类 HED 网络的热红外图像显著性人体检测深度网络

张 骏[1,2,3]，张 鹏[1,2,3]，张 政[1,2,3]，白云飞[1,2,3]
（1. 中航华东光电有限公司，安徽 芜湖 241002；2. 特种显示国家实验室，安徽 芜湖 241002；
3. 国家特种显示工程技术研究中心，安徽 芜湖 241002）

摘要：热红外图像中的人体目标易于观察显著性强，应用广泛，但受限于热红外设备的硬件，往往图像中的人体目标边缘模糊，检测效果较差，同时因为热红外的特殊成像原理，人体目标检测时极易受到发热物和遮挡物的干扰，检测的精度也无法得到保证。针对上述问题，本文提出了一种类 HED（holistically nested edge detection）的热红外显著性人体检测网络。网络采用类 HED 网络形式，通过将不同比例的空洞卷积编解码模块进行残差相加形式，完成人体目标的检测任务。实验证明该网络可以有效地检测人体目标，准确地预测边缘结构，同时在发热物及遮挡物等环境下也具有较高的检测精度。

关键词：HED；VGG；U-NET
中图分类号：TP183 **文献标志码：**A **文章编号：**1001-8891(2023)06-0649-09

Similar HED-Net for Salient Human Detection in Thermal Infrared Images

ZHANG Jun[1,2,3]，ZHANG Peng[1,2,3]，ZHANG Zheng[1,2,3]，BAI Yunfei[1,2,3]
(1. *Aviation Industry Corp Huadong Photoelectric Company Limited*, *Wuhu* 241002, *China*; 2. *State Special Display Engineering Laboratory*, *Wuhu* 241002, *China*; 3. *Land National Special Display Engineering Research Center*, *Wuhu* 241002, *China*)

Abstract: Human targets in thermal infrared images are easy to observe and have a wide range of applications. However, they are limited by the hardware of thermal infrared devices. The edges of human targets in the images are often blurred and the detection efficiency is poor. Simultaneously, because of the special imaging principle of thermal infrared, human target detection is vulnerable to the interference of heating and occlusion objects and the detection accuracy cannot be guaranteed. In response to the above issues, this study proposes a type of holistically nested edge detection (HED)-thermal infrared saliency human detection network. The network adopted the form of a similar HED network and detected human targets by adding the residuals of different proportions of the hole convolutional codec module. Experiments showed that the network can effectively detect human targets, accurately predict the edge structure, and also have high detection accuracy in an environments with heating objects and obstructions.

Key words: HED, VGG, U-Net

0 引言

显著目标检测（salient object detection， SOD）旨在分割出图像中最具吸引力的视觉目标。在视觉跟踪，图像分割等领域有着广泛的应用[1]。传统的显著目标检测算法多采用超像素相似度、直方图，像素梯度比[2-5]等手工特征的方法进行检测，但在小物体、物体被遮蔽的情况下，检测效果较差。近年来，随着卷积神经网络（convolutional neural networks , CNNs）的发展，特别是在图像分割领域的成功应用，显著目标检测也迎来了新的发展。目前主流的显著目标检测多使用全卷积网络（full convolutional network，FCN）[6]，该类方法采用现有的骨干网络进行不同深层的特征提取，如 VGG，ResNet，DenseNet[7-12]等，网络结构也多为 U 型的编解码结构（Encoder-Decoder）[13]。如 Nian Liu 等人提出的 PICANet[14]通过将多尺度特征与 U-NET 架构结构，整合全局上下文和多尺度的局部上下文

收稿日期：2021-03-29；**修订日期：**2021-04-27.
作者简介：张骏（1983-），男，工程师，硕士，研究方向：图像处理与模式识别。E-mail: zqiniop@163.com。
基金项目：安徽省科技重大专项项目。

来提升检测精度。Mengyang Feng 等人提出 AFNet[15]强调全局特征在显著目标检测中的作用，在网络结构中增加了全局感知模块和注意力反馈模块，以便更精确地探索目标的结构。Xuebin Qin 等人提出 BASNet[16]采用堆叠的 U-Net 形式，将目标检测分成预测-优化模块，使用边界感知的混合损失函数提升检测细节。Jiangjiang Liu 等人提出 PoolNet[17]在 U 型特征网络的基础上加入了基于池化改进的全局引导模块和特征整合模块，进而锐化细节特征。

2017 年 HED[18]网络被提出，网络用于端到端的边缘检测。网络使用多尺度的特征，并且在主干网络中，添加若干输出层网络，再通过一个训练的权重融合函数得到最终的边缘输出[18]。文献[18]强调了在生成输出的过程中通过不断地集成和学习得到更精确的边缘预测图的过程。

红外热图像，因为其特殊的成像原理，使得其应用广泛，尤其是在军事侦察、资源勘探等领域，但是红外热图像缺点明显，如穿透性强，对物体温度感知敏感、对比度低、区域边界模糊等。热红外图像中的人体目标往往显著性强，易于观察，但是边界模糊，极易受到周边发热物的干扰，在遮蔽情况下会出现漏检测和误检测。

