可见,图像 5 最为复杂且最难检测。

针对这 5 幅图像,使用表 1 中的 8 种检测算法开展小目标检测,发现:首先,文中算法及基于 CNN 的检测算法(MDvsFA-cGAN 算法、DNA 算法、LSPM 算法、ISTS-UT 算法)因具有较强的特征学习能力,可以更好地区分目标和背景,更大程度地抑制背景和噪声,进而能够更为准确地检测目标。与之相比,传统的小目标检测算法(Top-Hat 算法、MPCM 算法、MDBWCM 算法、IPT 算法)检测性能较差。其次,红外小目标图像的复杂程度会影响检测算法的性能。对于背景灰度分布比较均匀并且变化比较平缓的图像(如图 8 中图像 1 所示),其平均方差加权信息熵比较低,虽然小目标的 LSCR 也不高,但大部分算法对其都能达成良好的检测;而随着背景复杂度的提升,例如在局部区域增加了高亮且结构复杂的建筑物图像(如图 8 中图像 2 所示),虽然小目标的 LSCR 较高,但在使用传统算法检测时,可能会出现检测到目标的同时,误检率也急剧增大的情况。对于背景更为复杂,存在大范围复杂山体结构并包含大量噪声的图像(如图 8 中图像 3 所示),由于其

平均方差加权信息熵显著变大,同时目标的 LSCR 较低,采用基于形态学滤波和局部对比度测量的方法往往很难有效抑制噪声,因此检测效果有限;而基于 CNN 的检测算法因受到卷积运算机制制约,只关注了感受野范围内的局部特征,也可能遇到与传统算法类似的问题。随着图像复杂度的增大,甚至出现目标与复杂背景(建筑物、森林)重合的情况时,传统检测算法已无法从中有效提取小目标的特征了,基于 CNN 的几种对比算法也会因小目标融入背景、不具明显特征,而面临目标特征提取困难,影响检测效果。但文中基于 ISTS-UT 网络的检测算法由于引入了注意力机制,能通过对整幅图像像素点之间相关性信息的分析,有效识别图像背景,确保对小目标的分割,达成准确检测的目的。

此外,从算法的复杂度来看,文中所提算法也较 3 种基于 CNN 的对比检测方法更好,算法间的参数量和计算量对比结果如表 3 所示。

表 3 不同检测算法复杂度对比

算法	参数量	计算量
MDvsFA-cGAN	3 138 497	20 578 094 284
LSPM	31 140 801	61 706 938 368
DNA	4 696 837	14 291 800 896
ISTS-UT	36 332 711	10 205 670 144

可见,文中所提算法参数量最大。这与 Transformer 引入的全局自注意力计算及其编码器-解码器结构有关。不过,因为所提算法并未采用深层卷积,而且在每层的 Transformer 编解码模块中,还会根据一维化后的特征数量适当减少 Transformer 内部自注意力运算的循环次数,所以,即使文中算法的参数量最大,可是对比其他 3 种算法,其运算量却最小。

3 结 论

针对红外小目标检测中目标特征不明显且易受杂噪及背景干扰的问题,本文提出了一种基于深度学习的红外小目标检测算法。该算法借鉴图像分割网络 U-Net 的设计思想,构建了一种 U 型多尺度 Transformer 网络,用以对红外小目标图像进行粗粒度分析和高层语义分析,以便在强化小目标局部显著性特征的同时,获取图像在不同尺度下的全局特

征。通过这种对图像由细节至整体的分析,文中检测算法呈现出良好的动态注意力分析、全局信息融合能力,不仅稳定、有效地实现了复杂背景下的红外小目标的精准检测,还具备良好的鲁棒性。总之,与

需要预设目标特征的传统算法以及单纯依赖目标特征进行检测的基于神经网络的算法相比,即使被用于检测信噪比低、背景复杂的红外图像,本文所提方法依然呈现了令人满意的检测效果。

参考文献:

