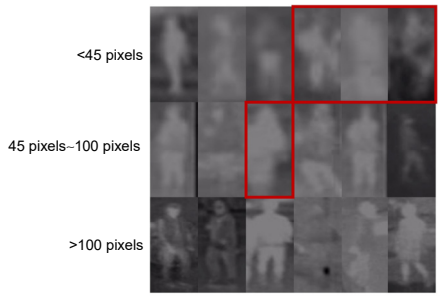
Opto-Electronic Engineering
光 电 工 程

Article

2021 年，第 48 卷，第 12 期

DOI: 10.12086/oee.2021.210358

基于多任务学习框架的红外行人检测算法



苟于涛[1,2,3]，马　梁[1,2,3]，宋怡萱[1,2,3]，
靳　雷[1,2]，雷　涛[1,2*]

[1] 中国科学院光电探测技术研究室，四川 成都 610209；
[2] 中国科学院光电技术研究所，四川 成都 610209；
[3] 中国科学院大学，北京 100049

摘要：与高质量可见光图像相比，红外图像在行人检测任务中往往存在较高的虚警率。其主要原因在于红外图像受成像分辨率及光谱特性限制，缺乏清晰的纹理特征，同时部分样本的特征质量较差，干扰网络的正常学习。本文提出基于多任务学习框架的红外行人检测算法，其在多尺度检测框架的基础上，做出以下改进：1) 引入显著性检测任务作为协同分支与目标检测网络构成多任务学习框架，以共同学习的方式侧面强化检测器对强显著区域及其边缘信息的关注。2) 通过将样本显著性强度引入分类损失函数，抑制噪声样本的学习权重。在公开 KAIST 数据集上的检测结果证实，本文的算法相较于基准算法 RetinaNet 能够降低对数平均丢失率(MR^{-2})4.43%。

关键词：红外行人检测；多任务学习；显著性检测

中图分类号：TP391.41；TN215　　**文献标志码：**A

苟于涛，马梁，宋怡萱，等. 基于多任务学习框架的红外行人检测算法[J]. 光电工程，2021，**48**(12): 210358
Gou Y T, Ma L, Song Y X, *et al*. Multi-task learning for thermal pedestrian detection[J]. *Opto-Electron Eng*, 2021, **48**(12): 210358

Multi-task learning for thermal pedestrian detection

Gou Yutao[1,2,3], Ma Liang[1,2,3], Song Yixuan[1,2,3], Jin Lei[1,2], Lei Tao[1,2*]

[1]Photoelectric Detection Technology Laboratory, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[2]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[3]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Compared with high-quality RGB images, thermal images tend to have a higher false alarm rate in pedestrian detection tasks. The main reason is that thermal images are limited by imaging resolution and spectral characteristics, lacking clear texture features, while some samples have poor feature quality, which interferes with the network training. We propose a thermal pedestrian algorithm based on a multi-task learning framework, which

收稿日期：2021-11-12；　收到修改稿日期：2021-11-30
作者简介：苟于涛(1997-)，男，硕士，主要从事基于深度学习的目标检测和多模图像融合识别的研究。
E-mail：gouyutao19@mails.ucas.ac.cn
通信作者：雷涛(1981-)，男，博士，研究员，主要从事基于传统方法及深度学习技术的图像处理与分析、复杂场景下目标检测识别与跟踪等方面的研究。E-mail：taoleiyan@ioe.ac.cn

makes the following improvements based on the multiscale detection framework. First, saliency detection tasks are introduced as an auxiliary branch with the target detection network to form a multitask learning framework, which side-step the detector's attention to illuminate salient regions and their edge information in a co-learning manner. Second, the learning weight of noisy samples is suppressed by introducing the saliency strength into the classification loss function. The detection results on the publicly available KAIST dataset confirm that our learning method can effectively reduce the log-average miss rate by 4.43% compared to the baseline, RetinaNet.

Keywords: thermal pedestrian detection; multi-task learning; saliency detection

1 引 言

目前，基于可见光图像的行人检测技术得到了飞速发展[1-2]，通过与行人重识别[3-4]等技术相结合，在安防监控、自动驾驶等领域中具有较大的应用价值。然而，受光照、烟雾、遮挡等干扰，仅依赖这类图像作为检测任务的解决方案难以在全天候复杂环境下实现较为鲁棒的检测。为此，文献[5-7]等提出基于多传感器信息融合的算法。但其数据获取难度较大，硬件成本较高。由于红外图像能够感知目标所发射的指定波段的热辐射信息，抗干扰能力强，不受环境光照的影响[8]。因此，本文基于红外图像，通过利用空间显著性信息，提升网络对红外行人的检测能力。

传统的行人检测算法主要通过滑动窗口产生大量候选区域，提取区域内手工特征，例如：HOG，SIFT等，再通过 SVM 等分类算法完成对候选区域内容的判别。但这类方法人工干扰较强，检测精度较差。随着深度学习技术的发展，R-CNN 系列[9]，Yolo 系列[10]等以不同的检测思路实现了较高精度的目标检测。面向基于可见光图像的行人检测算法，Zhang 等人[1]首先将 Faster R-CNN 在行人检测中的应用进行了相关研究。为了有效地感知不同尺度大小的行人样本，Li 等人[2]引入尺度感知模块。与上述方法相比，基于红外图像的检测算法性能距离实际应用存在较大的差距，主要存在以下几个原因：

1) 图像质量较差。由于红外物理特性以及硬件设备的限制，红外图像往往成像模糊，分辨率较低。目前大多数红外目标检测算法主要通过基于可见光图像的检测模型迁移而来，未能有效结合红外图像本身性质对检测算法进行优化。

2) 噪声样本。由于温度分布及拍摄环境的复杂性，红外图像中的部分样本并不具备良好的特征信息，如图 1(a)红框内所示。这些噪声样本因遮挡、成像距离、环境等因素产生，与背景特征较为接近，加大了网络学习的难度，容易使网络陷入较强的数据拟合而难以学习到具有普适性的红外行人特征。

针对问题 1)，John 等人[11]提出了一种自适应模糊 C-means 与卷积神经网络结合的检测模型，利用 C-means 分割算法对红外行人目标进行分割并筛选候选框。Devaguptapu 等人[12]通过 Cycle-GAN 将红外图像转化为伪彩色图像，并通过双目标检测器进行检测。同年，Ghose 等人[13]在保持原有纹理特征不变的情况下引入红外图像的显著信息，使其在不同时段的丢失

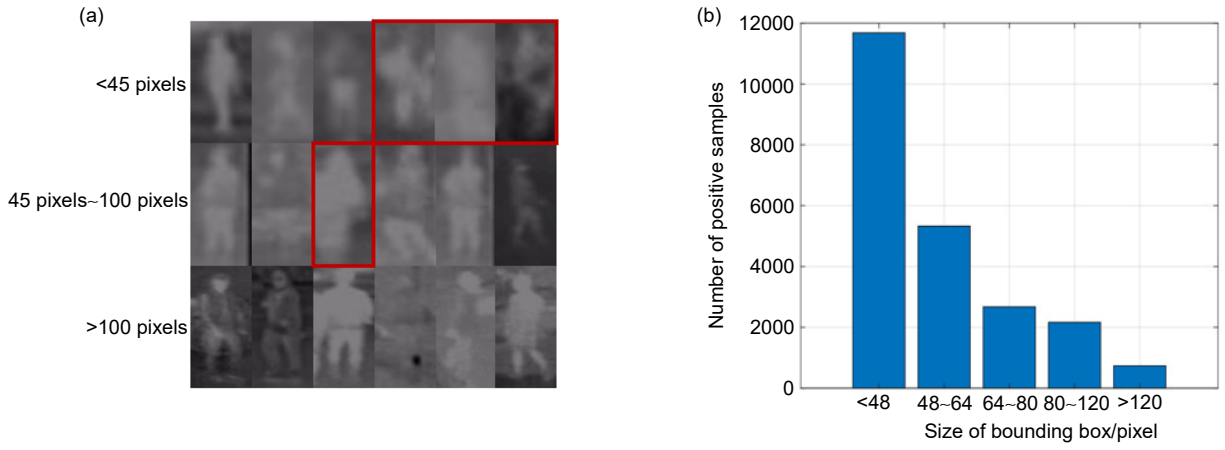


图 1 KAIST 行人样本可视化。(a) 不同尺度的部分行人样本；(b) 尺度分布情况

Fig. 1 The visualization of pedestrian samples in KAIST.
(a) Part of pedestrian samples with different scale; (b) Distribution of thermal pedestrian samples on KAIST

