基于深度注意力机制的多尺度红外行人检测

赵斌, 王春平*, 付强, 陈一超
陆军工程大学石家庄校区电子与光学工程系, 河北 石家庄 050003

摘要 针对多尺度目标检测问题,提出一种基于深度注意力机制的多尺度红外行人检测方法。首先,选取较为轻量级的 Darknet53 作为深度卷积特征提取的主干网络,设计四尺度的特征金字塔网络负责目标的定位和分类,通过引入更低层高分辨率的特征图来改善对小尺度行人目标的检测性能。其次,利用注意力模块替代特征金字塔网络中传统的上采样模块,生成基于卷积特征的局部显著图,可以有效抑制不相关区域的特征响应,突出图像局部特性。最后,利用 Caltech 行人数据集和 U-FOV 红外行人数据集进行两次迁移训练,以提高模型的泛化能力,丰富行人的样本特征。实验结果表明,所提方法在 U-FOV 数据集上的识别平均准确率达到了 93.45%,比 YOLOv3 高 26.74 个百分点,能检测到的最小行人像素为 6×13。在 LTIR 数据集上的定性实验结果验证,所提模型具有良好的泛化能力,适用于多尺度红外行人的检测。

关键词 探测器; 红外行人检测; 卷积神经网络; 超大视场; 特征金字塔网络; 注意力机制
中图分类号 TN215 **文献标志码** A **doi**: 10.3788/AOS202040.0504001

Multi-Scale Infrared Pedestrian Detection Based on Deep Attention Mechanism

Zhao Bin, Wang Chunping*, Fu Qiang, Chen Yichao
Department of Electronic and Optical Engineering, Shijiazhuang Campus of Army Engineering University, Shijiazhuang, Hebei 050003, *China*

Abstract In this paper, for multi-scale target detection, a multi-scale infrared pedestrian detection method based on deep attention mechanism is proposed. The lightweight Darknet53 is adopted as the backbone network for deep convolutional features extracting, and a four-scale feature pyramid network is constructed to classify and localize objects. The detection performance with respect to small-scale objects such as pedestrians is improved by introducing low-level and high-resolution feature maps. Furthermore, an attention module is designed to replace the traditional upsampling block in the feature pyramid network, which generate local saliency map based on convolution feature, thus suppress the feature responses of unrelated areas and highlight the local feature of the image. Finally, the Caltech pedestrian and U-FOV infrared pedestrian datasets are used to execute two-step transfer learning to ensure the generalization of the proposed model and improve the pedestrian features. The results show that the average precision of the proposed method is 93.45% on the U-FOV dataset, which is 26.74 percentage higher than that obtained using YOLOv3, and the minimum pixel size of the pedestrian that can be detected is 6×13. In addition, the qualitative experiment results obtained using the LTIR dataset validate the good generalization of the proposed model, which makes it suitable for multi-scale infrared pedestrian detection.

Key words detectors; infrared pedestrian detection; convolutional neural network; ultrawide field of view; feature pyramid network; attention mechanism
OCIS codes 040.1880; 040.3060; 100.4996

1 引 言

红外探测系统隐蔽性好、抗干扰能力强、受光线和恶劣天气影响小,具备全天时工作的能力,是目标检测跟踪、导航制导、安防监控、汽车夜视等军事和民用探测系统中的重要组成部分。其中,超大视场(U-FOV)红外相机具有成像视场大、覆盖面广的优点,能极大地改善夜间汽车驾驶视线受阻等问题,可以有效预防近距离侧方盲区由于行人突然闯入造成的事故。

收稿日期: 2019-09-23; **修回日期**: 2019-11-01; **录用日期**: 2019-11-27

* **E-mail**: wang_c_p@163.com

为了提取红外图像或视频中的目标信息，区分前景和背景的差异，针对单帧红外图像缺乏色彩和纹理信息的问题，Liu 等[1]和 Cai 等[2]利用局部显著性差异实现了目标检测。在处理序列图像时，背景的低秩性和目标的稀疏性是背景建模的依据，如 R2PCP[3]、ROSL[4]、RMAMR[5]、PG-RMC[6]、OSTD[7]等算法通过优化分解得到了背景模型。但是，运动缓慢或长时间静止的目标会被逐渐融入背景，从而无法处理相机载体运动的情况，而且这些方法的性能依赖于人工设计特征的鲁棒性和完备性，检测结果大都缺乏语义信息，不能自动获取目标的类别属性信息。

得益于深度卷积神经网络(CNN)的飞速发展，基于深度学习的目标检测技术在性能上取得了巨大突破，其中 Faster R-CNN[8-10]、YOLO[11-13]、SSD[14-15]等方法已逐渐成为当前目标检测的主流方法。不同于传统检测方法，深度学习的主干卷积网络能从大量数据中自动学习目标特征，更有利于挖掘目标在数据中隐含的统计规律和本质特征。然而在 U-FOV 红外相机捕获的图像中，由于相机的焦距较短，目标尺度随距离增加而快速变小，且受限于红外图像对比度低、成像模糊的缺陷，小尺度目标容易淹没在背景中。此外，深度 CNN 中存在较多的下采样操作，导致小尺度目标的有限特征被进一步压缩。因此，多尺度红外行人检测的难点在于如何利用有限特征准确地检测出小尺度目标。

注意力机制可以有效学习输入数据或特征图上不同部分的权重分布，减少背景信息带来的影响，提高模型的识别能力和鲁棒性。残差注意力网络[16]利用残差机制构造网络，在引入注意力结构的同时保证了网络的深度。SENet[17]通过学习得到特征通道之间的相互依赖关系及各通道的重要性，然后依据通道重要性增强有用特征、抑制无用特征。CBAM[18]利用特征图的通道信息和空间信息，设计了一种具有注意力能力的卷积模块，该模块能使模型聚焦在更有用的信息上，进一步增强了模型对图像的分类能力。Attention U-Net[19]提出一种用于医学成像的注意力门控模型，该模型在 U-Net 中集成软注意力模型，突出显著特征。这些模型将注意力结构集成在基础网络结构中，一定程度上增加了网络参数量和计算开销。

本文将深度学习目标检测方法运用于 U-FOV 红外行人检测任务中，提出一种基于注意力机制的多尺度红外行人检测(MS-IRPD)方法。MS-IRPD 方法解决了两个问题：针对深度学习方法中普遍存在的小尺度目标检测性能不高的问题，建立低层特征金字塔检测网络(LFPN)[20]，充分利用低层高分辨率特征图的细节信息来弥补普通卷积网络中小尺度目标特征不足的缺陷；针对 U-FOV 图像中小尺度目标特征缺失严重的问题，将注意力模块融入到特征金字塔中，用于学习不同尺度卷积特征图之间的内在联系，从而产生目标显著性特征，使得检测网络更加关注局部细节，这在一定程度上弥补了红外小尺度目标特征缺失的问题。同时，不同于以往将注意力模块增加在每个卷积层上的方式，MS-IRPD 方法仅在特定层间构建注意力模型，是一种轻量的连接方法。

