

基于差分特征注意力机制的无锚框多光谱行人检测算法

沈继锋¹, 刘岳¹, 韦浩², 左欣³, 杨万扣⁴

(1. 江苏大学电气信息工程学院, 江苏 镇江 212013;

2. 中电海康集团有限公司, 浙江 杭州 310000;

3. 江苏科技大学计算机学院, 江苏 镇江 212003;

4. 东南大学自动化学院, 江苏 南京 210003)

摘要: 针对多光谱行人检测系统存在特征融合质量低、模型超参数多且锚框匹配算法复杂等问题, 提出了一种基于差分特征注意力机制的无锚框多光谱行人检测算法。该算法首先采用差分特征感知融合方法挖掘多模态特征间的互补信息来优化通道特征; 然后利用具有高效无锚框机制的 CenterNet 检测框架大大降低了模型计算复杂度, 从而提升检测速度; 最后引入差分特征注意力机制, 改善特征融合质量, 进一步提升检测精度。在 KAIST、CVC14 和 FLIR 这 3 个公开数据集上的实验结果表明, 提出的算法和其他先进方法相比, 能够同时有效提升检测精度和速度, 具有较好的实际应用前景。

关键词: 多光谱行人检测; 无锚框机制; CenterNet 模型; 注意力机制

中图分类号: TP242

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202130

Anchor free multispectral pedestrian detection algorithm based on differential feature attention mechanism

SHEN Jifeng¹, LIU Yue¹, WEI Hao², ZUO Xin³, YANG Wankou⁴

1. School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China

2. China Electric Haikang Group Co., Ltd., Hangzhou 310000, China

3. School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China

4. School of Automation, Southeast University, Nanjing 210003, China

Abstract: Multispectral pedestrian detection system suffers with low feature fusion quality, high quantity of model hyper-parameters and complex anchor matching algorithm. To deal with these problems, an anchor free multispectral pedestrian detection algorithm based on differential feature attention mechanism was proposed. Firstly, differential modality aware fusion was used to obtain the complementary information between different modalities to optimize the channel features. Secondly, the CenterNet detection framework with anchor free mechanism was adopted to greatly reduce the computational complexity of the model and thus improve the detection speed. Finally, differential feature guided attention mechanism was introduced to improve the quality of feature fusion and further enhance the detection accuracy. Experimental results on three open datasets, KAIST, CVC14 and FLIR, show that the proposed algorithm can effectively improve the detection accuracy and speed compared with the current advanced methods, and has a good practical application prospect.

Key words: multispectral pedestrian detection, anchor free mechanism, CenterNet model, attention mechanism

收稿日期: 2021-07-21; 修回日期: 2021-08-12

通信作者: 沈继锋, shenjifeng@ujs.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61903164); 江苏省自然科学基金资助项目 (No.BK20191427)

Foundation Items: The National Natural Science Foundation of China (No.61903164), The Natural Science Foundation of Jiangsu Province (No.BK20191427)

1 引言

随着以深度学习为基础的人工智能技术的发展,我国正在全面布局与人工智能相关的产业,自动驾驶、城市大脑、智能监控等众多领域都成为当前比较火热的研究方向^[1-2],而行人检测技术在这些领域中又具有重要的研究意义和应用价值。考虑到安全性和生产效率,行人检测算法应该具备精度高、速度快和延迟低等性能指标,并且应当具有极低的漏报率和误判率,而实际应用中检测设备面对的场景和气候等因素复杂多变,不同行人目标差异性较大,因此复杂场景下的行人检测是一个极具挑战性的研究课题。尤其在全天候应用场景中,行人检测算法应该对光照变化具有较强的鲁棒性,以适应不同的照明条件。

可见光图像和红外图像在白天和夜间各有优点,而多光谱行人检测提供了一种可以结合两种图像优点的解决方案。已有研究成果表明^[3-4],多模态图像的行人检测结果明显优于单一模态图像,但仍存在一些不足。首先,图像特征质量对检测性能具有重要影响,然而当前对特征级融合质量的研究相对较少;其次,多模态特征中存在严重的信息冗余,行人信息中时常混杂着大量的不利于检测的冗余信息,容易导致性能低下;最后,检测速度取决于模型的推理速度,当前基于锚框的检测器推理速度较慢,很难满足实际场景的应用需求。上述这些问题都阻碍了当前高精度多光谱行人检测算法的实际落地应用。