率均有所下降，但推理时大量的计算消耗导致其难以应用于实际场景。针对问题 2)，最新的 TC-Det[14]通过引入分类网络分支，利用场景光照信息有效弱化噪声标签的干扰。

在深度学习技术中，多任务学习方式主要是通过共享相似任务间的有效信息，提升原有任务的表现。本文从多任务学习的角度出发，对比分析独立学习式及引导注意力式两类分支结构的设计，使其具有对红外图像显著区域的判别能力，最终以共享特征提取层的方式为检测分支提供场景显著信息，提升行人检测性能。此外，根据样本显著性分析可知，这些红外图像中所存在的噪声样本和背景的差异较小，具有较弱的显著性表达。因此，将协同分支所推理出目标的显著性信息引入至分类损失函数中，能够有效弱化网络对这些样本的关注，提升网络整体的泛化性能。

本文的主要贡献包括：

1) 在目标检测网络的基础上添加显著性检测分支，使网络具备红外图像显著性检测能力的同时，能以共同学习的方式，强化检测器对显著区域的关注。

2) 将显著性检测结果转换为每个样本的显著性得分，并结合手工设计的 Smooth Focal-Loss 函数计算网络分类损失，弱化噪声样本对网络学习的干扰。

3) 本文对整个网络结构进行消融测试，并通过横向对比主流的红外检测算法，证实了本文训练方式的有效性。最终，本文实现 KAIST 数据集上的 MR^{-2} 相较于基准算法 RetinaNet[15]显著降低 4.43%，且仅作为训练方式不增加计算消耗。

2 方法原理

本文提出了一种基于多任务学习框架的红外行人检测算法。首先将红外图像通过 ResNet50 完成不同尺度特征的提取，并基于尺度特征金字塔(feature pyramid network，FPN)完成高层语义特征到浅层细节特征的信息传播。然后将各层级特征分别送入协同分支与目标检测分支，分别完成显著性检测任务与行人检测任务。其中，协同分支利用显著性检测网络 R³Net[16]预测出像素级的显著标签 p_i 进行监督学习，并通过 $\varphi_{\text{score}}(y_i^s;\alpha,S_{\text{low}})$ 将显著性结果 y_i^s 转换为各个样本的显著性得分 S_i，用于辅助检测分支的训练。在 2.1 中，介绍引入显著性检测的多任务学习框架设计；在 2.2 中，介绍基于样本显著性的分类损失函数设计；在 2.3 中，介绍本文算法的整体计算步骤。网络整体框架如图 2 所示。

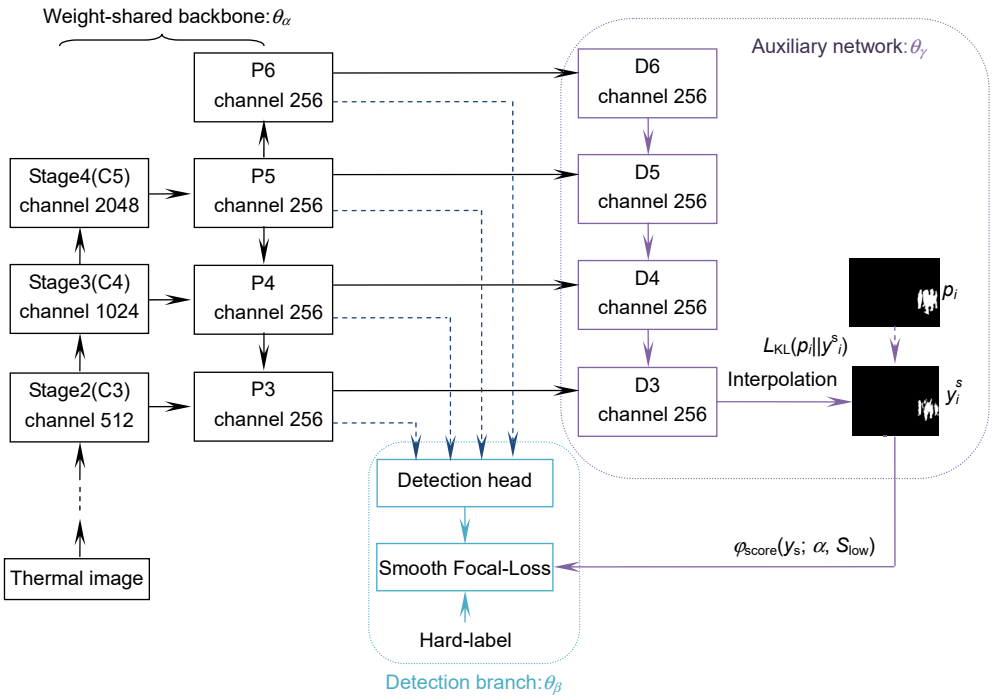


图 2 网络整体框架示意图

Fig. 2 The illustration of the network framework

2.1 引入显著性检测的多任务学习框架设计

Ulman 等人[17]将某一位置的显著性定义为该位置在颜色、方向、深度等方面与周围环境的差异程度，而图像所对应的显著图能够有效显示出该场景内的突出区域。Ghose 等人[13]首先提出将显著图通过通道替换的方式对红外图像进行加强，整体实验流程如图 3(a)所示。基于其实验结果分析可知，显著图作为一种显式的空间注意力，能够引导检测器学习显著区域。同时，相比于传统基于手工特征的显著性检测方法(如文献[18-19])，深度学习方法加强了对语义特征的关注，有效降低了大量背景噪声的干扰。该实验对训练集中的 1702 张图像以及测试集中的 362 张图像进行了像素级的显著区域标注，并通过 PICA-Net[20]和 R^3Net[16]两种深度显著性网络预测出数据集中所有的显著图并进行实验。虽然实验结果证实了利用显著图增强红外行人检测的有效性，但该方法作为一种数据增强手段，在实际应用时，需要通过额外的网络对测试图像进行显著性检测，严重影响了单帧行人检测的推理速度。

考虑到上述方法的局限性及显著图对红外目标检测的强化作用，本文设计了一种多任务学习方式，即在训练过程中同时完成目标检测及显著性检测两个任务，具体流程如图 3(b)所示。其中，协同分支在该框架中主要有两个作用：1) 学习红外图像显著区域的判别能力，以共同学习的方式替代原先的注意力强化手段，引导检测器关注显著区域；2) 显著性标签中包含显著目标精细的轮廓信息，与目标框标注相比，更有利于检测器的学习。下面本文将从协同分支结构的设计和学习方式进行分析。

2.1.1 协同分支结构设计

目前显著性检测网络大多数基于全卷积框架的设计，在采用特征提取网络进行不同层级的语义特征提取后，通过解码器框架对其进行解码，最后由像素级的标注信息进行监督学习。由于数据集中行人样本尺度差异较大，本文采用经典的单阶段多尺度目标检测算法 RetinaNet[15]作为实验的基准检测网络，特征提取部分采用 ResNet50。最终，本文设计并测试了两类不同共享层级的多任务学习框架，以判断最优共享方式的结构。

独立学习式框架。目前大多数多任务学习模型采用独立学习式的架构[21-22]，即不同分支共享特征提取模块，以独立并行的方式完成各自任务。这种架构要求共享的特征能够满足不同任务的需要，并通过分支任务信息改善主任务的训练效果。在此基础上，本文设计了以下三种模型架构，设计方案如图 4 所示。