2　U-FOV 红外图像与深度卷积网络特性

2.1　U-FOV 红外图像特点

夜间行人检测在先进驾驶员辅助系统(ADAS)中占据越来越重要的地位，它能自动为夜间行驶的车辆提供行人目标位置信息，有效防止事故发生。相比于传统汽车探照灯的模式，红外相机在夜间不受光照影响，具有更宽广的视野并且可以探测到更显著的目标特征。相比于小视场红外相机，U-FOV 红外相机对目标距离更敏感，视野盲区小，容易实现水平视场的全覆盖。然而，由于红外图像本身质量相对较低，加之 U-FOV 成像图像中包含丰富的小尺度目标，故在 U-FOV 红外图像中实现多尺度行人检测非常困难，目前，还未有相关的文献提出检测方法，也未有对应的数据集可以直接使用。因此自主采集图像制作 U-FOV 红外行人数据集，图 1 为水平视场约为 140°，垂直视场约为 110°，焦距为 5.56 mm，成像波段为 7～14 μm 的红外镜头采集到的分辨率为 800×600 的 U-FOV 红外图像。

图 1(a)展示了距镜头 2～8 m 范围内的行人目标，此时目标占据的像素较多，具有较为明显的结构和纹理特征，但靠近边缘时畸变严重。图 1(b)展示了距镜头超过 10 m 的行人目标，这些目标所占像素少，以边缘、轮廓及亮度等底层特征为主，很难在卷积网络中形成高层语义特征。因此，U-FOV 红外相机对目标距离的敏感性高，目标尺度随距离的急剧减小，呈现出多尺度特性，其中小尺度目标很难利用现有的模型进行识别检测。虽然尺度剧烈变化给行人检测带来了极大的困难，但好的一方面在于可以利用检测结果的尺寸信息粗略地推断出传感器与目标之间的距离，无需附加设备就能为系统提供目标在图像中的深度信息。

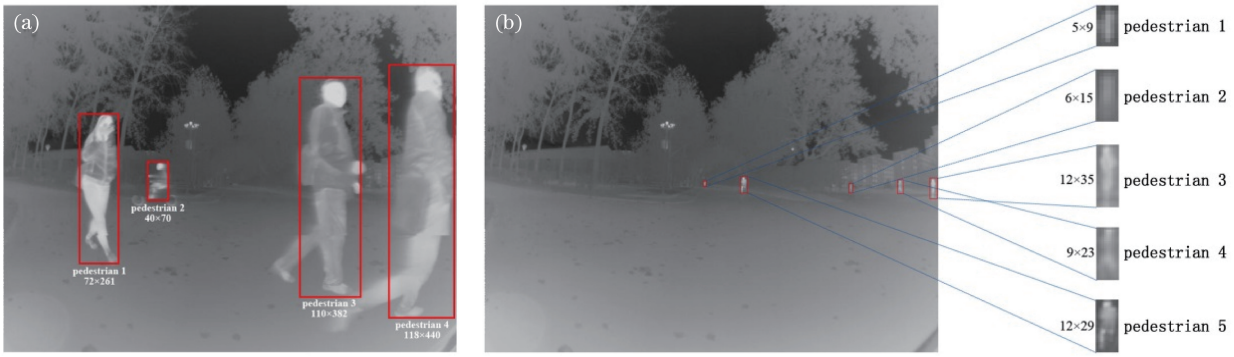

图 1　U-FOV 红外图像行人特性。(a)中大尺度目标;(b)小尺度目标

Fig. 1　Characteristic of pedestrian in U-FOV infrared images. (a) Large and medium scale pedestrians; (b) small scale pedestrians

2.2　深度卷积网络检测目标的特点

传统机器视觉的目标检测方法一般使用滑动窗口的结构,主要包括 3 个步骤:首先利用不同尺度的滑动窗口遍历图像以确定候选区域[21-22];然后提取候选区域内的视觉特征,例如 SIFT(Scale Invariant Feature Transform)[23]、HOG(Histogram of Oriented Gradient)特征[24]、Harr 特征[25]等;最后利用分类器进行分类识别,常见的分类器有随机森林、SVM、Adaboost 等。这类方法对于候选区域的选择缺乏针对性,会导致窗口冗余、时间复杂度高。并且人工设计的特征没有很好的鲁棒性,不能应对图像和目标多样性的变化。

深度卷积网络能从数据中自动学习目标特征[26],避免了显式的特征提取过程,可实现由端到端的目标检测。在网络中,依靠下采样增加感受野和抽象图像中的语义信息。随着下采样层和卷积层的堆叠,局部信息被不断综合成全局信息,边缘、纹理等低层特征被不断整合成高层语义特征,特征图分辨率变小、通道数增加,整个网络结构呈现出“倒尖锥”的形状。然而这种结构对小尺度目标的检测识别十分不友好,因为随着下采样的进行,小尺度目标映射到特征图上的像素会逐渐变少,最终不足一个像素,所以网络对于小尺度目标的定位能力急剧下降,这也是深度卷积网络中普遍存在的一个问题。例如 VGG19[27]、ResNet101[28]、GoogleNet[29]等网络都进行了 32 倍下采样,那么在最终特征图上 0.5 的量化误差反映到输入图像上则是 16 pixel,这有可能超过了某些小目标的大小,从而导致目标漏检,例如图 1(b)中 pedestrian 1 和 pedestrian 2 的像素都小于 16×16。

3　多尺度红外行人检测的深度学习框架与方法

深度卷积网络中,多尺度的目标检测是一个比较有挑战性的任务,其难点主要源于小尺度目标的低分辨率和有限的特征信息。SSD 从不同尺度的特征图上预测了目标分类与位置,提高了小尺度目标的检测性能。DSSD 使用反卷积层增加了大量的上下文信息,进一步提高了小物体的检测精度。FPN[30]利用深度卷积网络堆叠过程中固有的多尺度、多层级金字塔结构来构建特征金字塔,设计了一种具有横向连接的自顶向下架构,用于在所有尺度上构建高级语义特征图。为此,采用基于多尺度特征复用的方法构建 U-FOV 红外行人检测框架。

3.1　U-FOV 红外行人检测网络结构

深度卷积网络检测 U-FOV 中的行人需要克服两个问题:一是抵消特征图下采样对小尺度目标的不利影响;二是补充 U-FOV 红外图像中行人尺度过小造成的特征缺失。在构建检测网络时,设计两部分的网络结构:基于 YOLOv3 重构特征金字塔结构,增加一层更高分辨率的特征图来增加小目标的特征信息,可以分别从四个尺度上独立检测目标;提出一种轻量的注意力结构,仅在用于目标预测的相邻尺度特征图之间构建注意力模型,可以得到显著性特征,并将其与深度卷积特征相融合,获得有益于任务的特定局部特征。

整个网络的结构如图 2 所示,主要由特征提取主干网络、注意力生成模块及目标检测网络 3 部分构成。特征提取主干网络采用 Darknet53 完成对图像卷积特征的提取,形成多层级的特征表达。其中,

