激光与光电子学进展
先进成像

基于实时语义分割的红外小目标检测算法

邵斌[1,2,3]，杨华[1,2,3]，朱斌[1,2,3*]，陈熠[1,2,3]，邹融平[1,2,3]
[1]国防科技大学电子对抗学院，安徽 合肥 230037；
[2]脉冲功率激光技术国家重点实验室，安徽 合肥 230037；
[3]红外与低温等离子体安徽省重点实验室，安徽 合肥 230037

摘要 语义分割网络对图像进行像素级分类，相较于目标检测其对于目标的精准定位更有优势，因此在红外小目标检测中发挥着重要作用。针对红外小目标的特点，提出一种基于实时语义分割的红外小目标检测网络。该网络基于双分支特征提取结构，采用渐进式特征融合模块和改进的Dice损失函数，使红外小目标分割的速度与效果达到良好的平衡。实验结果表明，该算法在较小参数量和计算量的情况下相较于FCN、ICNet、BiSeNet V2、STDCNet、TopFormer等5种算法达到较高的精度，在实际采集的红外小目标数据集上，其推理帧率相较于传统的FCN提升44%，达到117 frame/s，且红外小目标的交并比相较于与其推理帧率相近的TopFormer提升49%，有利于语义分割在红外小目标检测的实际应用。

关键词 图像处理；红外小目标；实时语义分割；双分支特征提取；渐进式特征融合
中图分类号 P391.9 **文献标志码** A **DOI**：10.3788/LOP221958

Infrared Small Target Detection Algorithm Based on Real-Time Semantic Segmentation

Shao Bin[1,2,3], Yang Hua[1,2,3], Zhu Bin[1,2,3*], Chen Yi[1,2,3], Zou Rongping[1,2,3]
[1]*College of Electronic Engineering, National University of Defense Technology, Hefei* 230037, *Anhui, China*;
[2]*State Key Laboratory of Pulsed Power Laser Technology, Hefei* 230037, *Anhui, China*;
[3]*Key Laboratory of Infrared and Low Temperature Plasma of Anhui Province, Hefei* 230037, *Anhui, China*

Abstract Semantic segmentation network classifies images at the pixel level, which has more advantages for accurate target location than target identification, thus playing an essential role in infrared small target detection. According to the characteristics of an infrared small target, a novel infrared small target detection network based on real-time semantic segmentation is proposed. A good compromise between the speed and impact of infrared tiny target segmentation is achieved by the network, based on the dual branch feature extraction structure, using the progressive feature fusion module and enhanced Dice loss function. The experimental results demonstrate that the algorithm achieves high accuracy compared with five algorithms, namely FCN, ICNet, BiSeNet V2, STDCNet, and TopFormer for small parameters and calculation. The proposed algorithem is advantageous for the practical application of semantic segmentation in infrared small target detection because its reasoning frame rate on the actual collected infrared small target dataset is 44% higher than that of traditional FCN, reaching 117 frame/s, and the intersection and merging of infrared small targets are 49% higher than that of TopFormer with the similar reasoning frame rate.

Key words image processing; infrared small target; real-time semantic segmentation; dual branch feature extraction; progressive feature fusion

1 引 言

红外成像探测系统是城市安全防护、边境巡查、战场侦察的常用手段。探测系统通常距离目标较远，在空域背景下，具有潜在威胁的飞行物例如无人机通常只占有很少的像素。因此红外成像后，针对无人机等红外小目标进行快速检测对于安全行动的决策有着重要意义。传统的红外小目标检测算法[1-2]通常依赖手

收稿日期：2022-06-30；**修回日期**：2022-08-24；**录用日期**：2022-08-31；**网络首发日期**：2022-09-10
基金项目：国家自然科学基金(61307025)
通信作者：*zhubineei@163.com

工提取特征处理图像，当应对真实环境的复杂场景时鲁棒性不足。近年来随着深度学习在图像特征抽取上展现出的巨大优势，许多研究者将深度学习技术应用到红外小目标检测领域，提出基于目标检测网络的红外小目标检测算法和基于语义分割网络的红外小目标检测算法[3-5]。相较于目标级的检测，语义分割将图像逐像素分类对于红外小目标定位更为准确。

由于语义分割网络对图像逐像素处理，计算量巨大，算法处理图像的帧率无法满足在真实环境中部署的要求。为了提高模型的推理速度，通常有两种途径：限制输入图像的大小和通道裁剪。image cascade network（ICNet）[6]使用图像级联的方式加速算法，借助低分辨率的图像提高分割效率，使用高分辨率的图像提升分割效果，在GPU上达到实时的推理效果。bilateral segmentation network（BiSeNet V2）[7]设计双分支的特征提取网络，细节分支保留高分辨率的特征表达，语义分支获取深层语义信息，而后将细节信息和语义信息直接融合，最终网络达到精度和速度的平衡。short-term dense concatenate network（STDCNet）[8]在编码器中通过逐步降低特征图的维度减小参数量，在解码器中细节引导模块将空间细节信息整合到网络的浅层特征图中加快收敛，网络在分割准确性和推理延迟之间取得良好的平衡。token pyramid vision transformer（TopFormer）[9]将卷积提取的不同尺度的特征图作为token输入加快计算，并结合卷积结构，在多个公开语义分割数据集上优于基于卷积神经网络（CNN）和ViT[10]的网络。虽然上述方法在针对一般物体的公开大型数据集上有较好的表现，但是在红外小目标图像上由于缺乏针对红外小目标特征的设计，在快速下采样阶段获取特征图时容易丢失目标，在实际复杂环境中出现红外小目标的漏检和误检。