(a) 多尺度级联

考虑将 FPN 输出的每层特征沿通道方向进行级联，再将级联后的特征图通过 1×1×1024 的卷积核进行通道压缩，该框架使显著性分支的 loss 直接作用于原检测特征，对检测分支的归纳偏置较大，但由于特征压缩卷积核的通道数过多，网络学习难度较大。

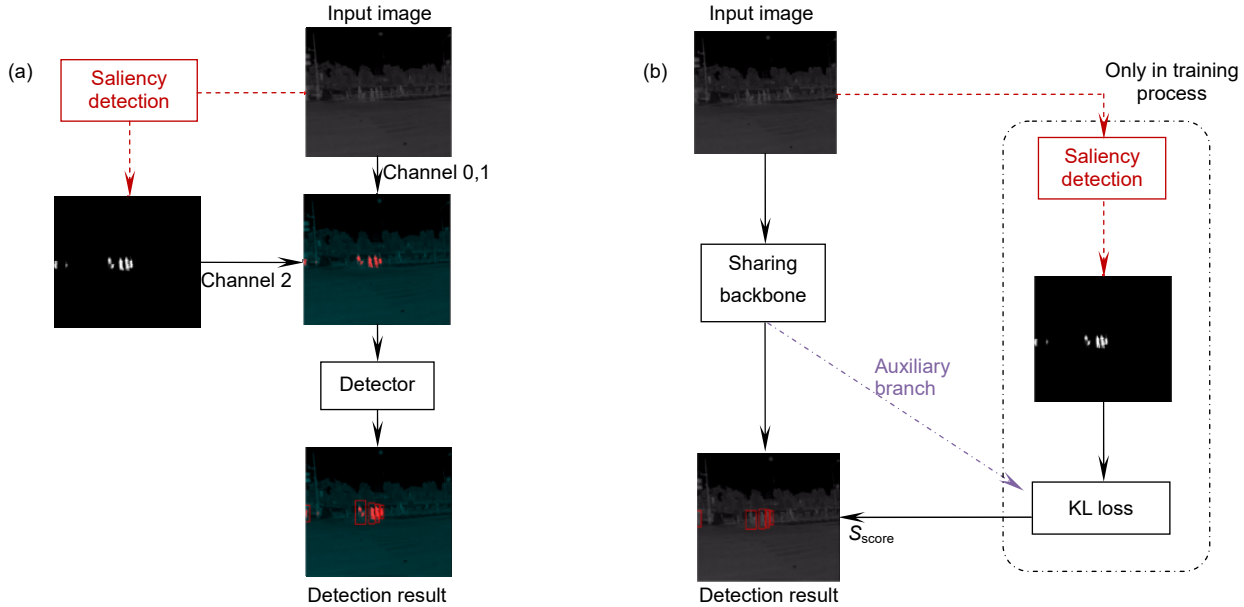


图 3 文献[13]的方法与本文方法的整体框架对比。(a) 文献[13]方法的整体检测框架；(b) 本文所提方法

Fig. 3 Comparison between the method in Ref. [13] and ours. (a) The framework of Ref. [13]; (b) Our multi-task frame work

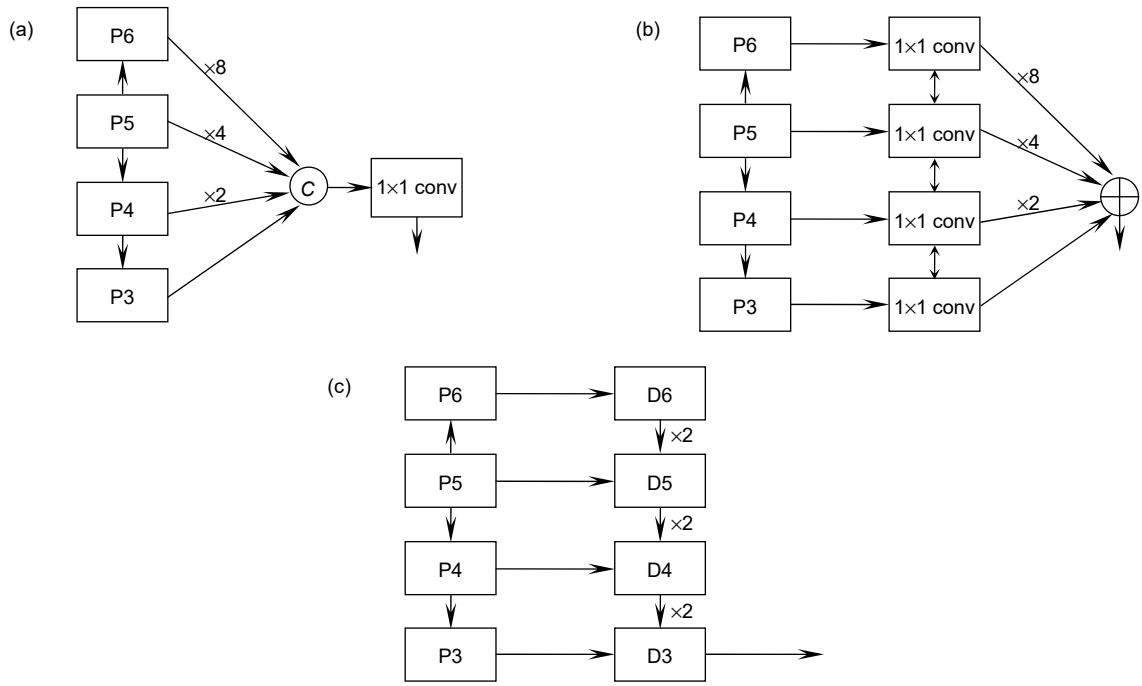


图 4 三种独立学习式网络结构设计方案。(a) 多尺度级联式; (b) 多尺度并行式(PAR); (c) 流注式(CAS)

Fig. 4 The illustration of three designed solutions.
(a) Multi-scale concatenation; (b) Multi-scale divide and conquered(PAR); (c) Multi-scale cascaded(CAS)

(b) **多尺度并行式框架(PAR)**

将不同尺度层级的特征独自进行特征压缩，卷积层参数共享，最后通过元素级相加得到最终结果。与框架(a)相比，框架(b)有效地降低了分支网络的卷积层参数，但由于每层特征最后相加时权重相同，在 loss 反传时难以考虑不同尺度目标的特征差异，因而造成精度的下降。

(c) **流注式框架(CAS)**

考虑到 Unet 框架的设计，本文将最高层语义特征 P6 通过双线性插值不断上采样，并将每次上采样后的结果 D4~D6 分别与 P3~P5 进行元素级相加及 1*1 卷积，最后 D3 特征通过卷积层降低维度，输出预测结果。相较于前两个模型，框架(c)充分利用了不同尺度层级的特征。

独立学习式框架模型在多任务学习中最为普遍，其要求特征提取模块具有容纳两种不同特征的能力，性能的提升主要通过分支网络额外的信息标注驱动主任务分支的特征提取。同时，由于两个分支完全独立，显著性分支与检测特征之间并未存在直接作用关系。

引导注意力式框架。

在引导注意力式框架中，协同分支在完成辅助任务的同时，会将网络中的特征表达作为空间或通道注意力对主任务模型中的特征进行强化。本文以级联模型为基础，将显著性分支特征或最后预测结果以元素

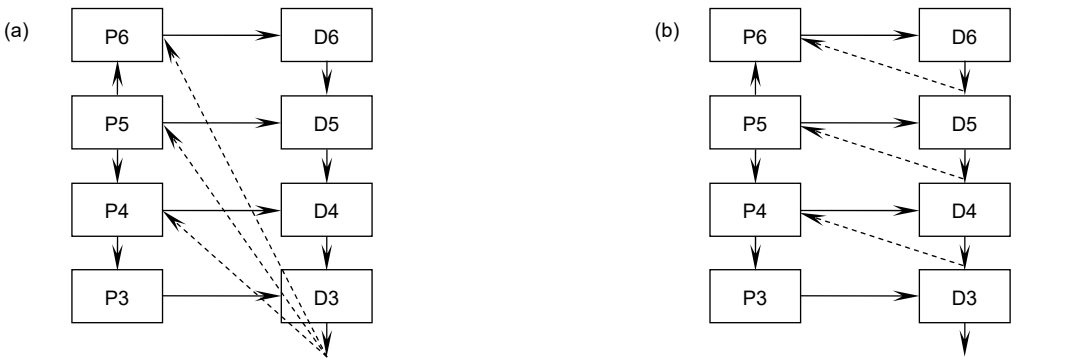


图 5 两种引导注意力式网络结构设计方案。(a) 结果强化式框架(Guided(a)); (b) 特征强化式框架(Guided(b))