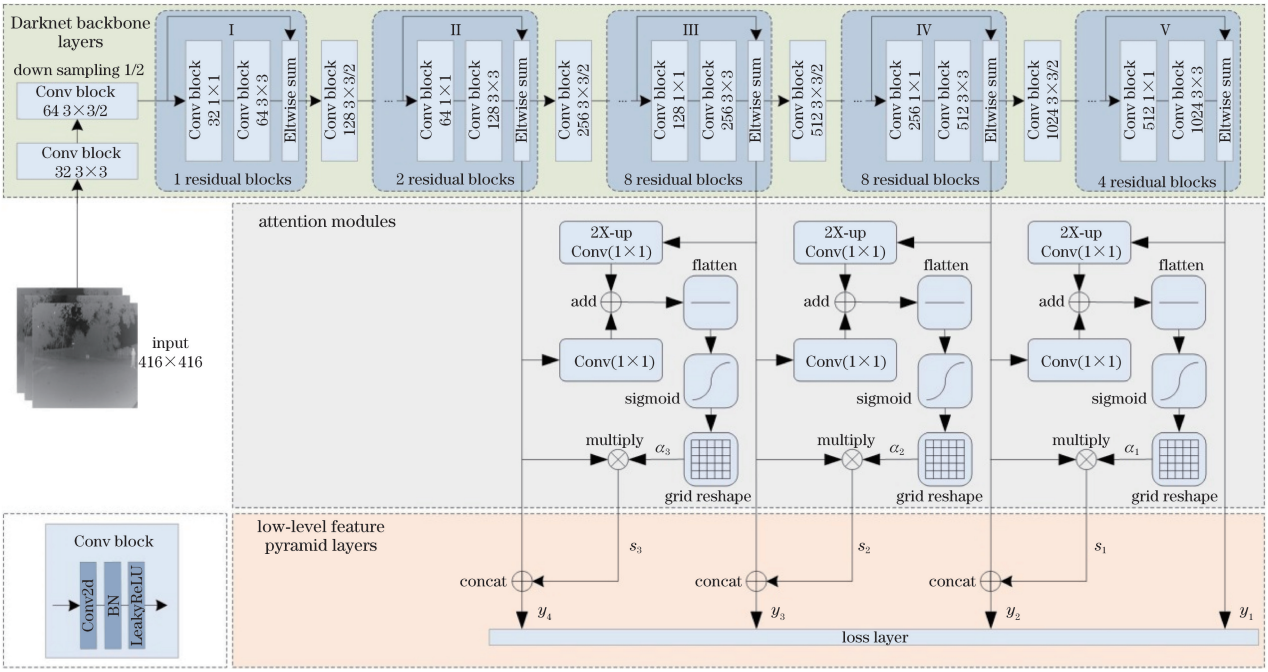

图 2　基于 Darknet53 的多尺度红外行人检测网络结构

Fig. 2　Architecture of multi-scale infrared pedestrian detection network based on Darknet53

α_1～α_3 为三个尺度上的注意力系数，s_1～s_3 为对应的注意力特征图，y_1～y_4 为用于最终目标预测的四尺度特征图，Conv 为卷积，2X-up 为 2 倍上采样，BN 为批归一化。

3.1.1　四尺度目标检测网络

原始的 YOLOv3 网络虽然改善了小目标的检测性能，但仍不足以处理 U-FOV 中的小尺度目标。该模型要求输入图像的分辨率为 416×416，用于预测目标的特征图的最大分辨率为 52×52，这之间存在步长为 8 的下采样，那么 YOLOv3 模型理论上能检测到的最小目标的分辨率在 8×8 左右。然而由图 1(b)可知，即使在分辨率为 800×600 的原始图像中也存在小于该理论尺度的目标，因此需要进一步增大可用于预测目标的特征图分辨率。基于这一考虑，重新设计目标检测网络，增加一组更低层、高分辨率的特征图用于预测目标，并将其融入到特征金字塔结构中，形成四尺度的目标预测网络，进一步提高小尺度目标的检测精度。训练时，将 y_1～y_4 的预测结果送入到损失层中计算损失，用于调整网络参数；检测时，直接在 y_1～y_4 上预测目标，得到四个尺度上的检测结果，将结果综合后输出。

3.1.2　注意力模型

相比于可见光图像，红外图像的对比度低，细节分辨能力较差，导致图像中的目标特征检测受限，尤其影响小目标的检测精度。由于红外系统依赖热辐射成像，目标在图像中基本呈现高亮的显著特性，因此可以引入注意力机制对这一特性进行强化。考虑到在增加预测特征图分辨率后，模型的计算负担有所增加，在构建显著性特征生成模型时，设计一种轻量的连接方法。不同于 CBAM 模型[18]将注意力模块固化到每个卷积层中的方式，本文利用卷积网络不同层级间固有的上下文结构，仅在两组不同尺度的特征图之间融入注意力模块，如此，针对四尺度的目标检测网络，仅需三个注意力模块。这种连接方式不仅可以生成图像显著性特征，还完成了特征金子塔结构的横向连接，实现了不同尺度特征图的融合。

注意力模块在两组不同尺度特征图间构建上下文联系，其结构如图 3 所示。两路分别输入 x_1 和 x_2，首先经过 1×1 卷积调整成相同通道数，并对低分辨率的特征图 x_1 进行 2 倍上采样，将其转换成与 x_2 通道数相同、分辨率相同的粗略特征图；然后将两路特征图按元素位相加的方式融合，经激活函数激活后输出，flatten 层将多维输入转换成 sigmoid 模块需要的输入形式，由此得到各特征点的显著性系数；最后将介于 0～1 之间的显著性系数 α 重新网格化为与输入 x_2 分辨率相同的系数图，并将之与 x_2 相乘，生成具有特定局部区域显著特征的特征图。

注意力模块计算了两组不同尺度特征图之间的目标相似性，当相似性较高时，即前后不同尺度特征图之间目标继承性较好，对应区域的显著性系数较

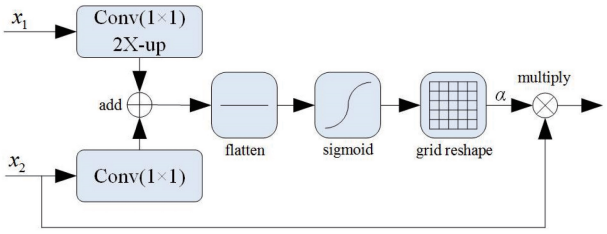

图 3　注意力模块结构

Fig. 3　Architecture of attention module

大，反之显著性系数较小。显著系数图和特征图相乘的过程可视为图像各成分权重重新分配的过程，通过突出重点区域、融入显著特征，可以有效弥补红外图像特征缺失的问题。

3.2　模型中的基本模块

3.2.1　卷积模块

卷积模块由二维卷积层、批归一化层及非线性激活层组成，图 2 Conv block 内展示了完整的卷积模块所包含的网络层。卷积层用于提取特征，该过程中最重要的是卷积核大小、步长及数量的选择。卷积核的大小影响网络结构的识别能力；步长则决定了卷积后特征图的大小；数量关乎特征提取的丰富程度，但数量越多，网络的复杂度也会随之增加。

批归一化层[31]用于将数据归一化至均值为 0、方差为 1 的数据，然后再输入到下一层。在训练深度网络时，网络参数必然会发生变化，如果不进行归一化处理，那么除了输入层外，网络后面每一层的输入数据分布都会一直发生变化。神经网络的本质就是学习数据的分布特性，一旦每批训练数据的分布各不相同，网络在每次迭代中就要去学习适应不同的分布，这会大大降低网络的训练速度，这也是需要归一化预处理数据的原因。