为解决上述问题，本文针对红外小目标的特点提出一种基于实时语义分割的红外小目标检测网络（RTIRSeg）。该网络采用双分支特征提取结构获取红外小目标图像的细节信息和多层次语义信息，然后采用渐进式特征融合模块对各阶段语义信息和细节信息进行通道以及空间融合，还通过改进损失函数缓解红外小目标图像中正负样本不均衡的问题，实现对于红外小目标图像推理速度和检测效果的平衡。

2　网络结构设计

RTIRSeg网络的总体结构如图1所示，网络基于编码-解码架构[11]，包括3个部分：用于提取红外小目标特征的骨干网络（backbone）；用于融合不同分支提取的特征的颈部（neck）结构；用于红外小目标图像分割的分割头（head）。骨干网络由两个特征提取分支组成，一个是细节提取分支，另一个是提取多层次语义信息分支；颈部结构的作用是将两种来自骨干网络提取的包含红外图像不同信息的特征在特征融合模块中进行融合；分割头用于特征插值以及像素语义分割。

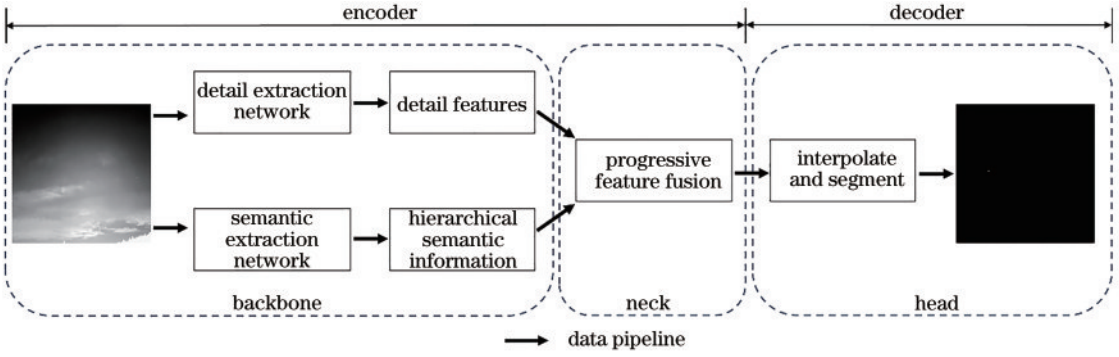

图1　RTIRSeg网络结构

Fig. 1　Network structure of RTIRSeg

细节提取分支为了尽可能地保留红外小目标图像中的细节特征，采取减少特征分辨率降低次数并且扩大特征通道容量的做法，该分支结构如图2所示。红外小目标尺寸远小于常规尺寸目标，使用池化降低分辨率可能会导致目标的丢失，因此细节提取分支对于输入网络的红外小目标图像，通过步长为2的3×3卷积使分辨率下降，同时将通道数扩展为64，再使用卷积核大小为3且步长为1的卷积对得到的特征进行空间维度的融合。重复上述特征分辨率减少及空间细节信息融合操作，最终输出通道数为128、大小为64×64的特征图。此部分基础计算模块为卷积模块ConvM，如式1所示：

$$\mathrm{ConvM}=\mathrm{ReLU}\left\{\mathrm{BN}\left[\mathrm{Conv}\left(c_{\mathrm{in}},c_{\mathrm{out}},ks,s,p\right)\right]\right\},\tag{1}$$

式中：Conv是普通卷积操作；BN[12]为批量归一化操作；ReLU[13]为激活函数；c_{in}为输入特征图的通道数；c_{out}为输出特征图的通道数；ks为卷积核的大小；s为卷积操作的步长；p为在特征图周围补充的像素的大小。

语义提取分支为了获得红外小目标图像的多种感受野特征进行快速下采样同时减少特征通道数保证模

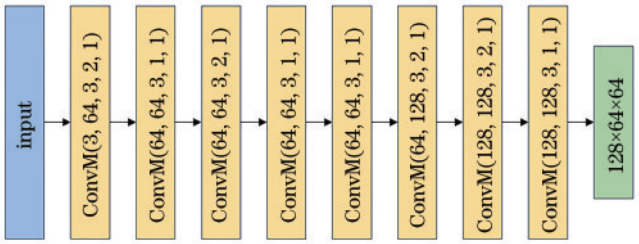

图 2　细节特征提取网络

Fig. 2　Detail feature extraction network

型轻量化和实时性，此部分结构如图 3 所示。该分支由 5 个不同的阶段组成，每个阶段由特征分辨率缩减和空间特征融合构成，每阶段都将上一阶段的特征图通道数加倍，然后将包含不同层次语义信息的特征图输出给特征融合模块。前两阶段特征图缩减时为减少特征信息的丢失，使用的是 3×3 的步长为 2 的卷积；后面 3 个阶段随着特征图通道数增加，为继续获得深层语义信息同时尽可能地缩减网络模型参数量，使用最大池化（MaxPool）进行特征图的下采样。