- [1] SUN Yang, YANG Jungang, AN Wei. Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model[J]. IEEE Trans on Geoscience and Remote Sensing, 2021, 59(5): 3737-3752
- [2] RIVEST J, FORTIN R. Detection of dim targets in digital infrared imagery by morphological image processing[J]. Optical Engineering, 1996, 35: 1886-1893
- [3] ZENG Min, LI Jianxun, ZHANG Peng. The design of top-hat morphological filter and application to infrared target detection[J]. Infrared Physics & Technology, 2006, 48: 67-76
- [4] WEI Yantao, YOU Xingge, LI Hong. Multiscale patch-based contrast measure for small infrared target detection[J]. Pattern Recognition, 2016, 58: 216-226
- [5] LU Ruitao, YANG Xiaogang, LI Weipeng, et al. Robust infrared small target detection via multidirectional derivative-based weighted contrast measure[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 19: 1-5
- [6] GAO Chenqiang, MENG Deyu, YANG Yi, et al. Infrared patch-image model for small target detection in a single image[J]. IEEE Trans on Image Processing, 2013, 22(12): 4996-5009
- [7] LIU Ming, DU Haoran, ZHAO Yuejin, et al. Image small target detection based on deep learning with SNR controlled sample generation[J]. Current Trends in Computer Science and Mechanical Automation, 2017, 1: 211-220 (in Chinese)
- [8] MCINTOSH B, VENKATARAMANAN S, MAHALANOBIS A. Infrared target detection in cluttered environments by maximization of a target to clutter ratio(TCR) metric using a convolutional neural network[J]. IEEE Trans on Aerospace and Electronic Systems, 2021, 57(1): 485-496 (in Chinese)
- [9] JU Moran, LUO Jiangning, LIU Guangqi, et al. ISTDet: an efficient end-to-end neural network for infrared small target detection[J]. Infrared Physics & Technology, 2021, 114(7): 103659
- [10] DAI Yimian, WU Yiquan, ZHOU Fei, et al. Asymmetric contextual modulation for infrared small target detection[C]//IEEE Winter Conference on Application of Computer Vision, 2021
- [11] LI Boyang, XIAO Chao, WANG Longguang, et al. Dense nested attention network for infrared small target detection[J]. IEEE Trans on Image Processing, 2022, 32: 1745-1758
- [12] ASHISH Vaswani, NOAM Shazeer, NIKI Parmar, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017
- [13] DOSOVITSKIY Alexey. An image is worth 16×16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations, 2021
- [14] YUE Liyuan, CHEN Yunpeng, WANG Tianwei, et al. Tokens-to-token ViT: training vision transformers from scratch on ImageNet[C]//Proceedings of International Conference on Computer Vision, 2021: 538-547
- [15] HUI Bingwei, SONG Zhiyong, FAN Hongqi, et al. A dataset for dim-small target detection and tracking of aircraft in infrared image sequences[EB/OL]. (2020-08-17)[2023-09-23]. <https://doi.org/10.11922/csdata.2019.0074.zh>

Infrared small target detection algorithm with U-shaped multiscale transformer network

DUAN Peipei¹, ZHANG Yan², LUO Mingshi¹, YAN Xiaoying¹

(1.School of Computer Science, Xi'an Shiyou University, Xi'an 710065, China;
2 School of Electric Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an 710021, China)

Abstract: To solve the problem of small targets feature extraction and the susceptibility of targets to being overwhelmed by noise and complex backgrounds, a detection method with U-shaped multiscale transformer network is proposed. Based on the U-shaped multiscale network architecture, the proposed method uses convolution operations to extract and enhance local salient features of small targets. Concurrently, it uses the Transformer mechanism to model global image features, facilitating the extraction and suppression of the image background. Subsequently, through self-attention operations on target confidence maps and feature maps, fusion of shallow and deep features in images is achieved. This accomplishes pixel-level segmentation of infrared small targets, fulfilling the purpose of target detection. Experiments demonstrate in infrared sequence image dim and small aircraft target detection and tracking data set, even when applied to infrared images with complex background and noisy, our method outperforms the state-of-the-art detection methods. The method shows good robustness and high detection accuracy. When the threshold is selected to maximize the average of FM, the detection rate of our method reaches 0.997 2, its false alarm rate is 2.82×10^{-7} , the precision rate is 0.912 7, and the recall rate is 0.921.

Keywords: infrared small target detection;image segmentation;deep learning;self-attention mechanism

引用格式:段沛沛, 张严, 雒明世, 等. 基于 U 型多尺度 Transformer 网络的红外小目标检测算法[J]. 西北工业大学学报, 2025, 43(1): 154-162
DUAN Peipei, ZHANG Yan, LUO Mingshi, et al. Infrared small target detection algorithm with U-shaped multiscale transformer network[J]. Journal of Northwestern Polytechnical University, 2025, 43(1): 154-162 (in Chinese)