非线性激活函数用于增强网络的非线性描述能力，建立输入与输出之间复杂的非线性映射关系。模型中采用的激活函数（LeakyReLU）是修正线性单元（ReLU）的一种特殊版本，解决了 ReLU 在输入为负值时，输出始终为 0 所导致的神经元不学习的问题。其数学表达式为

$$y=\max(0,x)+\lambda\ \min(0,x),\tag{1}$$

式中：x 为卷积结果；y 为激活输出；λ 为一个很小的常数，保留了负轴的值。

3.2.2　残差模块

残差模块由 2 个卷积核大小分别为 1×1 和 3×3的卷积模块及 1 条捷径连接构成。如图 4 所示，1×1 卷积层对输入进行降维，减少了参数量和计算量；3×3 卷积层用于提取特征，恢复特征维度，其输出为 $H(x)$；捷径连接用于构建残差，对冗余网络层进行恒等映射，解决了随着网络深度加深时可能出现的梯度消失、梯度爆炸及网络退化等问题。

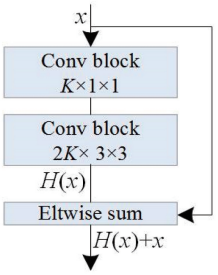

图 4　残差模块

Fig. 4　Residual module

3.3　行人检测原理

为了检测出红外图像中多尺度的行人，MS-IRPD 方法从四个尺度的特征图（即图 2 中 y_1～y_4）上生成锚框，每个尺度独立进行特征抽取、边框分类及回归，从而预测出目标的坐标、置信度及分类标签。图 5 展示了在分辨率为 13×13 特征图上实现行人检测的基本原理，其中，(t_x，t_y，t_w，t_h)分别表示边框的中心点坐标及宽高，$\sigma(t_0)$表示边框内容属于真实目标的概率分数，C_{person}和 C_{bg}表示边框内容分别分类成行人和背景的概率分数。将输入图像宽高都 13 等分，那么每个图像块对应特征图上的一个像素点，通过在特征图上产生锚框来生成目标候选边框。每个像素点位置上产生 3 种锚框，4 个尺度共 12 种锚框，锚框的宽高事先利用 k-means 聚类方法在数据集中聚类得到。根据锚框内图像特征计算边框回归参数和目标分类置信度，从而得到目标预测边框坐标和属性，再将预测结果映射到输入图像中，则可达到识别并定位目标的目的。

4　实验分析

4.1　实验设置与数据集

实验环境为 64 位 Windows 操作系统，NVIDIA GeForce GTX TITAN X GPU；软件采用 Keras，并以 Tensorflow 为后端进行卷积神经网络计算；编程语言为 Python3.6。以 YOLOv3 模型为基本框架，在经 ImageNet 和 COCO 数据集预训练的 Darknet53 上构建行人检测模型，使用 Adam 优化算法进行训练。经过两次迁移训练：第一次迁移训练在 Caltech 行人检测数据集上进行，旨在扩充数据量，增强模型对行人类目标检测的鲁棒性与泛化能力；第二次迁移训练在 U-FOV 红外行人数据集上进行，训练模型对红外图像目标的识别定位能

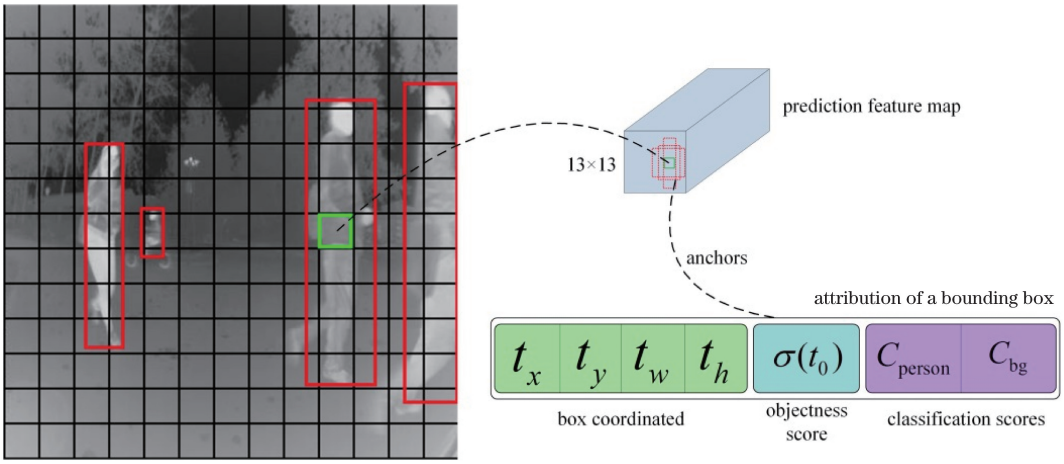

图 5　行人检测原理

Fig. 5　Principle of pedestrian detection

力。检测时利用 U-FOV 测试集验证模型的行人检测性能，并利用 LTIR(Linköping Thermal Infrared)数据集检验模型的泛化能力。

U-FOV 数据集由手工标注的 1000 张训练图像、200 张验证图像及 661 张测试图像组成，分辨率为 800×600。Caltech 行人数据集包含约 10 h、分辨率为 640×480、频率为 30 Hz 的视频，视频由车载摄像机在城市环境中拍摄得到，总计约 250000 帧图像、350000 标注框及 2300 个不同的行人。LTIR 是一个用于评价短时目标跟踪性能的热红外数据集，包含20个红外图像序列，每个序列平均含有 563 帧图像。

4.2　学习率与损失

训练分两个阶段：第一阶段固定主干网络卷积核参数不变，训练其他网络层参数；第二阶段开放训练整个网络模型参数，损失函数与文献[13]相同。在 Caltech 行人数据集上，每个阶段分别训练 10 个 epoch；在 U-FOV 红外行人数据集上，每个阶段分别训练 30 个 epoch。两次迁移训练的学习率及损失变化如图 6 所示，在 Caltech 数据集上的学习率一直保持 0.0001 不变，而在 U-FOV 数据集上的学习率则在第二阶段训练中随着损失降低而逐渐减小。

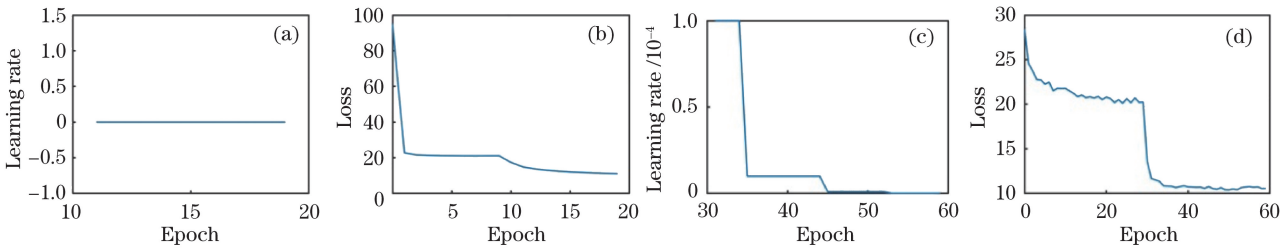

图 6　学习率和损失曲线。(a) Caltech 数据集上的学习率；(b) Caltech 数据集上的损失；(c) U-FOV 数据集上的学习率；(d) U-FOV 数据集上的损失