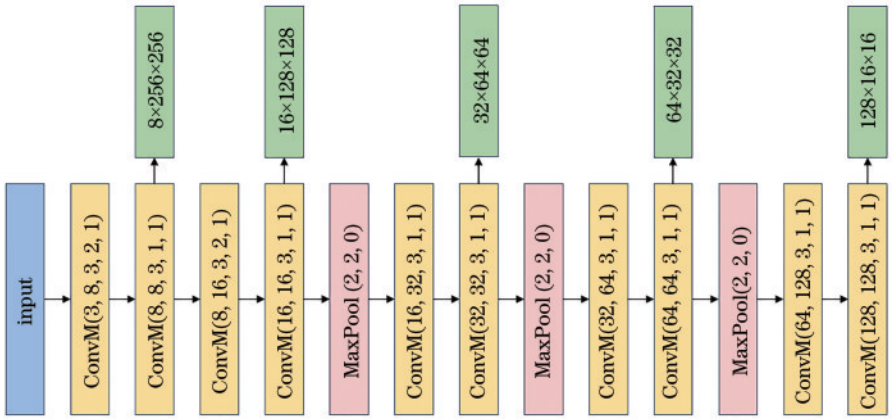

图 3　语义信息提取网络

Fig. 3　Semantic information extraction network

颈部结构为了充分利用提取的细节信息和语义信息，减少因特征丢失造成的漏检误检采取渐进式特征融合方式，其结构如图 4 所示，其中 BiInter 代表的是双线性插值，DSConv 代表的是深度可分离卷积[14]-BN-ReLU 组合的模块，深度可分离卷积可以减小模型的参数量。融合模块首先对语义提取特征分支提取的不同层次的语义信息进行由浅到深的渐进式融合，整体步骤为先使用深度可分离卷积扩展浅层语义的特征图