针对上述的难点，本文提出了一种类似 HED[18]的深度网络，该网络以 VGG16[7]作为主干网络提取多尺度人体目标特征，同时采用多个 U 型编解码模块残差相加，实现多层次全局/局部上下文特征的提取，用来提升边界感知的精度，最后使用特征融合模块完成目标检测。

1 网络结构

本章将对网络的结构以及损失函数进行介绍。

1.1 网络结构图

该网络由 3 层组成，分别为特征提取层（feature extractor），编解码层（encoder-decoder）以及融合层（fusion），如图 1 所示。

特征提取层，采用 VGG16[7]或 MobileNet V2[19-20]作为特征提取的主干网络。VGG16[7]网络被应用于诸多的视觉任务中，具有良好的特征提取能力[16]。相对于 VGG16 网络，MobileNet 网络使用深度可分离卷积代替了传统卷积操作，在保证特征提取的前提下，进一步减少运算量，是一种轻量化的特征提取网络[19-20]。

编解码层整体采用 FPN（feature pyramid networks）结构[21]，每一尺度的特征采用 U 型编解码网络。融合层采用类 HED 结构，保证边缘信息可以精确检测。

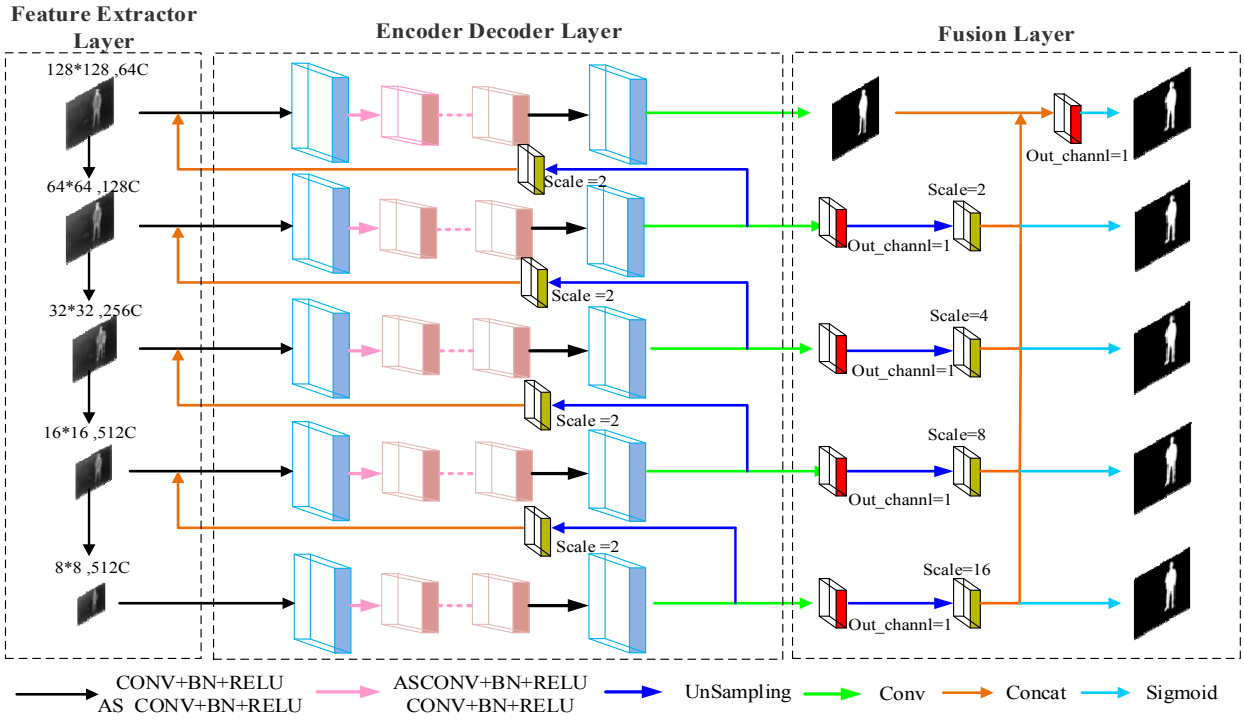

图 1 本文提出的网络架构图

Fig.1 The proposed network architecture

1.2 特征提取层

热红外图像目标易于观察，但轮廓模糊、噪点多、干扰目标多，对于特征提取的要求高。因此在主干网络的选择上，多尺度特征成为首选。文献[14-17]均表明，多尺度特征可以最大程度保证局部和全局上下文信息，对于目标边缘信息的感知提取有着重要的作用。

VGG 网络有着高效的运算效率，较小的存储空间，在诸多视觉任务中被作为多尺度提取的主干网络[16]。

本文中，采用预训练＋调参（Pre-training＋fine-tuning）的迁移学习完成 VGG16 主干网络的特征提取任务。通常来说，直接将预训练模型用来提取特征效果往往不佳，其主要原因在于目标数据集规模、数据内容与预训练数据集之间存在差异。热红外图像与可见光图像差异较大，同时热红外数据集规模较小。文献[8]针对此种情况给出解决方法，即冻结除 FC 层外其他网络的权重，只更新网络结构中的较高层和最后的全连接层，并且取得很好的效果。SSD，YOLO 等算法中均采用更新网络结构，使用 2 层卷积层替代全连接层的方式完成特征提取，效果显著[9]。本文也采用相同的做法，以输入 224×224×3 的图像举例。丢弃 VGG 网络最后的 FC（Full Connect）全连接层，改为增加 2 个卷积操作，分别为 Conv2d（input＝512，output＝1024，kernel_size＝3），Conv2d（input＝1024，output＝512，kernel_size＝1）。