Fig. 5 Two design schemes of the guided-attention network.
(a) Result enhanced framework(Guided(a)); (b) Feature enhanced framework(Guided(b))

级相加的方式作用于原有检测特征，具体模型结构如图 5 所示。

(a) **结果强化式框架(Guided(a))**

直接将协同分支的预测结果通过最大值池化后分别与(P4~P6)进行相加，为了使预测结果与原有特征的通道数相匹配，本文将预测结果在通道维度上复制 256 层。

(b) **特征强化式框架(Guided(b))**

本文将 FPN 上每层特征经过与高级特征元素级相加即等通道卷积后再作用回原特征，该方法将显著性分支中的特征整体作为注意力对(P4~P6)进行强化，其相加时两边通道数相对应。

引导注意力式框架扩展了两个分支所共享的网络，其将分支中的特征信息直接用于加强主网络特征，例如文献[14]。根据 3.2 的实验结果可知，流柱式框架(CAS)与引导注意力式框架 Guided(b)相较于原始模型均有所提升。考虑到 Guided(b)增加了推理阶段的计算消耗，最终本文采用流柱式框架作为后续优化的基础框架。

2.1.2 显著性检测标注及损失函数

本文基于文献[13]的标注，采用迁移学习的训练方法，完成协同分支的训练。网络训练框架如图 3(b)所示，首先将以 ResNext101 为特征提取结构的 R³Net[16] 显著性检测网络作为教师模型，通过已标注的部分显著性标签完成网络训练后，再对学生模型，即协同分支双线性插值后的结果进行像素级监督指导。本文通过 KL-Loss 计算两种网络检测结果分布的相似性，使协同分支的显著性检测结果与 R³Net 接近，其中 KL-Loss 的计算为

$$L_{\text{sal}} = F_{\text{KL-loss}}(p_i \,||\, y_i^s) = \sum_{i=1}^{N} p_i \cdot \log\left(\frac{p_i}{y_i^s}\right), \quad (1)$$

其中：p_i 为样本 x_i 经 R³Net 网络推理的结果，y_i^s 则为 x_i 经过显著分支得到的结果。显著性检测结果中只包含显著区域及背景区域的信息，并不能区分不同目标类别。因此计算过程中只需要对前景类与背景类进行 softmax，即类别数 N 为 2。

由于显著性分支与主任务分支相互独立，显著性标签并未直接作用于检测特征，因此该方式对显著区域的引导作用弱于直接作用于原始图像。本文采用的 R³Net 网络在 362 张测试图像中的 F_β-score 及 MAE 得分见表 1，部分测试图像如图 6 所示。由实验结果可知，R³Net 能感知图像场景中的显著区域，其大部分显著目标的边缘轮廓清晰，对协同分支的学习具有有效指导。少部分样本区域内显著性较低甚至无显著区域，证明不同目标在场景中由于特征质量的差异而具备不同强度的显著性表达。最终，将显著性检测损失 L_{sal} 引入网络主损失函数中，实现多任务框架的训练。

表 1 R³Net 显著性检测结果的定量分析

Table 1 Quantitative analyses of saliency detection results on R^3Net

Method	F_β-Score	MAE
R³Net	0.6875	0.0045

2.2 基于样本显著性的分类损失函数设计

RetinaNet 算法针对目标的分类损失采用 Focal-Loss 函数[15]，该函数将预测得分与交叉熵损失相结合，使网络更关注难分样本，忽略大量易分样本，

图 6 教师网络 R³Net 的部分显著性检测结果可视化。奇数列为红外图像，偶数列为显著性检测结果

Fig. 6 The visualization of part of R³Net saliency detection results. Odd-numbered columns are IR images, and even-numbered columns are detection results

从而缓解网络正负样本不平衡的问题。但在红外行人检测中，红外图像分辨率普遍较低，存在大量噪声样本。在 Focal-Loss 的影响下，网络过度关注这些特征空间中的离群点，而忽略了大量具有普适性特征的行人目标。这种现象严重影响了网络的泛化性能，导致网络产生大量误检结果。

本文对不同显著性强度的样本进行分析，部分样本如图 7 所示。本文发现这类特征质量较差的样本往往不具备良好的显著性表达。因此，得益于多任务学习框架，本文考虑将协同分支所预测的显著性检测结果转为显著性得分，并作为样本的先验信息引入目标检测的标签中，以合理方式降低显著性较差的样本权重，从而使网络学习到更加泛化的行人特征。本文将从样本显著性得分因子的计算和分类损失设计两个方面进行分析。

2.2.1 样本显著性得分因子计算

由于预测的显著图与原图像在空间上存在像素级的对应关系，因此本文可由式(2)计算得到样本显著性占比 C_i：

$$C_i=\frac{1}{H_i\times W_i}\sum_{j=1}^{H_i}\sum_{k=1}^{W_i}P_{\text{bbox}}(j,k)\ ,\tag{2}$$

其中：$P_{\text{bbox}}(j,k)$ 为标注框中第 j 行，k 列的显著值，可由协同分支的显著性检测结果得出，H_i 和 W_i 对应该标注框的高和宽。若直接将 C_i 当作样本权重，那么对于显著性较差的目标而言，大部分样本信息将被丢失，网络鲁棒性下降；对于显著性较强的目标，边缘的细节和目标框内背景的占比对显著性占比影响较大，微小的显著性变化都将对检测结果产生影响。因此，本文将 Sigmoid 函数通过缩放变换后作为显著性得分因子 S_i 的映射函数，弱化强显著性目标间的差异，而主要关注显著性较差的目标。其中 S_i 的计算公式为

$$S_i=\varphi_{\text{score}}(C_i;\alpha,S_{\text{low}})=\frac{w}{1+\mathrm{e}^{-\alpha C_i}}+b\ 。\tag{3}$$

为了降低超参数的计算，本文引入得分因子的限制条件。当显著性占比高于一定比例后，S_i 应趋近于 1，即分类标注接近于原始标注。而当显著性占比为 0 时，即目标框内没有显著性目标，本文设定最小阈值 S_{low}（S_{low} >0.5），保证该目标框信息仍然以 S_{low} 参与分类 loss 的计算中。因此，可得到公式(4)：

$$\begin{cases}w+b=1\\ \dfrac{w}{2}+b=S_{\text{low}}\end{cases}\Rightarrow\begin{cases}w=2(1-S_{\text{low}})\\ b=2S_{\text{low}}-1\end{cases},\tag{4}$$

其中：α 主要有两个作用：1) 目标框中始终存在一部分背景区域，需要通过 α 去除这一部分占比。2) 将显著性占比映射为 Sigmoid 函数的输入，当 α 越大，曲线越快接近于 1。不同参数下的映射函数如图 8(a)所示，其中横坐标为显著性占比，纵坐标为函数映射得到的显著性得分 S_i。当 C_i >0.4 时，显著性得分因子趋向于 1，此时相当于采用原始标注监督学习。最终，将每个正样本的 S_i 引入本文设计的 Smooth Focal-Loss 中，通过真实的目标框标注完成目标分类的监督学习。

2.2.2 Smooth Focal-Loss 函数

在红外行人检测中，为了降低噪声样本对网络学习的干扰，设计了 Smooth Focal-Loss 用于网络分类的训练。在基于 Focal-Loss 的基础上，本文针对非背景类目标，引入显著性 21 得分因子 S_i。当 S_i 较低即目标显著性较差时，网络损失会在原基础上有所下降，弱化噪声样本的训练损失，缓解网络陷入这些离群样本