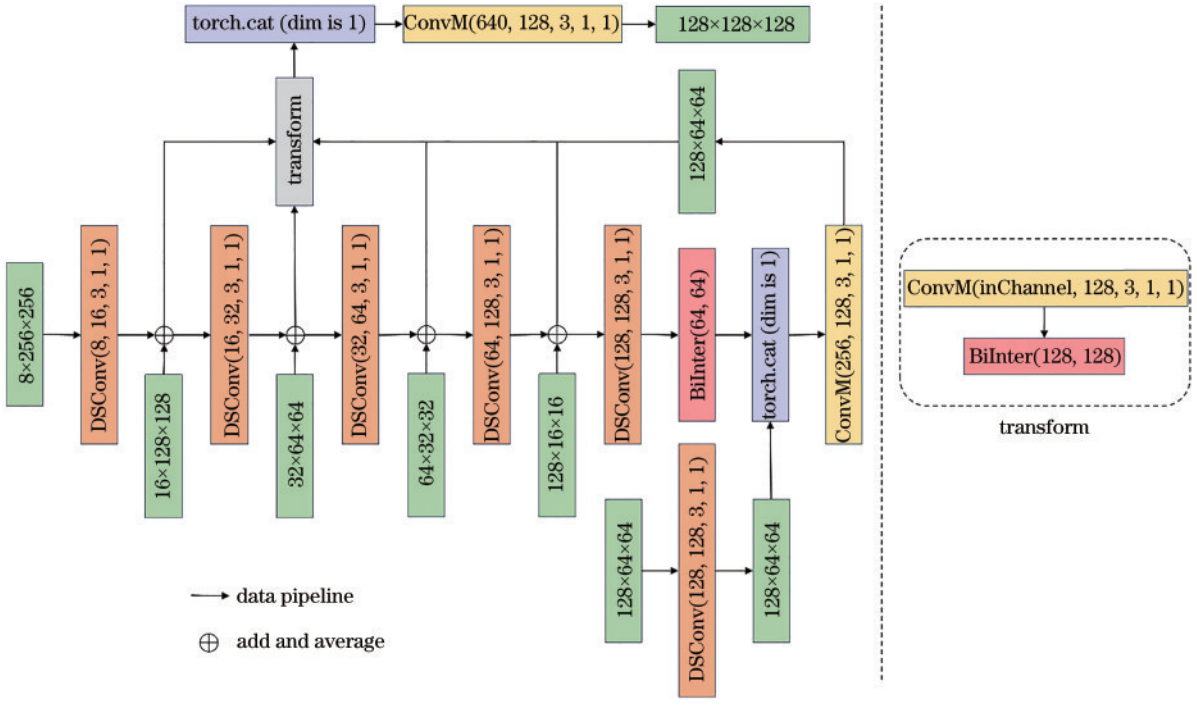

图 4　渐进式特征融合模块

Fig. 4　Progressive feature fusion module

通道数，使其与更深一层语义特征的通道数匹配同时降低特征分辨率，然后将变换后的特征图和更深一层特征图逐元素相加再平均，得到新的融合前两层的特征图，逐层重复此操作。最深层语义特征图插值到和红外小目标细节分支提取的特征图相同尺寸，此时将两者进行通道维度的拼接，再进行通道维度的特征融合。最后将每个阶段融合后的特征图通过卷积以及插值运算变换为相同大小的特征图，然后对上述相同大小的特征图进行通道维度的拼接，再使用卷积进行空间尺度的融合以及特征通道维度缩减，生成用于分割的特征图。

最后分割头对来自融合模块的特征使用双线性插值将分割图上采样至和原图一样大小，并求取损失。由于红外图像中小目标和背景所占像素比例差距很大，会造成正负样本不平衡，模型将会更倾向于将像素预测为负样本即背景以减小网络损失，造成漏检。因此为了缓解红外小目标图像固有性质对于像素分类预测的不利影响，对 Dice 损失函数[15]进行修改以适应红外小目标的分割检测：

$$L_{\mathrm{imp}}=1-2\frac{\sum_{c=1}^{2}w_c\sum_{n}T_{cn}P_{cn}}{\sum_{c=1}^{2}w_c\sum_{n}T_{cn}+P_{cn}},\tag{2}$$

式中：T_{cn}代表类别 c 在位置 n 的真实值；P_{cn}代表类别 c 在位置 n 的预测概率；w_c代表类别 c 的权重。这里按照 SPIE 对红外小目标的定义中背景和目标像素数的关系确定目标和背景类别的权重，取背景与目标权重比为 1∶808。

3 实验与分析

3.1 实验设置

3.1.1 数据集

所使用的数据集是在外场真实环境采集的红外小目标数据，使用 FLIR sc7000 红外热像仪采集无人机在郊外山脉背景、纯净天空背景、城市建筑背景、含云天空背景下的红外小目标视频数据，目标类型包括 DJI MINI2 和 DJI 御 2 行业双光版无人机。经筛选清洗得到 1412 幅图像，红外小目标数据图像大小裁剪为 512×512，数据组织成 pascal voc 格式，标注生成图像的 mask 图。按照 8∶1∶1 的比例将数据集划分为训练集、验证集和测试集。

3.1.2 参数与环境

所有实验均在 Ubuntu 系统下使用深度学习框架 PyTorch 进行实验，具体信息如表 1 所示。实验均采用 Adamw 优化器，每个批次样本数设置为 4，learning rate 为 0.00006，权重衰减为 0.01，迭代次数设置为 80000。

表 1 实验环境

Table 1 Experimental environment

Environment	Configuration
CPU	Intel(R)Xeon(R)CPU W-2223
GPU	NVIDIA GeForce RTX 2080Ti × 1
Operating system	Ubuntu 18.04
Memory	32 GB
Framework	PyTorch v1.9.1
Library	CUDA 10.2、CUDNN 7.6.5

3.1.3 评价指标

采用多种分割相关评价指标，从多方面评价分割网络的表现。

Dice 系数：

$$c_{\mathrm{mDice}}=\frac{1}{c}\sum_{1}^{c}\frac{2N_{\mathrm{TP}}}{2N_{\mathrm{TP}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}},\tag{3}$$

式中：N_{TP}为标签及模型预测均为正样本的像素；N_{FP}为标签为负样本模型预测为正样本的像素；N_{FN}为标签为正样本模型预测为负样本的像素。

pixel accuracy(Acc)为预测类别正确的像素数占总像素数的比例。intersection-over-union(IoU)一定程度上可以反映红外小目标定位的准确性。frame per second(FPS)表示图像分割的速度。floating point operations per second(FLOPs)可以用于衡量网络的计算复杂度。Param 表示模型的参数复杂度。receiver operating characteristic curve(ROC)表示的是模型在选取不同阈值时其对于所分割物体的敏感性。

3.2 实验结果与分析

3.2.1 模块实验验证

采用消融实验来证明所提算法改进部分的有效性，选择 BiSeNet V2 为基线网络，实验结果如表 2 所示。表 2 第 1 行数据表示的是基线网络在红外小目标测试集上的性能，算法推理帧率达到 107 frame·s^{-1}，目标类别的 IoU 为 65%。将基线网络的复杂语义分支换更换为新设计语义信息提取网络，相较于基线网络，其

表 2 消融实验

Table 2 Ablation experiment

Method	IoU of target /%	FPS /(frame·s^{-1})	FLOPs /10^9	Param /10^6
Baseline	60.9	107	15.38	14.76
+ Improved Feature Extraction	61.06	158	14.88	3.93
+ Improved Feature Fusion	68.32	118	11.41	3.04
+ Improved DiceLoss	68.63	117	11.41	3.04