Fig. 6　Learning rate and loss curves. (a) Learning rate on Caltech dataset; (b) loss on Caltech dataset; (c) learning rate on U-FOV dataset; (d) loss on U-FOV dataset

4.3　显著性系数和特征图

图 7 为对应于图 2 第Ⅲ个残差模块组的输出 y_3 的显著系数图 α_2 和显著特征图 s_2，它们的分辨率为 52×52。第 1 列为输入图像，每幅输入图像包含不同数量、尺度及背景的行人。第 2 列为显著系数图，其元素值都介于 0～1 之间，表征特征图不同区域的重要性，为了形象直观的表征，采用 matplotlib 中的“jet”模式对其进行展示，下方部分颜色越深表示越接近于 0，重要性越低。对比输入图像和显著系数图可以发现：显著系数图中下方部分区域基本都对应着输入图像中的道路等具有较高相似性，较低显著性的区域；上方部分对应的输入图像区域则是结构比较复杂，显著性较高的区域。这说明注意力模块在网络中确实起到了抑制输入图像中不相关区域，突出显著特征的作用。第 3～9 列是经过显著系数处理后的特征图，与目标无关的背景区域的特征响应逐渐被抑制，模型能更加智能地聚焦于特定的局部区域。

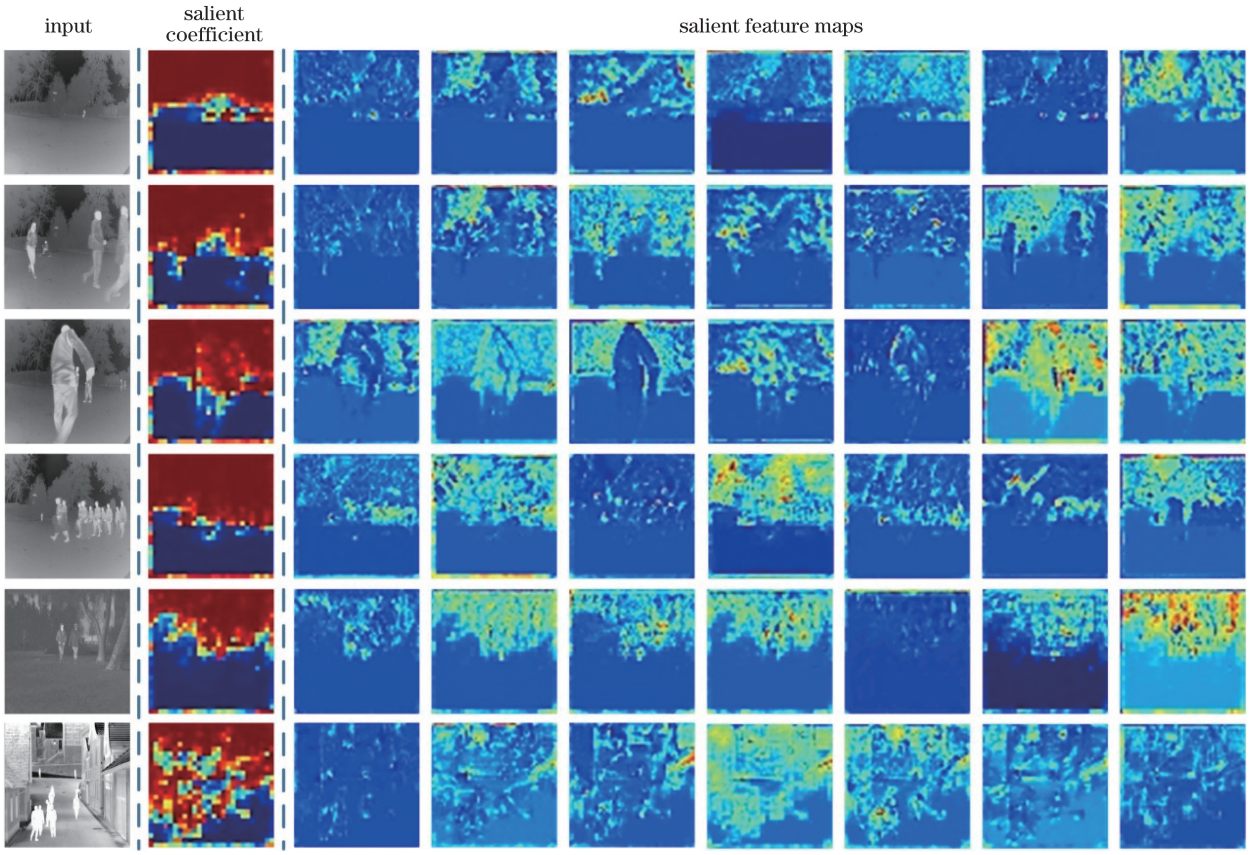

图 7　显著性系数和显著特征图

Fig. 7　Salient coefficient and feature maps

4.4　U-FOV 数据集上的行人检测结果

图 8 为测试集中行人目标尺度的分布情况，横、纵坐标分别表示归一化目标的宽和高。可以看出，绝大部分行人的归一化宽和高都小于 0.1，甚至集中于 0 附近，因此数据集中存在较多的小尺度目标，检测难度较大。引入 Caltech 行人数据集进行多次迁移训练，能为模型提供更为丰富的行人特征，提高模型的鲁棒性和泛化能力。注意力模块的加入丰富了模型的显著特征，起到了联系前后不同尺度上下文信息、增强局部显著性及抑制不相关区域的作用，进一步提高了模型的行人检测性能。

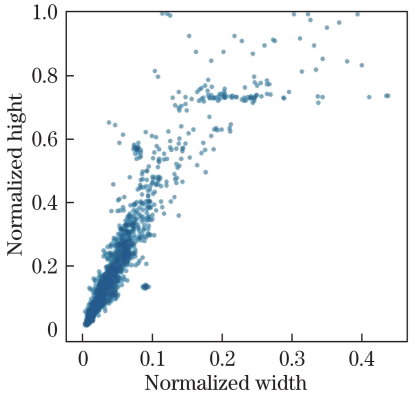

图 8　U-FOV 测试集中的行人尺度分布

Fig. 8　Distribution of pedestrian size in U-FOV test set

图 9 为 5 幅不同尺度测试图像的可视化检测结果。第 1 行是输入图像，第 2～5 行是不同深度学习模型的检测结果，最后一行是利用所提方法得到的检测结果。对比图 9 第 1、2 列处理结果，所提方法取得了很好的小尺度目标检测效果。对于图 9 第 1 列中的 5 个小尺度目标，SSD 方法由于采用了效率更高的 MobileNet，在卷积特征提取上不具备优势，导致小目标全部漏检；其他方法仅能检测到中间较为显著的一个目标，存在四个漏检目标；而所提方法能检测出全部目标，且检测结果的最小边框大小为 6×13。对比图 9 第 3、4、5 列，Faster R-CNN 在图 9 第 3 列中存在漏检，在图 9 第 5 列中存在将多个目标回归到一个目标边框的问题；SSD 对于密集行人则存在较多的漏检；R-FCN[32] 在处理行人遮挡问题时，也容易存在漏检；YOLOv3 和所提算法处理密集行人检测问题的能力略强。总体而言，所提方法在保证中、大尺度目标检测率的基础上，极大提高了小尺度目标的检测性能，适合用于处理 U-FOV 图像数据中的行人目标。