图 7 协同分支的部分显著性检测结果可视化。(a) 显著性较强样本；(b) 显著性较差样本

Fig. 7 The visualization of part of saliency object detection results on the auxiliary network. (a) High salient object samples; (b) Low salient object samples

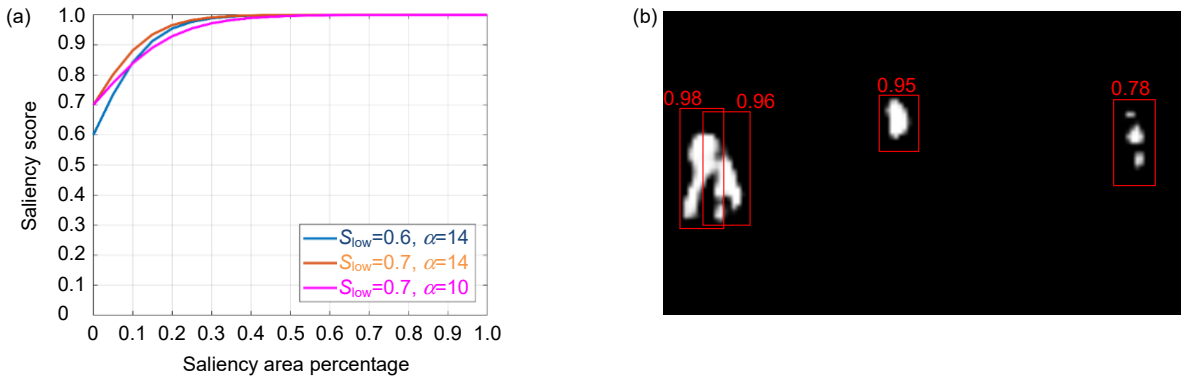


图 8 (a) 不同参数下显著性得分因子的映射函数曲线;
(b) 部分映射结果可视化。红框为检测 label，数字为计算的显著性得分因子 S_i

Fig. 8 (a) Mapping function curves of significance score factors with different parameters;
(b) Visualization of partial mapping results. The red box refers to the object detection label, and the number refers to the S_i

的判别中。同时，由于背景类未引入显著性得分，强化了网络对误检情况的判别，需通过 α_2 对正负样本比例进行控制。Smooth Focal-Loss 计算公式如下：

$$L_{\rm cls}=\begin{cases}-\alpha_2(1-\hat{y}_i)^{\gamma}S_i\log(\hat{y}_i) & y_i=1\\ -(1-\alpha_2)(\hat{y}_i)^{\gamma}\log(1-\hat{y}_i) & y_i=0\end{cases}, \quad (5)$$

其中：y_i 为分类标签，$\hat{y}_i$ 为检测分支计算结果，α_2 为正负样本平衡因子，在实验中本文将其设置为 0.25，γ 设置为 2。网络最终的损失函数计算公式为

$$L_{\rm total}=L_{\rm cls}+L_{\rm reg}\ , \quad (6)$$

其中：$L_{\rm reg}$ 为目标检测分支的回归函数，用于监督学习目标框的坐标偏移，本文采用 Smooth L1-Loss 进行计算。最终，通过显著性损失函数 $L_{\rm sal}$，回归函数 $L_{\rm reg}$ 以及分类函数 $L_{\rm cls}$ 计算得到网络整体损失函数 $L_{\rm total}$，并以此完成整体网络的训练。

2.3 算法整体计算步骤

算法计算步骤：θ_α，θ_β，θ_γ 分别为网络特征提取，协同分支，目标检测分支各部分参数。

输入：图像样本 x_i，目标框标注 y_i；

输出：网络训练整体损失 $L_{\rm total}$。

1) x_i 输入预训练后的教师网络 R³Net，并由 R³Net 给出显著性检测标签 p_i；

2) x_i 输入特征提取网络 $F(x_i;\theta_\alpha)$，得到中间特征 $\hat{x}_i$；

3) $\hat{x}_i$ 输入协同分支 $F(\hat{x}_i;\theta_\beta)$ 计算得到显著性检测结果 y_i^s；

4) 将 y_i^s 与 p_i 通过式(1)计算得到显著性检测损失 $L_{\rm sal}$；

5) 将 y_i^s 与 y_i 通过式(2)转换为每个样本的显著性占比 C_i；

6) 设置参数 $S_{\rm low}$ 与 α，由式(3)建立显著性得分映射函数 $\varphi_{\rm score}(C_i;\alpha,S_{\rm low})$；

7) C_i 输入 $\varphi_{\rm score}$ 得到样本显著性得分 S_i；

8) $\hat{x}_i$ 输入目标检测分支 $F(\hat{x}_i;\theta_\gamma)$，得到目标框检测结果 $\hat{y}_i$；

9) 将 $\hat{y}_i$，y_i 及 S_i 带入 Smooth Focal-Loss 即式(5)进行计算，得到样本分类损失 $L_{\rm cls}$；

10) 将 $\hat{y}_i$，y_i 带入 Smooth L1-Loss 计算，得到样本回归损失 $L_{\rm reg}$；

11) 将 $L_{\rm sal}$，$L_{\rm cls}$，$L_{\rm reg}$ 等权重相加，得到网络整体损失 $L_{\rm total}$。

3 实验结果

3.1 实验细节

3.1.1 实验环境

本文采用 Pytorch 框架完成所有算法的训练和测试。网络 ResNet50 部分参数在 ImageNet 数据集中完成预训练，其余参数采用 Xavier 方法进行初始化。本文采用 Adam 优化器以 0.0001 的学习率在单个 NVIDIA TITANX GPU 上训练 40 轮。学习过程中，本文将单批数量设置为 8，锚框长宽比为 0.42，并在 4 个不同尺度上分别设置[1, 1.4, 1.7]三个不同大小的框。在训练过程中，本文采用数据增强方法对样本进行随机增强，包括：随机裁剪、缩放、翻转、归一化等方法，并通过随机通道对比度、亮度等模拟红外成像所产生的噪声干扰。在测试过程中，本文采用阈值为 0.3

的非最大值抑制以去除预测过程产生的大量重复框。针对显著性检测网络 R³Net，本文采用 0.9 动量，学习率为 0.001 的 SGD 优化器，单批数量为 10 进行 9000 次迭代训练。

3.1.2 数据集

本文在 KAIST 多光谱数据集上进行实验测试。其中该数据集包含 95328 张配准的可见光—远红外图像对，并包含 1182 个独立的行人样本。本文仅采用红外部分图像用于本文的实验。本文采用与文献[14]一致的实验方案，即训练集采用文献[24]中提供的清洗后的训练标注，而测试集采用文献[6]提供的测试标注，测试集按照行人检测的合理设置[5]进行测试。其中测试图像有 2252 张图像样本组成，其包含 1455 张白天图像与 797 张夜晚图像供实验分析。为了完成显著性检测任务且保证实验的合理性，本文采用 Ghost 等人提供的 1701 张像素级标注。这些标注均从训练集中采集而不包含任何测试集信息。

3.1.3 评估指标

针对行人检测，本文借助于 KAIST 标准评估工具对行人检测结果进行评估，其中采用对数平均丢失率(log-average miss rate，MR^{-2})对检测性能进行量化。该指标计算方式为在$[10^{-2}, 10^{0}]$中的单张图片误检数(false positive per image，FPPI)按对数间隔均匀取 9 个点，并由每个点所对应的最小丢失率(miss rate，MR)的对数平均值计算所得。FPPI 和 MR 的计算式如下：

$$FPPI=\frac{N_{\mathrm{FP}}}{N_{\mathrm{img}}}\ , \tag{7}$$

$$MR=\frac{N_{\mathrm{FN}}}{N_{\mathrm{FN}}+N_{\mathrm{TP}}}\ , \tag{8}$$

其中：N_{TP}, N_{FP}, N_{FN} 分别表示检测过程中的真阳性样本数，假阳性样本数，假阴性样本数，N_{img} 表示检测图片总数。FPPI 定量平均每张图片的误检个数，而 MR 为正样本中未能够检出的比例。MR^{-2} 指标统一且有效衡量了目标误检及丢失情况，通常用于行人检测等目标密集场景的检测性能评估。MR^{-2} 越小即表示整体检测的丢失率及误检情况越少。为了更全面地分析本文算法的性能，本文还在一部分实验中，采用经典目标检测指标 mAP 用于定量检测精度及召回率的表现情况。AP 主要通过计算精确度(Precision，P)和召回率(Recall，R)曲线与坐标轴围成的面积得到。其中 Precision 和 Recall 计算公式如下：