调整后网络结构如表 1 所示。

表 1 中 Conv_add1 和 Conv_add2 分别代表新增的 2 个卷积操作。

5 种不同尺寸的 Feature Map，可以兼顾不同尺寸人体目标的检测，其中 Conv_4 和 Conv_7 观察野较大，用来检测尺寸较小的人体目标，比如远距离的人体目标和遮挡率较大情况下的人体目标，Conv_10 观察野中等，用于检测常用尺寸的人体目标，Conv13 、Conv_add2 观察野较小，用于检测近距离的大尺寸目标，比如近距离出现的人体目标。同时不同尺度的也可以兼顾到浅层和深层特征，浅层特征主要针对目标的边缘，角点等信息，这一类信息对于全局上下文信息依赖大，而深层特征难以描述，更多被认为是局部信息。本网络将前 3 个尺度的特征看作浅层特征代表全局上下文信息，而后两层特征看作深层特征代表局部信息。

表 1 主干 VGG16 网络结构

Table 1 Backbone VGG16 network structure table

Operation	Input Size	Output Size	Output
Input Data	224×224×3		No
Conv_1 (3×3)	224×224×3	224×224×64	No
Conv_2 (3×3)	224×224×64	224×224×64	No
MaxPool(2×2)	224×224×64	112×112×64	No
Conv_3 (3×3)	112×112×64	112×112×128	No
Conv_4 (3×3)	112×112×128	112×112×128	Yes
MaxPool (2×2)	112×112×128	56×56×128	No
Conv_5 (3×3)	56×56×128	56×56×256	No
Conv_6 (3×3)	56×56×256	56×56×256	No
Conv_7 (3×3)	56×56×256	56×56×256	Yes
MaxPool（2×2）	56×56×256	28×28×256	No
Conv_8 (3×3)	28×28×256	28×28×512	No
Conv_9 (3×3)	28×28×512	28×28×512	No
Conv_10(3×3)	28×28×512	28×28×512	Yes
MaxPool(2×2)	28×28×512	14×14×512	No
Conv_11 (3×3)	14×14×512	14×14×512	No
Conv_12 (3×3)	14×14×512	14×14×512	No
Conv_13 (3×3)	14×14×512	14×14×512	Yes
MaxPool(2×2)	14×14×512	7×7×512	No
Conv_add1(3×3)	7×7× 512	7×7×1024	No
Conv_add2 (3×3)	7×7×1024	7×7×512	Yes

1.3 编解码层

该层是整个网络的重要组成部分，通过编解码层，不仅需要完成不同尺度的 Feature Map 的编解码过程，同时也要将浅层特征和深层特征进行合并。

这里我们采用了两种设计思路，第一种是单一尺度的 U 型编解码方式；第二种是不同尺度之间的 FPN 连接方式。

首先是单一尺度的 U 型编解码方式，其结构如表 2 所示。

表 2 单一尺度 U 型编码网络结构

Table 2 Single-scale U-encoded network structure

Operation	Parameters
Conv_1	*K*=3, stride=1, padd=1
Dilation Conv_1	*K*=3, dilation=2**i*, padd=2*(5-*i*) (*i*=1,2,3,4)
Dilation Conv_2	*K*=3, dilation=2**i*, padd=2* (5- *i*) (*i* =1,2,3,4)
Conv_2	*K*=3, stride=1, padd=1
Conv_3	*K*=1, stride=1

表 2 中空洞卷积 Dilation Conv 中空洞系数采用

2×(5−i)的形式。其中 i 表示尺度的层数。

这种方式类似于 U-Net 网络，与 U-Net 网络的不同之处在于卷积操作的选择上以及网络的深度。如图 2 所示。

图 2 中(a)设计的 U 型编解码网络结构，其中 $W\times H$ 为每次操作的尺寸大小，(b)为传统的 U-Net 网络结构，可以清晰地看到(b)网络结构丢弃了 Skip-Connections，使用空洞卷积（Dilation Conv）代替了 Conv，减少了网络的深度。传统的 U-Net 网络使用 Skip-Connections 用于增加全局上下文的信息，兼顾浅层特征和深层特征，而代价就是网络的数据量在不断的增大[11]。而这里丢弃了 Skip-Connections，但为了保证全局上下文信息不会丢失，本文采用 Dilation Conv 替代 Conv[22]，通过设置逐步增大的空洞系数来增加感受野的范围，以达到增加全局上下文信息的效果[18]。其原因如下：首先：弥补特征提取层 MaxPool 操作引起的模糊效果。池化操作能扩大感受野，但会造成模糊，特别是在边缘信息的特征提取上，而 Dilation Conv 操作通过设置空洞系数可以最大程度弥补池化操作（Pooling）带来信息损失[21]；其次，不增加模型参数的同时建立每个像素与周边像素之间的关联性。传统的 Conv 操作如果想增大像素间的关联性，最直接的方法就是扩大卷积核的尺寸，但会带来更多的参数，使得计算变慢，而 Dilated Conv 操作通过设置空洞系数就可以实现这一功能，同时运算量不变[22]；最后：适当变换的空洞系数更适合于热红外图像。在多尺度方法中，每个特征尺度的网络结构和相关参数通常都是相同的，从直观上而言，这是一种自然的选择，因为每个尺度下，都在解决同一问题。同时如果每个尺度上参数都在变换会导致参数不稳定，解决方案空间受限，算法鲁棒性降低[23]。而热红外图像的成像原理的特殊性，从本质上就决定了图像空间易变性。而实验证明不同比例的空洞系数更适合热红外图像。