推理帧率增加 51 frame·s^{-1}，参数量减小 10.83×10^{6}，分割精度略有增加。第 3 行网络中添加了渐进式特征融合方式，增加了对于多层次特征图的操作，推理帧率减少 40 frame·s^{-1}，但分割精度提升了 7.26%。最后增加改进后的 Dice 损失后，缓解了红外小目标图像中正负样本不均衡的问题，在参数量和帧率基本不变的情况下，分割精度也提升了 0.31%。

3.2.2　实验结果对比

为了验证所提 RTIRSeg 网络的有效性，选取经典的 FCN 语义分割算法和其他针对常见公开数据集的实时语义分割算法 ICNet、BiSeNet V2、STDCNet、TopFormer 作为对比。为了模拟真实红外小目标检测的应用场景，选取 6 种不同背景下的红外小目标图像，图 5(a)～(e)分别代表的是人工建筑背景、含天空和山脉的郊外背景、少云天空背景、纯净天空背景、含城市建筑天空背景，图 5(f)、(g)代表多云天空背景条件下的红外小目标图像。为了便于观察分析，算法分割出的绿色掩模图和对应场景的原图按 0.5 的透明度进行叠加，并对于红外小目标周围区域进行放大展示。

图 5 展示了不同背景下，各算法的红外小目标检

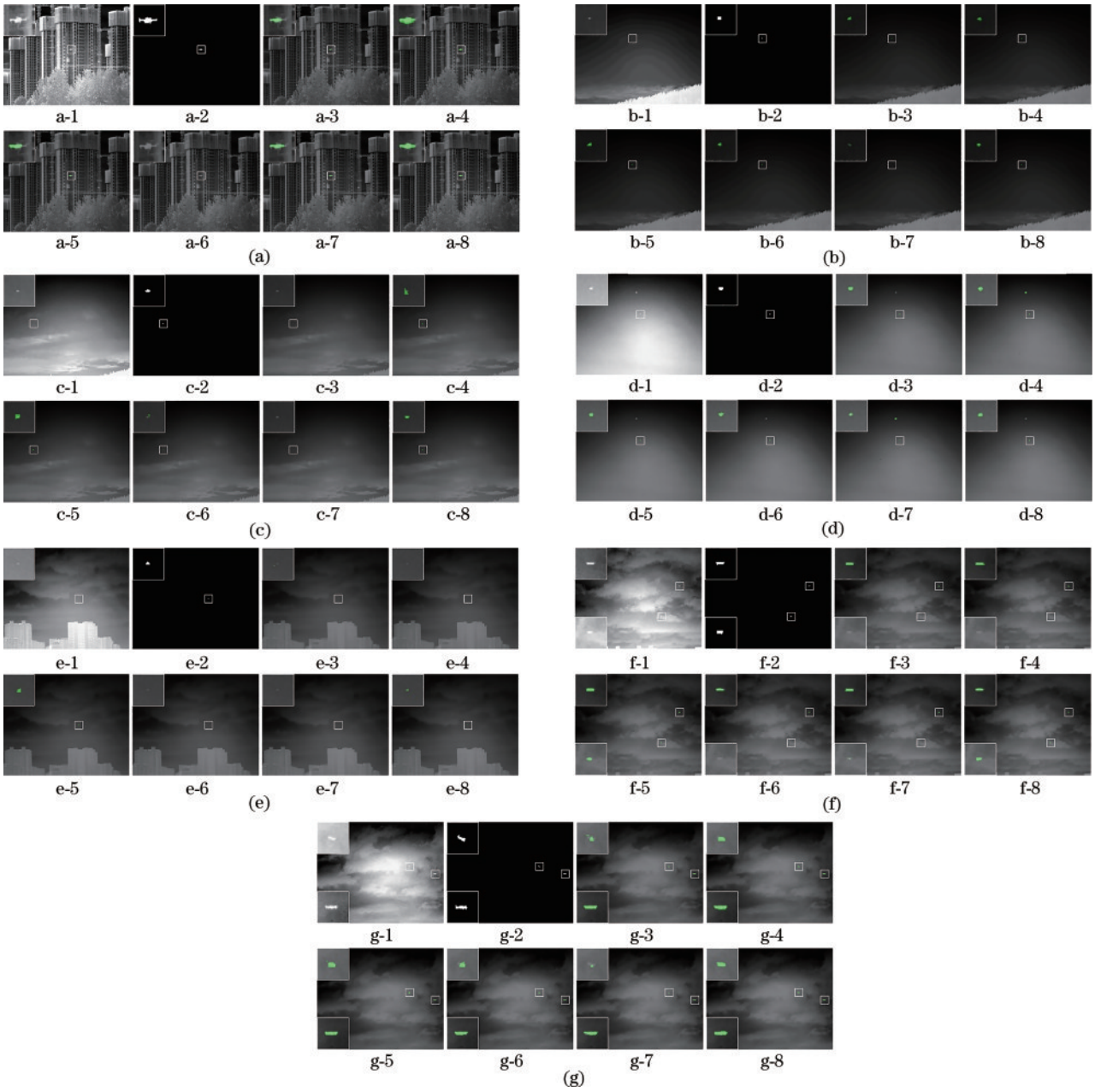

图 5　多场景下所提算法分割结果。(a)人工建筑背景；(b)含天空和山脉的郊外背景；(c)少云天空背景；(d)纯净天空背景；(e)含城市建筑天空背景；(f)(g)多云天空背景