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}\ , \tag{9}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}\ 。 \tag{10}$$

Precision 为所有预测为正的样本中，实际为正的样本比例。Recall 则为所有实际为正的样本中能够有效检出的比例。其中 AP 指标主要用于统一衡量 Precision 及 Recall 的整体情况，AP 越大表明目标检测的综合性能越强。由于本文仅针对行人单类目标进行分析，因此 mAP 与 AP 值相同。

3.2 消融实验分析

3.2.1 多任务学习框架性能测试

本文在数据集中完成了 2.1 中设计的 PAR、CAS、Guided(a)、Guided(b)这 4 种方案的性能测试，其中测试结果如表 2、表 3 及图 9。

通过对表 2、表 3 及图 9 的数据分析，本文可得到以下结论：

1) 采用流注式的多任务网络(CAS)丢失率低于基准网络，这说明引入合理的显著性检测分支结构能够从侧面强化行人检测的性能。

2) 引导式注意力模型(Guided(a))由于采用单通道复制的方法直接与原特征通道数匹配，破坏了原有的特征分布情况，MR^{-2} 上升 7.86%。而模型(Guided(b))通过将与注意力特征重新结合，强化了特征提取网络，对原有检测分支添加了近似自注意力的结构，MR^{-2} 下降 0.21%。但在推理时仍需要保留分支网络，加大了计算消耗。

3) 在独立学习式框架中，多尺度并行框架相较于

表 2 独立学习式框架性能测试

Table 2 Individual learning framework contrastive test

Algorithm	Reasonable-all		Reasonable-day		Reasonable-night	
	MR^{-2}/(%)	mAP/(%)	MR^{-2}/(%)	mAP/(%)	MR^{-2}/(%)	mAP/(%)
RetinaNet(baseline)	24.68	83.20	29.41	80.45	15.41	89.11
PAR	25.16	84.00	30.35	80.80	13.36	91.43
CAS	22.15	86.39	27.77	82.75	9.63	94.16

表 3 引导注意力式框架性能测试

Table 3 Guided-attention framework contrastive test

Algorithm	Reasonable-all		Reasonable-day		Reasonable-night	
	MR^{-2}/(%)	mAP/(%)	MR^{-2}/(%)	mAP/(%)	MR^{-2}/(%)	mAP/(%)
RetinaNet(baseline)	24.68	83.20	29.41	80.45	15.41	89.11
Guided(a)	30.01	79.96	34.68	76.86	19.40	87.23
Guided(b)	21.94	85.74	26.97	82.72	11.86	92.38

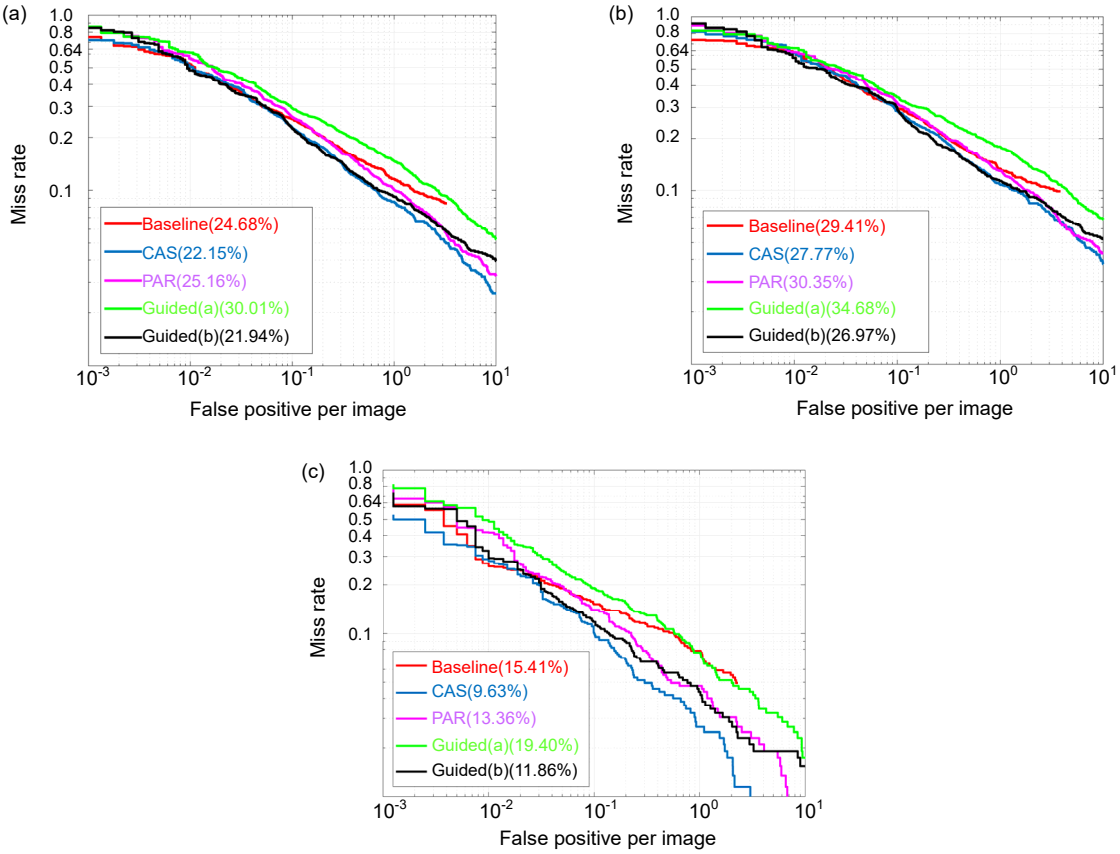


图 9 不同时段不同模型的 MR-FPPI 曲线可视化。(a) 全天候；(b)白天；(c) 夜晚

Fig. 9 The visualization of MR-FPPI curves with various models on different periods within (a) all day, (b) only day, and (c) only night

基准网络 MR^{-2} 反而上升 0.48%，其精度的损失主要来源于不同尺度目标特征分布的差异性，由于不同层级分支网络等权重的反向传播分支损失，使其难以适应这种差异性而进行等效的优化，造成了性能的下降。考虑到以上三点，本文将采用 CAS 模型作为多任务学习的基本方案，并在此基础上完成对基于显著性的损失函数性能研究。

3.2.2 基于样本显著性的分类损失函数性能研究

本节主要验证显著性得分映射函数 $\varphi_{\text{score}}(y_i^s;\alpha,S_{\text{low}})$ 以及损失函数 Smooth Focal-Loss 的有效性。针对 φ_{score} 中的参数 α 和 S_{low}，本文采用网格搜索法进行实验，在不同参数值下进行测试，网络其余设置均保持不变，实验结果如表 4 所示。

由于网格搜索法搜索参数需要进行大规模的网络学习，本文仅采用 S_{low} 间隔为 0.5，α 间隔为 2 的参数列表对一部分参数进行训练。由实验结果可知，当 S_{low} =0.7，α =14 时，检测精度在 9 次实验中性能最优，且最优效果强于未采用该训练策略的 CAS 模型。然而，由于网格搜索法的局限性，该结果仅能证明训练方法的有效性，并不能证明该参数为全局最优参数。因此，在针对特定数据集时，需根据实际情况对参数进行调整。当 S_{low} 过小时，非显著目标样本的权重抑