Dilated Conv 的系数选择上，深层特征代表了局部信息，本身难以描述，因此本文在选择时，采用小数值的空洞系数，适度地扩大相邻像素之间的关联，而浅层网络，则是逐步增大空洞系数，扩大与远距离像素之间的关联，以获取更多的全局信息，提取更多的边缘特征[20]。同时我们也发现如果一味地增大空洞系数会出现网格效应，在边缘信息预测中尤为突出，同时网络会出现退化现象。通过实验验证逐步递增小数值的空洞系数效果更好。

传统的 U-Net 网络输入为图片形式，通过 4 层深度的网络提取足够的特征，保证图像分割的精度，而本层的所有网络输入是通过主干网络提取的。Feature map 本身就已经执行了一定深度的卷积运算，如果再执行一个深层网络，网络会出现退化，为了避免网络退化，对网络深度进行缩减[12-14]。通过实验证明，上述的优化对于人体目标的边缘细节预测起到了提升作用。

其次是不同尺度之间采用 FPN 连接方式。FPN 结构主要用于多尺度变化的目标检测。如图 3 所示。

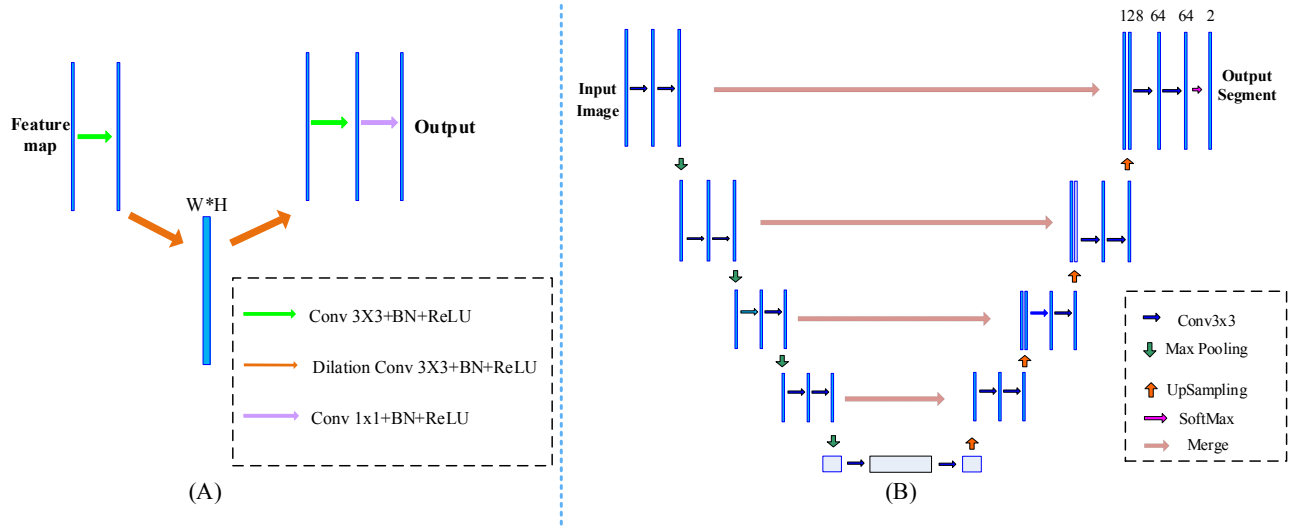

图 2 U 型编解码网络与 UNet 网络对比：(a) 本文提出的 U 型编解码码网络；(b)UN et 网络结构

Fig.2 Comparison between U-Net and U-shaped encoder-decoder networks : (a) U-shaped encoder-decoder network proposed in this article; (b) U-Net network architecture

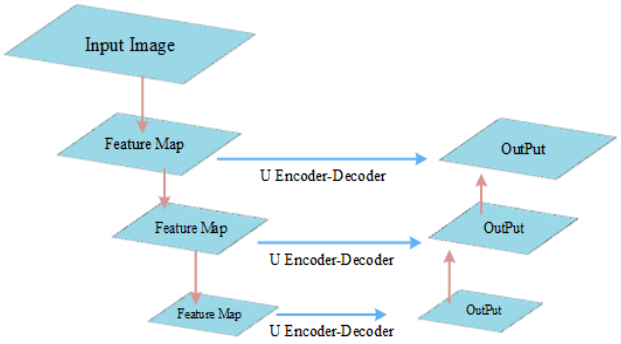

图 3 FPN 结构

Fig.3 Feature pyramid network (FPN) architecture

人体目标的尺寸变化较大，随着网络深度增加和多次下采样运算，小目标信息损失严重，无法在像素级别进行准确区分，易导致出现误检测和漏检测。FPN 结构自顶向下，深层特征流向浅层网络，可以兼顾深度特征和浅层特征，弥补了小目标信息丢失的缺陷，极大程度避免了误检测和漏检测的情况出现[24]。