Fig.5　Segmentation results of proposed algorithm in multiple scenes. (a) Artificial construction background; (b) suburban background with sky and mountains; (c) less cloudy sky background; (d) clear sky background; (e) sky background with buildings; (f) (g) cloudy sky background

测结果，其中，图 5(a-1)～(g-1)、图 5(a-2)～(g-2)分别代表的是原图和作为真实分割标签的掩膜图，同一子图下 3～7 分别代表的是 FCN、ICNet、BiSeNet V2、STDCNet、TopFormer在同一背景下的红外小目标检测效果，带有数字标号 8 的图像则是所提算法检测效果。

从不同的红外小目标检测场景来看，图 5(a-1)代表的是复杂建筑背景，图中含有人工建筑以及复杂的树木轮廓模拟，此时基于卷积的实时语义分割算法能够较好分割目标，而 TopFormer模型为了减小计算量，对于图像经下采样后的特征进行自注意力计算，模型可学习参数少，对于特征的表达不够，造成红外小目标的漏检。图 5(b)、(d)中目标点周围干扰少，目标的局部对比度高，所选取的算法都能够对红外小目标进行分割。图 5(c)中由于少量的云层干扰，图像的局部对比度减小，FCN 和 TopFormer出现了漏检现象，如图 5(c-3)、(c-7)所示。图 5(e)中目标属于极小情况，并且目标接近云层，局部对比度进一步减弱，此时大部分语义分割算法失效，基于两个分支的网络可保留细节信息和语义信息，完成了分割。图 5(f)、(g)中多云天气条件下，云层干扰情况使得部分算法出现误分割的情况。

从算法之间的角度看，FCN作为语义分割的开山之作，其对于图像的建模能力不足，输出分割图时上采样倍率过大，分割效果不够精细。从图 5(a-3)～(g-3)可以看出，其在云层干扰、含人工建筑等含有外界非目标物体干扰的场景下出现严重漏检和误检。分析对比图 5(a-4)～(g-4)、(a-6)～(g-6)、(a-7)～(g-7)时发现，ICNet、STDCNet、TopFormer代表的实时语义分割网络，相较于 FCN，在面对复杂地面、云层干扰、多目标能够表现较好的性能，但是由于缺乏细节信息，并且在特征融合时未能将各阶段特征图充分利用造成红外小目标图像特征丢失，从而出现漏检和误检。基于双分支架构的 BiSeNet V2，结合细节特征信息和语义特征信息，在极弱小目标天空中出现漏检现象。所提实时语义分割算法基于语义和细节双分支设计，简化语义特征提取网络，保证图像推理的速度，并重新设计特征融合模块，对各阶段特征进行渐进式融合，充分利用特征信息，在图 5(a-8)～(g-8)中常见红外小目标检测场景中均展现了较好的分割效果，总体分割效果最优。

表 3 为不同网络参数及推理帧率对比。可以看出，随着 FLOPs 和参数量的减小，网络的推理帧率呈现递增的趋势。FCN 相较于其他经过剪枝量化得到的实时语义分割算法参数量是最多的，因此推理速度并没有优势。实时语义分割算法 TopFormer虽然通过极度简化网络模型在所选算法中获得最高的推理帧率，但过少的网络参数使得其无法对特征进行良好的提取、利用，从图 5(a-7)～(g-7)可以看出，其分割的效果相较于其他算法处于劣势。ICNet简化了浮点计算量但由于使用金字塔池化模块[16]使得模型仍保留比较大的参数量，推理帧率相较于 FCN 有所提升。BiSeNet V2、STDCNet 以及所提实时红外小目标语义分割算法在参数量、浮点计算量和推理帧率达到相对的平衡，在参数量较小的情况下有较高的推理帧率。RTIRSeg 在视频连续帧的分割效果如图 6 所示。

表 3　不同网络参数及性能对比

Table 3　Comparison of parameters and performance of different networks

Network	FLOPs /10^9	Param /10^6	Inference speed /(frame·s^{-1})
FCN	247.13	49.49	81
ICNet	19.27	47.82	97
BiSeNet V2	15.38	14.76	105
STDCNet	10.57	8.57	103
TopFormer	0.62	1.37	121
RTIRSeg	11.41	3.04	117