表 4 不同参数下的检测性能对比实验

Table 4 Contrastive testing experiment with different parameters

S_{low}	w	b	α	Reasonable-all		Reasonable-day		Reasonable-night	
				MR^{-2}/(%)	mAP/(%)	MR^{-2}/(%)	mAP/(%)	MR^{-2}/(%)	mAP/(%)
0.75	0.5	0.5	16	20.63	86.50	25.15	83.55	10.63	93.19
			14	21.88	85.57	26.66	82.80	12.67	91.53
			12	22.37	85.30	27.64	81.97	11.25	92.58
0.7	0.6	0.4	16	22.72	85.36	28.51	81.86	10.84	92.78
			14	**20.25**	**86.13**	**25.18**	**82.81**	**9.57**	**93.57**
			12	21.76	85.82	26.36	82.90	11.46	92.39
0.65	0.7	0.3	16	21.13	85.89	26.06	82.83	10.37	92.80
			14	22.35	84.58	27.04	81.62	12.52	91.30
			12	21.05	86.15	24.99	83.63	12.18	91.97

制程度将严重增大，导致网络对这些目标信息的利用率较低，造成信息大规模丢失，网络性能可能反而低于 baseline。而 α 过小时，映射函数将以更慢的速度收敛于 1。此时对于显著性较高的目标，轻微的显著性差异都将导致不同的样本权重。因此只有合理设定 S_{low} 和 α 才能在不丢失有效目标信息的同时，有效抑制噪声样本的影响。

3.3 与主流红外行人检测算法的对比分析

本文将该算法与目前主流的红外行人检测算法 Faster RCNN-T[13]，Faster RCNN+SM[13]，Bottom up[25]，TC-thermal[14]，TC-Det[14]，RetinaNet[15](baseline)，RetinaNet+SM 进行了对比，对比结果如表 5 所示。RetinaNet+SM 在 RetinaNet 基础上采用与文献[13]相同的方式对原图进行显著图的堆叠。

表中 MR^{-2}-all, MR^{-2}-day, MR^{-2}-night 分别代表全天、仅白天、仅夜晚情况下的行人检测丢失率，(CAS+Smooth FL)表示采用本文 CAS 分支模型且通过设计的 Smooth Focal-Loss 损失函数进行优化，根据表中数据可得本文的多任务学习相较于 baseline 能够有效降低 4.43%，其中白天下降 4.23%，夜晚下降 5.84%。由于本文设计的网络框架强化了检测器对显著目标的关注，误检现象大幅度减小，最终检测效果如图 10 所示。实验结果表明，采用本文多任务学习方式的检测结果优于直接对原图进行增强的方法(MR^{-2} 分别为 20.25%与 23.47%)，且在测试阶段无需通过额外的网络进行显著性图的预测。

表 5 KAIST 红外行人检测算法性能测试对比，其中+SM 表示采用文献[13]的方式引入显著图

Table 5 Contrastive experiments on various thermal pedestrian detection methods, where +SM represents introducing Saliency Map in the way of Ref. [13]

Detectors	MR^{-2}-all	MR^{-2}-day	MR^{-2}-night
Faster RCNN-T[13]	47.59	50.13	40.93
Faster RCNN+SM[13]	—	30.4	21
Bottom up[25]	35.2	40	20.5
TC-thermal[14]	28.53	36.59	11.03
TC-Det[14]	27.11	34.81	10.31
RetinaNet(baseline)	24.68	29.41	15.41
RetinaNet+SM	23.47	30.30	9.85
Ours(CAS)	22.15	27.77	9.63
Ours(CAS+Smooth FL)	**20.25**	**25.18**	**9.57**

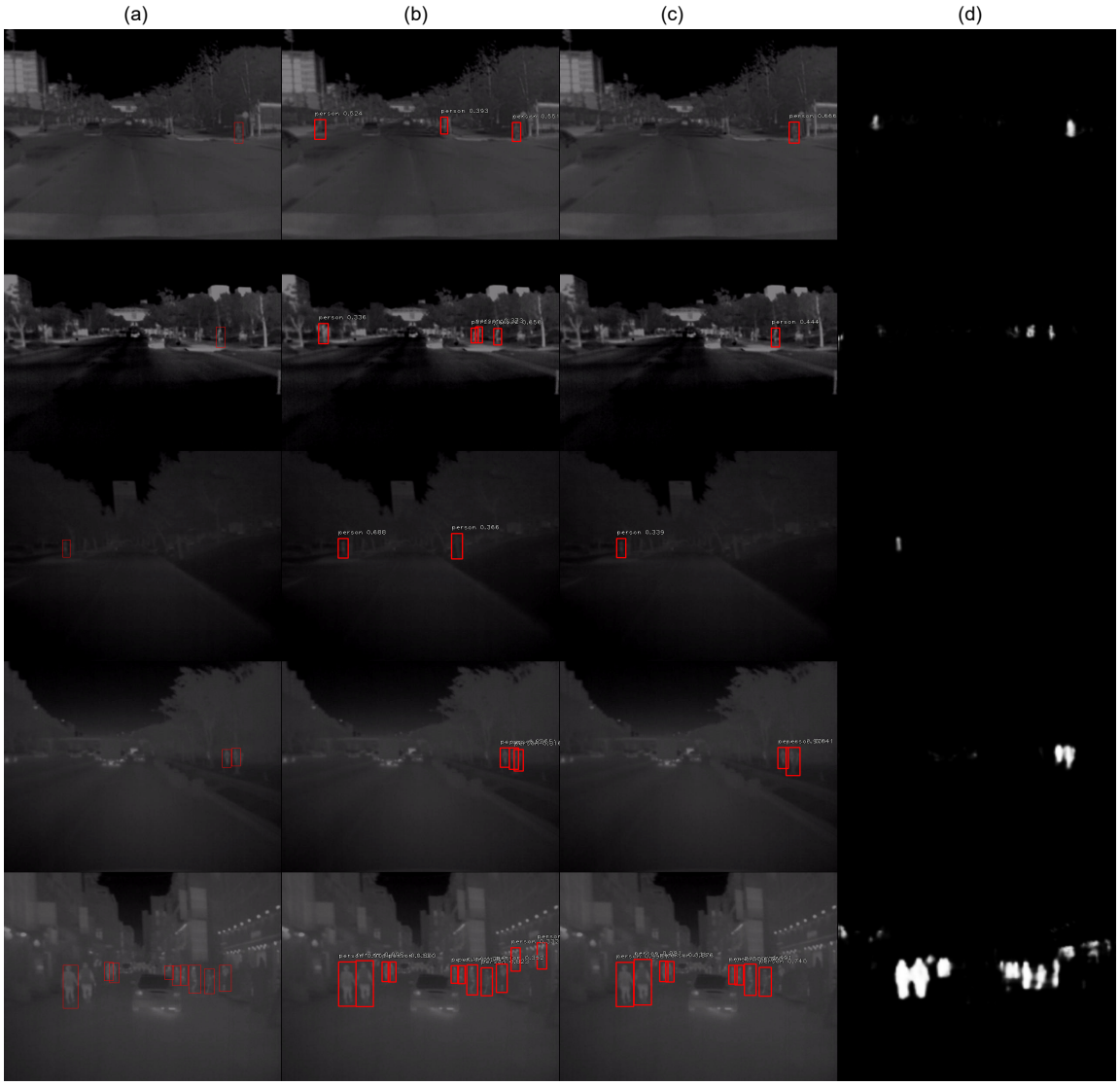


图 10　5 个场景下真实值及不同模型的检测结果。
(a) 真实值; (b) RetinaNet; (c) 本文模型检测结果; (d) 协同分支显著性检测结果

Fig. 10　Partial test results.
(a) Ground-truth; (b) Baseline; (c) Ours detection result; (d) Saliency detection result of the auxiliary network

4　结　论

本文提出了一种用于红外行人检测的多任务学习框架。针对红外图像质量较差，缺乏样本色彩及细节信息的问题，引入显著性检测任务，从侧面引导检测网络对强显著区域的关注。同时，针对红外图像中存在大量噪声样本的问题，将协同分支显著性检测的结果映射为每个样本的显著性得分因子，在分类损失中抑制噪声样本对网络学习的影响。最终，实验测试结果证实了方法的有效性，并能够在不增加推理计算消耗的同时，相较于基准算法 RetinaNet 有效降低 4.43 MR^{-2}。但是，本文方法仍受限于大量手工设计的参数。如何使网络以自适应的方式适应各种复杂场景将作为下一步研究的重点。

参考文献

[1] Zhang L L, Lin L, Liang X D, *et al*. Is faster R-CNN doing well for pedestrian detection?[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 443–457.