1.4 融合层

融合层借鉴 HED 网络，对多路分层输出进行融合，再结合特定的损失函数预测出精确的边缘信息。

通过图 2(a)的展示，可以清晰地发现单一尺度的编解码层的输入和输出尺寸一致。而每一尺度的编解码层的输出尺寸与输入图像的尺寸相比都不相同，其原因是特征提取层中 5 个 MaxPool（2×2）的运算。因此输出融合是，必须先进行上采样（UpSample）运算将所有的输出尺寸统一到输入图像的尺寸大小。

设输入图像为 I，宽度为 W_{input}，高度为 H_{input}，通过特征提取，编解码层运算后，输出的结果分别为 O_i，$i\in[1,5]$，对应的尺寸如下：

$$\begin{cases} W_i = W_{\text{input}} / 2^i \\ H_i = H_{\text{input}} / 2^i \end{cases}, \quad i\in[1,5]$$

式中：W_i、H_i 分别为每层编解码层输出的宽和高，因此每一层输出的上采样率分别为 2^i，$i\in[1,5]$，上采样运算后得到输出张量尺寸为$[1,W_{\text{input}},H_{\text{input}}]$。

再使用 Concat 运算，所有上采样后的特征合并成一个$[5,W_{\text{input}},H_{\text{input}}]$的张量。再通过一个核为 1×1 Conv 进行降维操作，最后使用 Sigmoid 作为激活函数输出最终的检测结果。

1.5 网络损失函数

本网络的损失函数采用多层混合损失函数的加权线性和形式，如公式(1)所示：

$$L=\sum_{k=1}^{K}\alpha_k l^k \tag{1}$$

式(1)中：l^k 表示第 K 输出结果的损失，K 表示输出的个数；α_k 是每个损失函数的权重。本网络中，$\alpha_k=1/K$，$K=5$ 包括编解码层输出结果和融合后结果。

本文使用二值交叉熵作为单层的损失函数，如公式(2)所示：

$$l^k = l_{\text{bce}}^k \tag{2}$$

式(2)中：$l_{\text{bce}}{}^k$ 对应 Pixel-level，其中 $l_{\text{bce}}{}^k$ 是常用的二值交叉熵，如式(3)：

$$l_{\text{bce}} = -\sum_{(x,y)}[G(x,y)\log(S(x,y)) + (1-G(x,y))\log(1-S(x,y))] \tag{3}$$

式(3)中：$G(x,y)\in\{0,1\}$ 表示 (x,y) 位置像素是否为 Ground Truth Label（GT Label）；$S(x,y)$表示预测出(x,y)像素点为检测物的概率。

l_{bce} 通过计算每个像素的二值分类熵，区分前景和背景的概率。

因此，公式(1)可以用公式(4)表示：

$$L=\sum_{k=1}^{K}a_k l^k = \sum_{k=1}^{K}a_k(l_{\text{bce}}^k)=\sum_{k=1}^{K}\frac{(l_{\text{bce}}^k)}{K} \tag{4}$$

在公式(4)中，损失函数以线性组合的形式，对多输出结果进行调整，最大程度地保证了融合后的结果在边界上的精准性。

2 实验结果

2.1 数据集

训练集数据：目前国内外没有公开用于显著检测使用的热红外图像数据源，因此，本网络训练采用自建热红外数据集，我们选购了艾睿 T3 热红外模组和 FLIR One Pro 作为采集设备，分别在室内、室外、楼梯等场景以人作为检测目标，分别针对单人，多人，障碍物遮蔽，发热物干扰等多类构建训练数据集，数据集总计 5000 张，图像尺寸为 384×288。图 4 是本数据集的部分数据展示。采用 8:1:1 的比例设置训练，验证和评估数据。

2.2 评估指标

目前显著目标检测算法主流的评估指标有精确度（Precision），召回率（Recall），F 度量（F-measure）以及平均绝对误差（mean absolute error, MAE）[1]。Precision 反映正确分配给提取区域的所有像素的显著像素的比例，而 Recall 则反映检测到显著像素对应 GT 的比例。通常使用范围在[0,255]的阈值对显著性图像进行二值化以曲线的形式表现。F-measure 通过 Precision 和 Recall 的加权谐波计算整体性能，如公式(5)所示：

$$F_\beta=\frac{(1+\beta^2)\text{Precision}\times\text{Recall}}{\beta^2\text{Precision}+\text{Recall}} \tag{5}$$

式中：$\beta^2=0.3$。MAE 采用绝对误差的平均值，可以

更好地反映预测误差的实际情况。如公式(6)所示：

$$\mathrm{MAE}=\frac{1}{W\times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\left|\mathrm{Pred}(x,y)-\mathrm{Label}(x,y)\right| \tag{6}$$