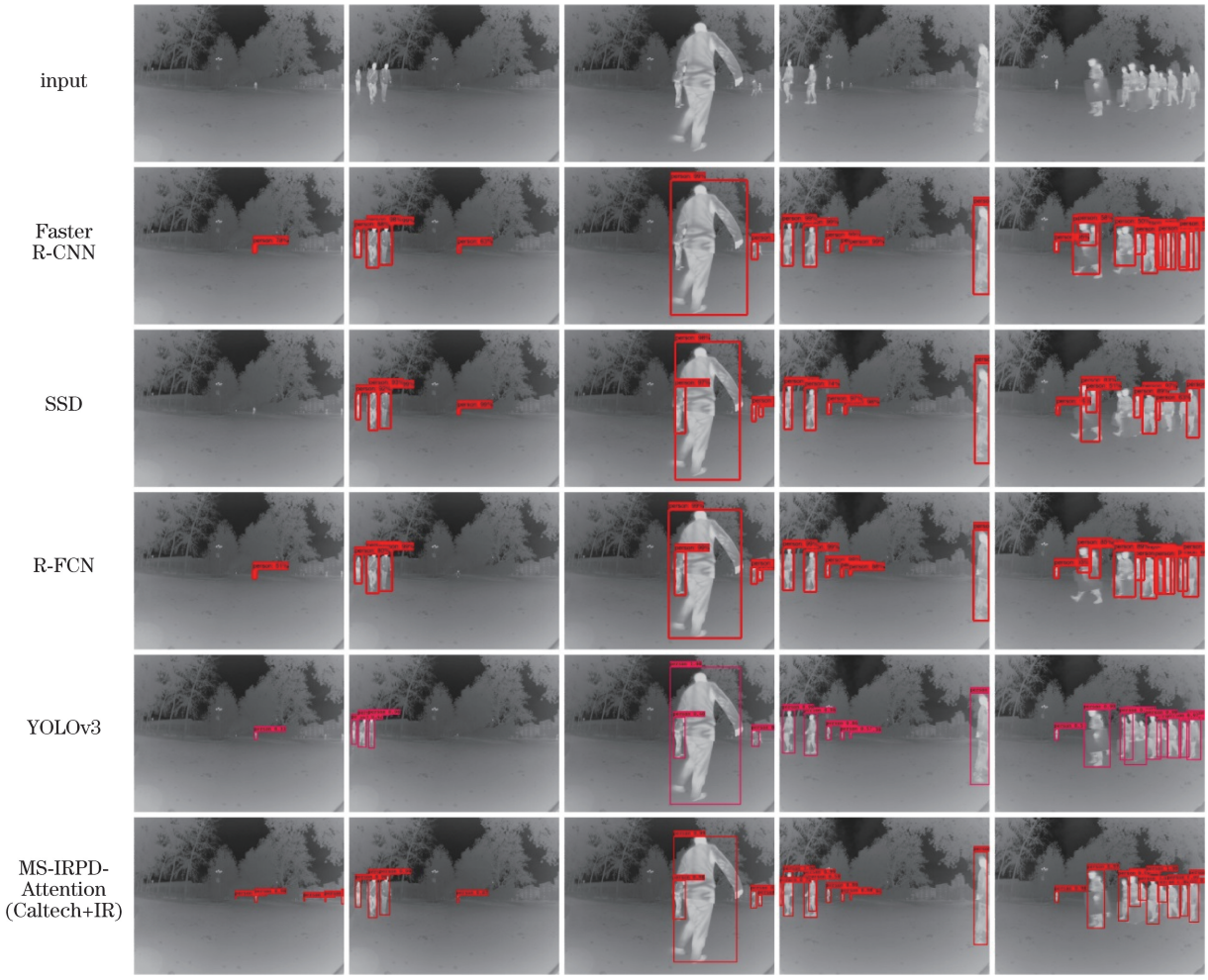

图 9　红外行人检测可视化结果

Fig. 9　Visualization results of infrared pedestrian detection

为进一步定量评估检测方法性能，选择 P-R (precision-recall)曲线作为评价指标，P-R 曲线刻画了查准率和查全率之间的关系。准确率和召回率定义为

$$P_{\text{precision}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}},R_{\text{recall}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}},\quad(2)$$

式中：N_{TP}为正样本被正确识别为正样本的数量；N_{FP}为负样本被识别为正样本的数量；N_{FN}为正样本被识别为负样本的数量。改变置信度阈值计算对应的准确率和召回率，可以得到 P-R 曲线。如果检测器分类性能好，那么在 R_{recall}增长的同时，$P_{\text{precision}}$应当保持在高水平。曲线下的面积代表检测器的平均准确率(AP)，表征检测器对该类目标的检测性能。

图 10 和表 1 分别给出了不同 IoU(intersection over union)阈值下的测试集上检测结果的 P-R 曲线及其对应的 AP 值。MS-IRPD 表示未加入注意力模块的四尺度特征图行人检测模型；IR 表示仅在 U-FOV 红外行人数据集上进行迁移训练；Caltech+IR 表示先在 Caltech 数据集上进行第一次迁移训练，再在自建数据集上进行第二次迁移训练；MS-IRPD-Attention 表示加入了注意力模块的行人检测模型。在测试集上检测行人时，通过设置不同的 IoU 阈值得到不同的 P-R 曲线，并保存了置信度分数大于 0.3 的检测结果。IoU 阈值越大，表示对预测边框的要求越严格，但容易出现更多的漏检目标；IoU 阈值越小，表示对预测边框的要求越宽松，但容易出现虚警目标。经对比发现，IoU 阈值处于 0.5 附近时检测性能较好，当 IoU 为 0.45 时，检测性能最好，因此将 0.45 设置为默认检测阈值。此时，MS-IRPD-Attention 的行人检测的 AP 达到了 93.45%，比 YOLOv3 高 26.74 个百分点。这些增益中，四尺度预测网络贡献了 21.99%，注意力模块贡献了 2.67%，两次迁移训练方式贡献了 2.08%。