[2] Li J N, Liang X D, Shen S M, *et al*. Scale-aware fast R-CNN for pedestrian detection[J]. *IEEE Trans Multimed*, 2018, **20**(4): 985–996.

[3] Zhang B H, Zhu S Y, Lv X Q, *et al*. Soft multilabel learning and deep feature fusion for unsupervised person re-identification[J]. *Opto-Electron Eng*, 2020, **47**(12): 190636.
张宝华, 朱思雨, 吕晓琪, 等. 软多标签和深度特征融合的无监督行人重识别[J]. 光电工程, 2020, **47**(12): 190636.

[4] Zhang X Y, Zhang B H, Lv X Q, *et al*. The joint discriminative and generative learning for person re-identification of deep dual attention[J]. *Opto-Electron Eng*, 2021, **48**(5): 200388.
张晓艳, 张宝华, 吕晓琪, 等. 深度双重注意力的生成与判别联合学习的行人重识别[J]. 光电工程, 2021, **48**(5): 200388.

[5] Hwang S, Park J, Kim N, *et al*. Multispectral pedestrian detection: Benchmark dataset and baseline[C]//*Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015: 1037–1045.

[6] Liu J J, Zhang S T, Wang S, *et al*. Multispectral deep neural networks for pedestrian detection[Z]. arXiv preprint arXiv:1611.02644, 2016.

[7] Wang R G, Wang J, Yang J, *et al*. Feature pyramid random fusion network for visible-infrared modality person re-identification[J]. *Opto-Electron Eng*, 2020, **47**(12): 190669.
汪荣贵, 王静, 杨娟, 等. 基于红外和可见光模态的随机融合特征金子塔行人重识别[J]. 光电工程, 2020, **47**(12): 190669.

[8] Zhang R Z, Zhang J L, Qi X P, *et al*. Infrared target detection and recognition in complex scene[J]. *Opto-Electron Eng*, 2020, **47**(10): 200314.
张汝榛, 张建林, 祁小平, 等. 复杂场景下的红外目标检测[J]. 光电工程, 2020, **47**(10): 200314.

[9] Ren S, He K, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2016, **39**(6): 1137–1149.

[10] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016: 779–788.

[11] John V, Mita S, Liu Z, *et al*. Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks[C]//*2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, 2015: 246–249.

[12] Devaguptapu C, Akolekar N, Sharma M M, *et al*. Borrow from anywhere: pseudo multi-modal object detection in thermal imagery[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019: 1029–1038.

[13] Ghose D, Desai S M, Bhattacharya S, *et al*. Pedestrian detection in thermal images using saliency maps[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019: 988–997.

[14] Kieu M, Bagdanov AD, Bertini M, *et al*. Task-conditioned domain adaptation for pedestrian detection in thermal imagery[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 546–562.

[15] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection[C]//*Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017: 2999–3007.

[16] Deng Z J, Hu X W, Zhu L, *et al*. R^3Net: recurrent residual refinement network for saliency detection[C]//*Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018: 684–690.

[17] Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry[J]. *Hum Neurobiol,* 1985, **4**(4): 219–227.

[18] Hou X D, Zhang L Q. Saliency detection: a spectral residual approach[C]//*2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007: 1–8.

[19] Montabone S, Soto A. Human detection using a mobile platform and novel features derived from a visual saliency mechanism[J]. *Image Vis Comput,* 2010, **28**(3): 391–402.

[20] Liu N, Han J W, Yang M H. PiCANet: learning pixel-wise contextual attention for saliency detection[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 3089–3098.

[21] Li C Y, Song D, Tong R F, *et al*. Illumination-aware faster R-CNN for robust multispectral pedestrian detection[J]. *Pattern Recognit,* 2019, **85**: 161–171.

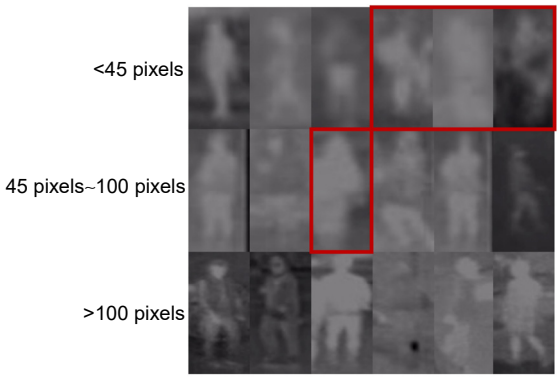
[22] Li C Y, Song D, Tong R F, *et al*. Multispectral pedestrian detection via simultaneous detection and segmentation[Z]. arXiv preprint arXiv:1808.04818, 2018.

[23] Guo T T, Huynh C P, Solh M. Domain-adaptive pedestrian detection in thermal images[C]//*2019 IEEE International Conference on Image Processing (ICIP)*, 2019: 1660–1664.

Multi-task learning for thermal pedestrian detection

Gou Yutao[1,2,3], Ma Liang[1,2,3], Song Yixuan[1,2,3], Jin Lei[1,2], Lei Tao[1,2*]

[1]Photoelectric Detection Technology Laboratory, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[2]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[3]University of Chinese Academy of Sciences, Beijing 100049, China



The visualization of pedestrian samples in KAIST

Overview: In recent years, pedestrian detection techniques based on visible images have been developed rapidly. However, interference from light, smoke, and occlusion makes it difficult to achieve robust detection around the clock by relying on these images alone. Thermal images, on the other hand, can sense the thermal radiation information in the specified wavelength band emitted by the target, which are highly resistant to interference, ambient lighting, etc, and widely used in security and transportation. At present, the detection performance of thermal images still needs to be improved, which suffers from the poor image quality of thermal images and the interference of some noisy samples to network learning.

In order to improve the performance of the thermal pedestrian detection algorithm, we firstly introduce a saliency detection map as supervised information and adopt a framework of multi-task learning, where the main network completes the pedestrian detection task and the auxiliary network satisfies the saliency detection task. By sharing the feature extraction modules of both tasks, the network has saliency detection capability while guiding the network to focus on salient regions. To search for the most reasonable framework of the auxiliary network, we test four different kinds of design from the independent-learning to the guided-attentive model. Secondly, through the visualization of the pedestrian samples, we induce noisy samples that have lower saliency expressions in the thermal images and introduce the saliency strengths of different samples into the classification loss function by hand-designing the mapping function to relieve the interference of noisy samples on the network learning. To achieve this goal, we adopt a sigmoid function with reasonable transformation as our mapping function, which maps the saliency area percentage to the saliency score. Finally, we introduce the saliency score to the Focal Loss and design the Smooth Focal Loss, which can decrease the loss of low-saliency samples with reasonable settings.

Extensive experiments on KAIST thermal images have proved the conclusions as follows. First, compared with other auxiliary frameworks, our cascaded model achieves impressive performance with independent design. Besides, compared with the RetinaNet, we decrease the log-average miss rate by 4.43%, which achieves competitive results among popular thermal pedestrian detection methods. Finally, our method has no impact on the computational cost in the inference process as a network training strategy. Although the effectiveness of our method has been proven, one still needs to set the super-parameters manually. In the future, how to enable the network to adapt to various detection conditions will be our next research point.

Gou Y T, Ma L, Song Y X, *et al.* Multi-task learning for thermal pedestrian detection[J]. *Opto-Electron Eng*, 2021, **48**(12): 210358; DOI: 10.12086/oee.2021.210358

* E-mail: taoleiyan@ioe.ac.cn