式中：W、H 表示图像的长、宽；Pred(x,y)，Label(x,y)分别表示预测图像和标记图像。

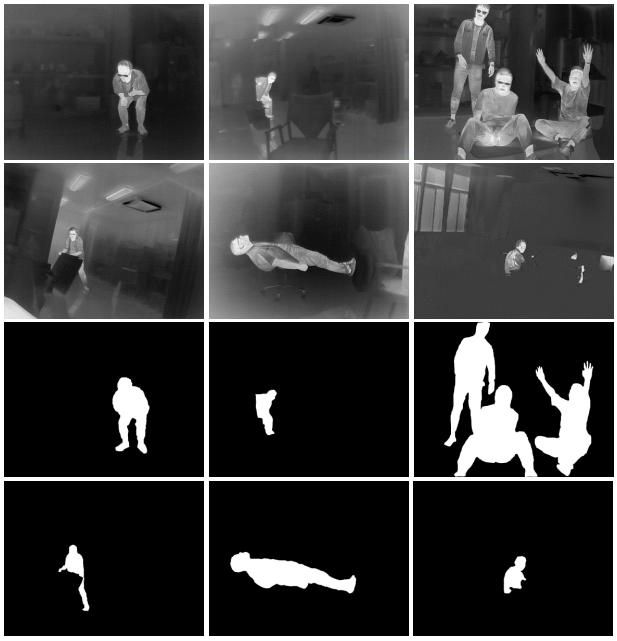

图 4 自建训练集数据：上 2 层为热红外图像，下 2 层为标记图（GT）

Fig.4 Self-built Training Dataset: The upper two layers are thermal infrared images, and the lower two layers are labeled maps (GT)

2.3 实验数据及结果分析

本实验采用自建数据集，通过采集设备获取的原始图像尺寸为 384×288。在数据增强中，分别使用了垂直/水平翻转，色彩调整（主要为对比度，色度和亮度调整）以及比例为 0.3 的灰度变换调整。通过数据增强后，训练样本个数为 4000，验证样本为 500，评价样本集为 500。网络中卷积层权重采用 Xavier 进行初始化，优化器（Optimizer）选择 Adam，学习率（learning rate）设置为 1e-3，Batch-Size 设置为 4，迭代次数为 400，整个训练用时大约 17 h。网络框架采用 Pytorch 1.4，训练平台为 Intel Core i7-7700HQ CPU @2.80 GHz，内存为 16 GB，显卡为 NVIDIA GeForce GTX 1050 Ti（4G）。

2.3.1 编解码层优化实验

实验用于验证，使用不同尺寸的空洞卷积结合普通卷积方式的编解码形式（Dilation Conv＋Conv）的优化效果，分别与普通卷积（Conv），空洞卷积（Dilation Conv），空洞尺度为 2 的空调卷积＋普通卷积（Dilation Conv＋Conv，Dilation＝2），小尺度（2, 2, 4, 6, 8）递增的空洞卷积+普通卷积方式（Dilation Conv＋Conv，Dilation＝2, 2, 4, 6, 8），大尺度（4 , 4 , 16, 32, 64）的空洞卷积＋普通卷积方式（Dilation Conv＋Conv ，Dilation＝4, 4, 16, 32, 64）以及小尺度非递增的空洞卷积＋普通卷积方式（Dilation Conv＋Conv, Dilation＝2,2,8,16,32）进行实验对比。评估指标为最大 F 度量值（maxF_β），MAE。实验结果如表 3 所示。

从表 3 可以清晰地看出，采用 Dilation Conv＋Conv（Dilation＝2），Dialtion Conv＋Conv（Dialtion＝2, 2, 4, 6, 8）在相关指标中都有较好的表现，但后者的效果更好。采用大数值空洞系数的 Dilation＋Conv（dilation＝4 ,4 ,16 , 32 ,64）指标最低，甚至低于仅采用 Conv 的情况，使用了空洞系数（dilation＝2,2,8,16,32）的情况，指标对比逐步变换的小数值空洞系数也有所下降。进一步验证了 1.3 节在编解码层的优化，能更好地兼顾全局/局部上下文特征，突出图像边界的细节检测。

表 3 编解码网络优化实验对比

Table 3 Comparative experiment of encoder-decoder network optimization

Operation	maxF_β	MAE
Conv	0.8279	0.01052
Dilation Conv (Dlation=2)	0.8526	0.00987
Dilation Conv+Conv (Dlation=2)	0.8884	0.00616
Dilation Conv+Conv (Dlation=2, 2 ,4 ,6, 8)	0.8934	0.00607
Dilation Conv+Conv (Dilation=4 , 4 , 16 , 32 , 64)	0.7891	0.01421
Dilation Conv+Conv (Dilation =2 ,2 ,8 , 16, 32)	0.8491	0.00979

2.3.2 多层特征融合对比实验

实验用于验证，多层特征融合多层输出（multi-Layered fusion multi output，MFMO）与多层特征融合单一输出（multi-layered fusion single output, MFSO）在损失函数上的对比表现，选择迭代 100 次时单一的 BCE 的损失函数值作为评估指标。实验结果如表 4 所示。