表 4 为 FCN、ICNet、BiSeNet V2、STDCNet、TopFormer以及所提算法在测试集上的部分评价指标。TopFormer在 Dice、Acc、IoU 的数据表现均并不理想。结合图 5(a-7)～(g-7)分析是因为红外小目标是局部特征，TopForme结构针对特征图关注长距离依赖，缺少对于局部信息的关注。而 ICNet、BiSeNet V2 表现得较好，其中，ICNet得益于金字塔池化模块保持了特征图的分辨率，有利于精细化分割，使其在像素分类的精度上达到最优，平均像素精度达到 91.45%。所提 RTIRSeg网络由于重新设计了语义提取分支以及渐进式特征融合模块，在保证网络参数不增加太多的情况下，有较好的分割效果，对于小目标部分，在常见的分割评价指标 Dice 和 IoU 上分别达到 81.17% 和 68.3%，取得了领先的结果。

图 7 为 6 种算法在低虚警率时的 ROC 曲线。ROC 反映的是不同阈值下算法分割的敏感度。横坐标为目标像素虚警率，纵坐标为目标像素正确分类概率，可以看出，随着分类阈值降低，像素分类正检率上升速度快于虚警率，体现了 6 种算法的有效性。TopFormer和 STDCNet由于参数量少，FCN结构过于简单，在虚警率达到 0.000015 时，模型分类的准确率暂时停止增加，维持较低的正检率。ICNet 和 BiSeNet V2 相较于

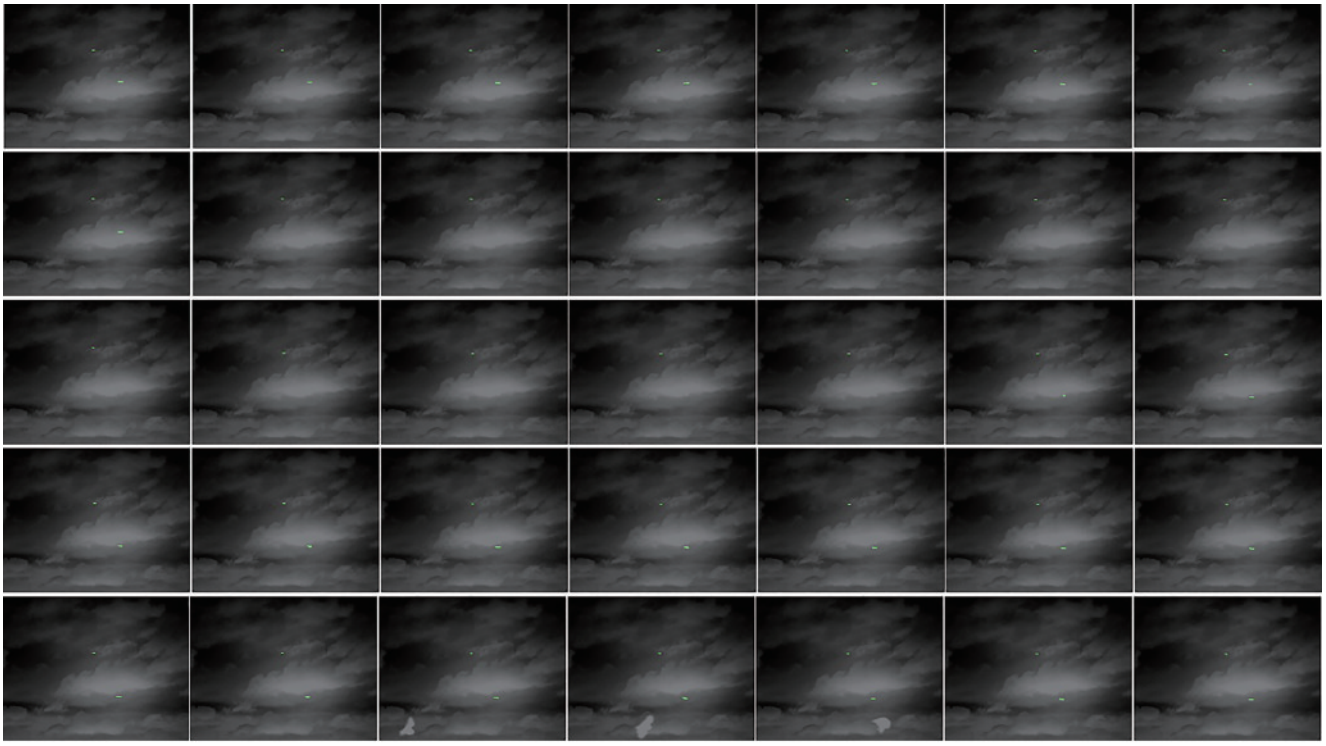

图 6　RTIRSeg在视频连续帧的分割效果

Fig. 6　Segmentation effect of RTIRSeg in continuous frame of a video

表 4　不同方法分割结果评价

Table 4　Evaluation of segmentation results of different methods　　unit: %

Network	Class	Acc	mAcc	Dice	mDice	IoU	mIoU
FCN	Background	99.99	81.69	99.99	84.86	99.99	76.76
	Target	63.38		69.73		53.52	
ICNet	Background	99.99	**91.45**	99.99	87.84	99.99	80.44
	Target	**82.90**		75.69		60.89	
BiSeNet V2	Background	99.99	90.55	99.99	87.85	99.99	80.45
	Target	81.11		75.7		60.9	
STDCNet	Background	99.99	80.09	99.99	83.13	99.99	74.77
	Target	60.19		66.26		49.54	
TopFormer	Background	99.99	78.60	99.99	81.29	99.99	72.77
	Target	57.21		62.59		45.55	
RTIRSeg	Background	99.99	90.42	99.99	**90.58**	99.99	**84.15**
	Target	80.85		**81.17**		**68.30**	