由不同方法的性能评价结果可以看出，对加入低层高分辨率的特征图进行四尺度目标预测后，检测性能得到明显改善，相比于YOLOv3原始模型，

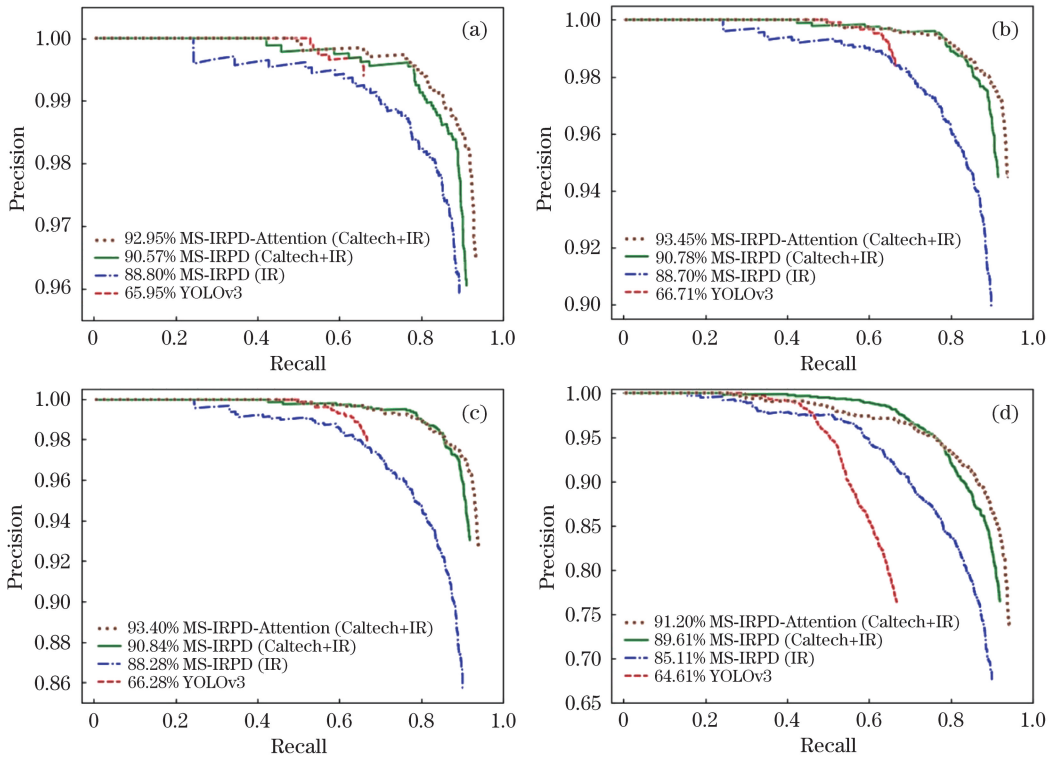

图 10　不同 IoU 阈值下的 P-R 曲线。(a) IoU 阈值为 0.3;(b) IoU 阈值为 0.45;(c) IoU 阈值为 0.5;(d) IoU 阈值为 0.7

Fig. 10　P-R curves under different IoU thresholds. (a) IoU threshold is 0.3; (b) IoU threshold is 0.45; (c) IoU threshold is 0.5; (d) IoU threshold is 0.7

表 1　不同 IoU 阈值下的行人检测平均准确率

Table 1　Average precision of pedestrian detection under different IoU thresholds

Method	Backbone	Dataset	AP			
			IoU is 0.3	IoU is 0.45	IoU is 0.5	IoU is 0.7
Faster R-CNN	ResNet101	U-FOV IR			0.5932	
SSD	Mobilenet_v1	U-FOV IR			0.5584	
R-FCN	ResNet101	U-FOV IR			0.6312	
YOLOv3	Darknet53	U-FOV IR	0.6595	0.6671	0.6628	0.6461
MS-IRPD	Darknet53	U-FOV IR	0.8880	0.8870	0.8828	0.8511
MS-IRPD	Darknet53	Caltech+IR	0.9057	0.9078	0.9084	0.8961
MS-IRPD-attention	Darknet53	Caltech+IR	0.9295	0.9345	0.9340	0.9120

其 AP 提高超过 20 个百分点，主要原因在于 U-FOV 数据集存在大量尺度过小的行人目标。

由于每幅测试图像包含的内容不同，所需的时间不同，因此在表 2 中给出了在本系统中遍历 10 次 U-FOV 测试集 661 张测试图像的平均总耗时和平均处理帧速(FPS)。YOLOv3＋Attention 表示在 YOLOv3 模型的基础上仅增加注意力模块，评估注意力模块带来的计算负担可得，平均处理时间仅提高了约 5.04%，说明设计的注意力模块具备轻量特性。所提方法是通过增加更底层高分辨的特征图来增加小目标的检测性能，故需要处理的锚框数量大幅增加，计算量也因此增加。

表 2　U-FOV 测试集的总处理时间

Table 2　Total times of U-FOV test set

Method	YOLOv3	YOLOv3＋Attention	MS-IRPD-Attention
Total time /s	90.75	95.32	125.21
FPS	7.28	6.93	5.28

4.5　扩展实验

为了进一步验证模型对于红外行人目标检测的

鲁棒性和泛化能力，选择分辨率较高的 LTIR 数据集作为测试对象，对其中存在行人的红外图像进行检测，图 11 给出了不同场景下的行人检测结果。该结果是在不利用 LTIR 数据集训练的情况下，直接利用 MS-IRPD-Attention 模型检测得到的结果。从可视化结果可以看出，尽管行人在不同背景中呈现出不同的红外特性，甚至存在遮挡等问题，但所提方法仍能有效检测出图像中存在的行人目标，证明了所提模型具有较好的鲁棒性和泛化能力。一方面是因为 Darknet53 预训练模型本身就具备一定的泛化能力；另一方面得益于 Caltech 数据集的迁移训练过程，从而模型获得了充足的行人特征样本。