表 4 中 MFMO-layer$i$ i＝1,2,3,4,5 分别表示对应层输出的结果，其中 MFMO-layer5 表示最终输出结果，

对比 MFMO 与 MFMO-layer5 在 BCE 的值不难发现，在相同迭代次数上，后者的损失函数下降幅度更快。而在损失函数的选择上，采用混合形式的损失函数要比单一的损失函数更具优势，鲁棒性更高，对边界细节检测更为敏感。

2.3.3 网络对比实验

实验以 BASNet、PICNet、PoolNet 3 种不同类型的网络进行对比，实验结果如图 5 所示。

图 5 可以得出 5 种网络均实现了热红外图像的显著目标检测，(b)为标注图像（ground truth , GT）。从图 5 中，可以看出(c)、(e)、(f)的检测效果较为理想。其中(c)最为接近 GT 图像，但在第 4 幅图像中也出现了错分类的情况，在细节处理上，第 3，7 幅图像也尤为明显，在手指的区分，握拳的空洞上均能较为清

表 4 多层特征融合对比实验

Table 4 Multilayer feature fusion contrast experiment

Fusion operation	BCE
MFSO	0.85367
MFMO-layer1	0.92691
MFMO-layer2	0.89786
MFMO-layer3	0.87981
MFMO-layer4	0.86286
MFMO-layer5	0.84326

(a) (b) (c) (d) (e) (f) (g)

图 5 实验结果：(a) 输入图像；(b)标注图像；(c) Ours(VGG16)；(d) Ours(MobileNet-v2)；(e) PoolNet(ResNet50)；(f) PICNet(VGG16)；(g) BASNet(ResNet34)

Fig.5 Experimental results (a) input image; (b) GT image; (c) Ours(VGG16); (d) Ours(MobileNet-v2); (e) PoolNet(ResNet50), (f) PICNet (VGG16); (g) BASNet(ResNet34)

晰地检测出来。(e)、(c)检测结果较为接近，但是在第 7 幅图像中(e)出现了检测误差。同时在第 2，3 幅图发现(e)在细节检测上没有(c)清晰。(f)的大部分检测结果也较为清晰，但是细节上明显不如(c)、(e)。(d)、(g)的检测结果对比其他 3 种网络就略有不足，在多幅图像的检测中出现了较大的误差，(d)的误差尤为明显，(g)的结果稍好。同时在第 2 幅图中出现了小尺度遮挡情况，使用本算法检测出的结果接近 GT 图像，而在第 4 幅图像中，出现了发热物体的干扰，很明显图(f)、(g)检测效果较差，出现了严重干扰的情况，而本算法检测的结果虽然也有部分的干扰，但是可以接受。

通过 PR 曲线（如图 6 所示），maxF_β，MAE、模型尺寸、实时性 5 个评价指标，对模型进行评价。

表 5 的数据可以清晰得出 Ours（VGG16）网络在 maxF_β，MAE 指标上有不错的表现，而 Ours（MobileNet-V2）网络表现较差，主要原因体现在骨干网络的选择上，同等约束条件下 VGG16 网络在多尺度特征提取的表现上要优于 MobileNet-V2 网络，但 MobileNet-V2 网络的优势体现在模型尺寸和实时性上，Ours（MobileNet-V2）的网络是 5 个网络中最轻量的，实时性也是最好的。PoolNet（ResNet50），PICNet（VGG16）和 Ours（VGG16）网络在 maxF_β，MAE 两个指标上较为接近，都体现出了较强的边界检测能力，从 PR 曲线上也验证了这一点。BASNet（ResNet34）网络在 maxF_β，MAE 指标上要略低于上述 3 种网络，但是整体性能上依然很优越。

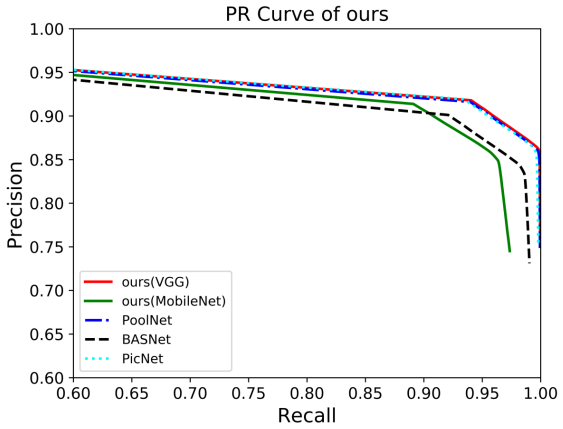

图 6 模型 PR 曲线

Fig.6 Model PR curves

表 5 多算法验证对比

Table 5 Multialgorithm validation comparison

Models	Evaluation metrics			
	maxF_β	MAE	Model size /MB	Running time/ms
BASNet(ResNet 34)	0.88087	0.01154	348.5	467.3
PICNet(VGG16)	0.88732	0.00633	153.3	178.2
PoolNet(ResNet50)	0.89066	0.00623	273.3	578.7
Ours(VGG16)	0.89146	0.00603	101.2	111.7
Ours(MobileNet-V2)	0.84066	0.01325	19.4	86.1