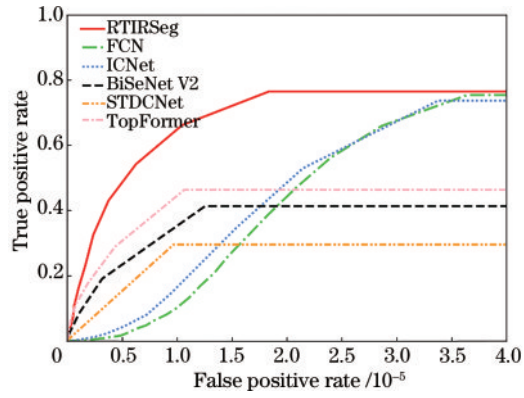

图 7　6种算法的ROC曲线

Fig. 7　ROC curves of six algorithms

前两个模型的参数量和浮点计算量都有所增加，模型的特征提取及表达能力变强，随着阈值降低，在虚警略有增加的情况下像素分类正确率有较大的提升。最后通过代表RTIRSeg的红线可以看到，所提算法在虚警率小的情况下取得良好探测概率，并且随着分割阈值的改变，较好展现模型的泛化能力。

4　结　　论

针对红外小目标检测时使用语义分割算法实时性的问题，提出一种满足多种实际场景检测需求的红外小目标实时语义分割算法。在编码阶段，网络的细节信息提取分支产生分辨率大通道数少的特征图，保留

特征信息;语义提取分支快速下采样获取语义信息,产生多阶段分辨率相对较小、通道数多的特征图,产生多尺度感受野信息。渐进式特征融合模块对提取到的红外小目标图像的细节信息和多层语义信息进行充分融合,提高网络的分割性能。改进的损失函数对红外小目标图像中的样本不均衡进行有效缓解。由于网络设计得合理,在采集的实际场景红外小目标图像中,所提算法能够实现快速语义分割,达到精确检测定位红外小目标的目的。目前,所提算法可在GPU上实现实时分割,实现探测设备轻量化、小型化是接下来努力的方向。

参 考 文 献

[1] Gao C Q, Meng D Y, 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.

[2] He Y J, Li M, Zhang J L, et al. Small infrared target detection based on low-rank and sparse representation[J]. Infrared Physics & Technology, 2015, 68: 98-109.

[3] 宋子壮, 杨嘉伟, 张东方, 等. 基于无监督域适应的低空海面红外目标检测[J]. 光学学报, 2022, 42(4): 0415001.
Song Z Z, Yang J W, Zhang D F, et al. Low-altitude Sea surface infrared object detection based on unsupervised domain adaptation[J]. Acta Optica Sinica, 2022, 42(4): 0415001.

[4] Dai Y M, Wu Y Q, Zhou F, et al. Asymmetric contextual modulation for infrared small target detection [C]//2021 IEEE Winter Conference on Applications of Computer Vision, January 3-8, 2021, Waikoloa, HI, USA. New York: IEEE Press, 2021: 949-958.

[5] Dai Y M, Wu Y Q, 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.

[6] Zhao H S, Qi X J, Shen X Y, et al. ICNet for real-time semantic segmentation on high-resolution images[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11207: 418-434.

[7] Yu C Q, Gao C X, Wang J B, et al. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129(11): 3051-3068.

[8] Fan M Y, Lai S Q, Huang J S, et al. Rethinking BiSeNet for real-time semantic segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 9711-9720.

[9] Zhang W Q, Huang Z L, Luo G Z, et al. TopFormer: token pyramid transformer for mobile semantic segmentation[EB/OL]. (2022-04-12)[2022-05-04]. https://arxiv.org/abs/2204.05525.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//9th International Conference on Learning Representations, May 3-7, 2021, Virtual Event, Austria. [S.l.]: OpenReview.net, 2021.

[11] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[12] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]//ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37, July 6-11, 2015, Lille, France. New York: ACM Press, 2015: 448-456.

[13] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, April 11-13, 2011, Fort Lauderdale, USA. Copenhagen: MLR Press, 2011: 315-323.

[14] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2022-05-03]. https://arxiv.org/abs/1704.04861.

[15] Milletari F, Navab N, Ahmadi S A. V-net: fully convolutional neural networks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV), October 25-28, 2016, Stanford, CA, USA. New York: IEEE Press, 2016: 565-571.

[16] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.