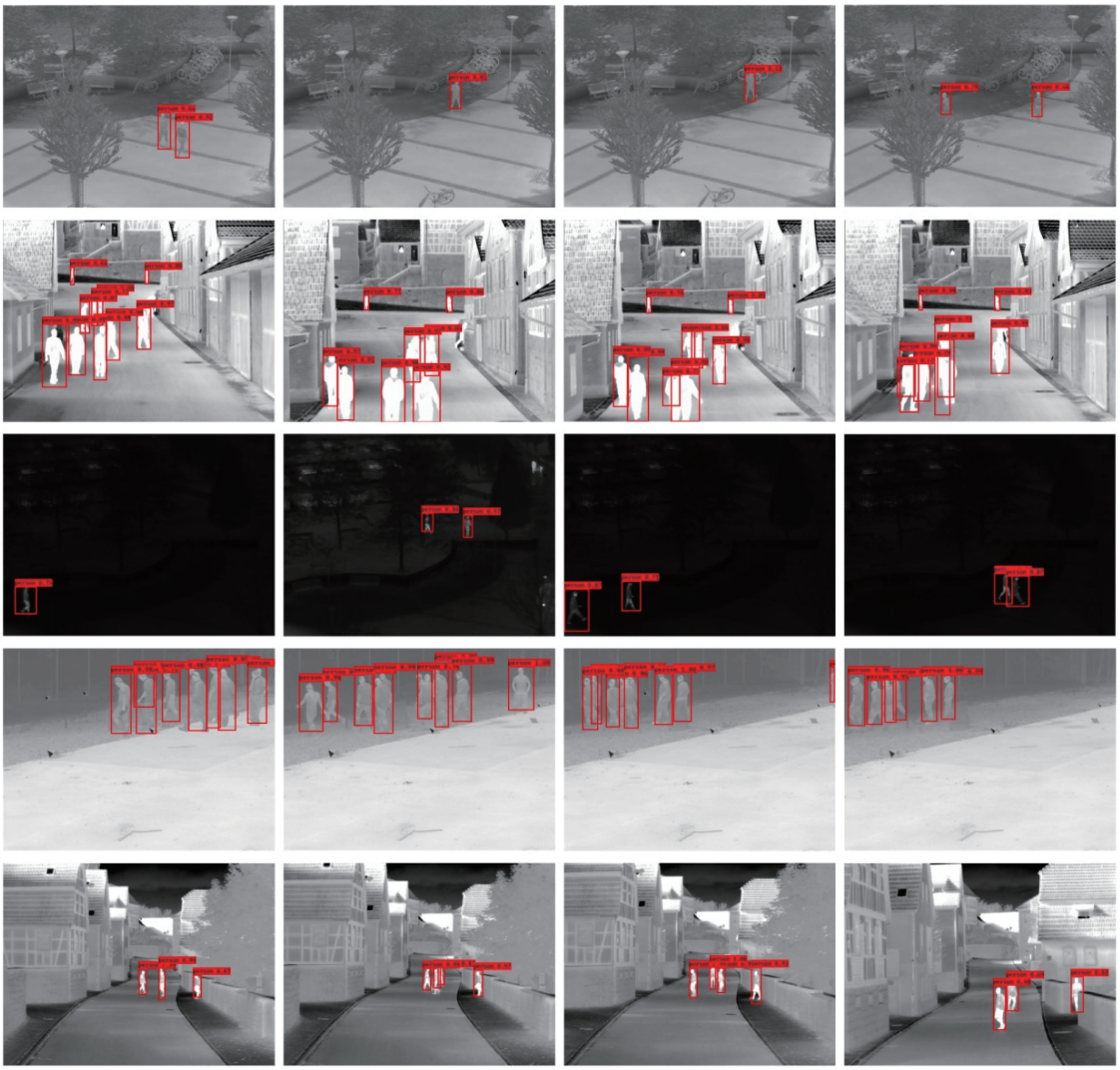

图 11　不同场景下 LTIR 数据集上的红外行人检测可视化结果

Fig. 11　Visualization results of infrared pedestrian detection on LTIR dataset at different scenes

5　结　　论

为实现更宽范围内的夜间行人自动检测，文中率先利用超大视场红外相机收集路况信息，为行人检测领域引入更具挑战性的新任务，为此提出一种基于深度注意力机制的多尺度红外行人检测方法。在四个尺度的特征图上构建特征金字塔预测多尺度目标，增强了小尺度目标的检测能力。设计注意力模块，产生显著系数图并对前三个尺度的预测特征图进行处理，突出重点局部区域，进一步提高了检测性能。同时，利用两个行人数据集进行迁移训练，补充了丰富的行人样本特征，改善了模型的泛化能力。实验结果表明，所提方法相比于 YOLOv3，对多尺度行人检测的 AP 增加了 26.74 个百分比，具有较强的泛化能力，适合用于检测多尺度红外行人目标。

参　考　文　献

[1]　Liu S T, Jiang N, Liu Z X, et al. Saliency detection of infrared image based on region covariance and global feature [J]. Journal of Systems Engineering

and Electronics, 2018, 29(3): 483-490.

[2] Cai Y F, Liu Z, Wang H, et al. Saliency-based pedestrian detection in far infrared images[J]. IEEE Access, 2017, 5: 5013-5019.

[3] Hintermüller M, Wu T. Robust principal component pursuit via inexact alternating minimization on matrix manifolds[J]. Journal of Mathematical Imaging and Vision, 2015, 51(3): 361-377.

[4] Shu X B, Porikli F, Ahuja N. Robust orthonormal subspace learning: efficient recovery of corrupted low-rank matrices[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 3874-3881.

[5] Ye X C, Yang J Y, Sun X, et al. Foreground-background separation from video clips via motion-assisted matrix restoration[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(11): 1721-1734.

[6] Cherapanamjeri Y, Gupta K, Jain P. Nearly optimal robust matrix completion[C]//Proceedings of the 34th International Conference on Machine Learning, August 6-11, 2017, Sydney, NSW, Australia. USA: MIT Press, 2017, 70: 797-805.

[7] Sobral A, Javed S, Jung S K, et al. Online stochastic tensor decomposition for background subtraction in multispectral video sequences[C]//2015 IEEE International Conference on Computer Vision Workshop (ICCVW), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 946-953.

[8] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 580-587.

[9] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1440-1448.

[10] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems, December 7-12, 2015, Montreal, Quebec, Canada. New York: Curran Associates, 2015: 91-99.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[12] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 6517-6525.

[13] Redmon J, Farhadi A. Yolov3: an incremental improvement[J/OL]. (2018-04-08)[2019-09-22]. https://arxiv.xilesou.top/abs/1804.02767.

[14] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[15] Fu C Y, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector[J/OL]. (2017-01-23)[2019-09-22]. https://arxiv.xilesou.top/abs/1701.06659.

[16] Wang F, Jiang M Q, Qian C, et al. Residual attention network for image classification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6450-6458.

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.

[18] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[19] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-Net: learning where to look for the pancreas[J/OL]. (2018-05-20)[2019-09-22]. https://arxiv.xilesou.top/abs/1804.03999.

[20] Tang X, Du D K, He Z Q, et al. PyramidBox: a context-assisted single shot face detector[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11213: 812-828.

[21] Qin J, Wang M H. Fast pedestrian proposal generation algorithm using online Gaussian model[J]. Acta Optica Sinica, 2016, 36(11): 1115001.
覃剑, 王美华. 采用在线高斯模型的行人检测候选框快速生成方法[J]. 光学学报, 2016, 36(11): 1115001.

[22] Zhao P R, Wu X Y, Tang X Y, et al. An algorithm of small object detection region proposal search based on GN splitting[J]. Acta Optica Sinica, 2018, 38(9):

0915005.
赵沛然, 吴新元, 汤新雨, 等. 基于 GN 分裂的小目标检测区域推荐搜索算法[J]. 光学学报, 2018, 38(9): 0915005.

[23] Cheung W, Hamarneh G. *n*-SIFT: *n*-dimensional scale invariant feature transform[J]. IEEE Transactions on Image Processing, 2009, 18(9): 2012-2021.

[24] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005: 8588935.

[25] Zhang C J, Liu J, Liang C, et al. Image classification using Harr-like transformation of local features with coding residuals[J]. Signal Processing, 2013, 93(8): 2111-2118.

[26] Ye G L, Sun S Y, Gao K J, et al. Nighttime pedestrian detection based on faster region convolution neural network [J]. Laser & Optoelectronics Progress, 2017, 54(8): 081003.
叶国林, 孙韶媛, 高凯珺, 等. 基于加速区域卷积神经网络的夜间行人检测研究[J]. 激光与光电子学进展, 2017, 54(8): 081003.

[27] Aimar A, Mostafa H, Calabrese E, et al. NullHop: a flexible convolutional neural network accelerator based on sparse representations of feature maps[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(3): 644-656.

[28] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[29] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 2818-2826.

[30] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 936-944.

[31] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//32th International Conference on Machine Learning, July 6-11, 2015, Lille, France. USA: MLR Press, 2015: 448-456.

[32] Dai J, Li Y, He K, et al. R-FCN: object detection via region-based fully convolutional networks[C]//Advances in Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. New York: Curran Associates, 2016: 379-387.