3 总结

针对热红外图像中人体目标边界模糊、易受发热物和遮挡物干扰，本文提出了一种基于类 HED 网络的热红外图像显著目标检测网络，该网络使用 VGG16 作为骨干网络提取多尺寸特征，同时使用多个 U 型编解码网络扩大感受野范围，最后，通过融合多层输出得到最终的检测结果。实验结果证明，本网络可以准确地预测热红外图像人体目标的边界细节。同时避免了发热物和遮挡情况下出现漏检和误检的现象。但是本网络仍存在不足处，如数据集数量过少、对比实验集过少等。后续针对不足处，进行深入研究。

参考文献：

[1] ZHAO Z Q, ZHANG P, XU S T, et al. Object detection with deep learning: a review[J]. *IEEE Transactions on Neural Networks and Learning Systems of IEEE*, 2018, **30**(11): 3212-3232.

[2] ZHANG Y, GUO L, CHENG G. Improved salient objects detection based on salient points[C]//35th *Chinese Control Conference* (CCC) *of IEEE*, 2016. DOI.:10.1109/ChiCC.2016.7554008.

[3] ZHAN Jin, HU Bo. Salient object contour detection based on boundary similar region[C]//*Fourth International Conference on Digital Home IEEE Computer Society*, 2012. DOI: 10.1109/ICDH.2012.74.

[4] Yuna Seo, Donghoon Lee, Yoo C D. Salient object detection using bipartite dictionary[C]//*IEEE International Conference on Image Processing*, 2014. DOI: 10.1109/ICIP.2014.7025228.

[5] Nouri F, Kazemi K, Danyali H. Salient object detection via global contrast graph[C]//2015 *Signal Processing and Intelligent Systems Conference* (SPIS) *Of IEEE*, 2016. DOI: 10.1109/SPIS.2015.7422332.

[6] Long J, Shelhamer E, Darrell T. Fully convolutional networks for

semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **39**(4): 640-651.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-sale image recognition[J]. *Computer Science*, 2014. DOI: 10.48550/arXiv.1409.1556.

[8] Sewak M. *Practical Convolution Neural Networks*[M]. Birmingham: Published by Packt Publishing Ltd. 2018.

[9] LIU Wei, Dragomir Anguelov, Dumitru Erhan, et al. SSD: single shot multiBox detector[C]//*IEEE European Conference on Computer Vision* (ECCV), 2016, DOI: 10.1007/978-3-319-46448-0_2.

[10] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//*IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2016, DOI: 10.1109/CVPR.2016.90.

[11] HUANG G, LIU Z, Laurens V D M, et al. Densely connected convolutional networks[C]//2017 *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, DOI: 10.1109/CVPR.2017.243.

[12] REN Qinghua, HU Renjie. Densely connected refinement network for salient object detection[C]//*International Symposium on Intelligent Signal Processing and Communication Systems* (ISPACS), 2018, DOI: 10.1109/ISPACS.2018.8923354.

[13] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C]//2015 *MICCAI*, DOI: 10.1109/ACCESS.2021.3053408.

[14] LIU N, HAN J, YANG M H. PiCANet: learning pixel-wise contextual attention for saliency detection[C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition*(CVPR) *of IEEE*, 2018: DOI: 10.48550/arXiv.1708.06433.

[15] FENG M, LU H, DING E. Attentive feedback network for boundary-aware salient object detection[C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, DOI: 10.1109/CVPR.2019.00172.

[16] QIN Xuebin, ZHANG Zichen, HUANG Chenyang. et al. BASNet: boundary-aware salient object detection[C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition*(CVPR) *of IEEE*, 2019, DOI: 10.1109/CVPR.2019.00766.

[17] LIU Jiangjiang, HOU Qibin, et al. A simple pooling-based design for real-time salient object detection[C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition of IEEE*, 2019, DOI: 10.1109/ CVPR.2019.00404..

[18] XIE S, TU Z. Holistically-nested edge detection[J]. *International Journal of Computer Vision*, 2017, **125**(5): 3-18.

[19] Mark Sandler, Andrew Howard, et al. MobileNet V2: inverted residuals and linear bottlenecks[C]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 4510-4520, DOI: 10.1109/ CVPR. 2018.00474.

[20] Andrew Howard, M Zhu, B Chen, et al. MobileNets: efficient convolution neural networks for mobile vision application[J/OL]// *Computer Science*, arXiv:1704.04861, https://arxiv.org/abs/1704. 04861.

[21] YU Fisher, Koltun V. Multi-scale context aggregation by dilated convolutions[C]//*The International Conference on Learning Representations*, 2016, DOI: 10.48550/arXiv.1511.07122.

[22] CHEN L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[C]//*Computer Science*, 2017. arXiv:1706.05587, https: //arxiv.org/abs/1706.05587.

[23] CHEN Q, XU J, Koltun V. Fast image processing with fully convolutional networks[C]//*ICCV of IEEE*, 2017, DOI: 10.1109/ ICCV. 2017.273.

[24] LIN Tsungyi, Piotr Dollar, R Girshick, et al. Feature pyramid networks for object detection[C]//2017 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) *of IEEE*, 2017, DOI: 10.1109/ CVPR.2017.106.