因此,本文从提升特征融合质量和算法运行速度两方面展开研究,提出了一种基于差分特征注意力机制的无锚框多光谱行人检测算法。针对特征融合质量问题,本文借鉴差分模态感知融合(differential modality aware fusion, DMAF)^[5]模块提出了一种差分特征注意力(differential feature attention, DFA)机制,聚焦可见光和红外图像特征的差异信息,更好地实现异质特征的互补性,从而实现性能提升。针对运行速度问题,本文借鉴了CenterNet检测框架,该框架去除了烦琐的锚框超参数设置和锚框匹配过程,大大提升了算法的运行速度。

近年来,已有一些学者在多光谱行人检测领域取得了一些成果。2011年,Grassi A P等人^[3]通过对比长波和短波红外图像发现,在夜间条件下长波

红外图像的特征更加明显,因此将长波红外图像作为可见光图像的补充。但是这种双模态图像未做到精密的光学配准,严重影响了检测器性能。2015年,Hwang S等人^[4]在莫里斯(Morris)搭建的可见光和红外图像采集系统的基础上进行了改进,增加了分光镜和三轴夹具,使得可见光和红外图像得到精密的光学配准。该系统采集到的全天候行人检测数据集被命名为KAIST。与此同时,作者提出了ACF+T+THOG多光谱行人检测算法。实验结果表明,将红外图像作为补充数据集训练出的模型的检测性能明显优于可见光图像行人检测算法。但该算法属于传统的机器学习方法,特征表达能力有限,从而导致性能出现瓶颈。

2016年,Jörg W等人^[6]首次利用深度卷积神经网络搭建了基于R-CNN体系的多光谱行人检测模型,其检测结果明显优于ACF+T+THOG多光谱行人检测算法。与此同时,作者详细研究了不同的特征融合策略对检测算法性能的影响,并提出了两种融合方法。实验结果表明,后期融合方法要优于早期融合方法。但是该模型使用的是Fast R-CNN框架,没有实现端到端可训练,因此会极大地影响模型的性能。2016年,Liu J J等人^[7]提出了双流Faster R-CNN,该网络用VGG16骨干网络提取可见光和长波红外图像的特征,再将两者特征级联融合并送入全连接层进行后续检测,实验结果表明,相较于当时的其他算法,该算法的检测性能得到明显提升。此算法虽然检测精度高,但是模型复杂度也高,单张图像检测时间过长,很难应用在实际领域。2018年,Vandersteegen M等人^[8]凭借YOLO^[9]算法的超高实时检测性能提出了基于YOLO框架的多光谱行人检测算法。该算法用到的检测器虽然在速度上达到了预期效果,但是在检测精度上明显低于前文所述的方法,因此同样难以应用在实际的交通以及自动驾驶领域。2020年,Zhou K L等人^[5]分析总结出了多光谱行人检测中存在的两大不平衡问题:光照模态不平衡问题和特征模态不平衡问题。针对这两个问题,作者提出了光照感知特征对齐模块、差分模态感知融合模块、光照门模块等来缓解模态不平衡问题。作者设计的一阶级联双流检测网络MBNet检测精度高,运行速度快,但它仍然是基于锚框的检测算法,没有摆脱基于锚框的复杂的超参数设计和计算,其检测速度无法有效满足实时检测算法的需求。

2 本文算法

2.1 算法结构设计

基于差分特征注意力机制的无锚框多光谱行人检测算法总体框架如图1所示。整个框架包含两个部分：特征提取模块^[10]和无锚框检测头模块。该算法将 VGG16^[11]作为可见光和红外图像的特征提取网络，并在 VGG16 的每层之间嵌入 DMAF^[5]模块和 DFA 模块。前者被用来解决多模态特征盲目融合的问题，而后者是一种注意力机制，其目的是提升特征的表达能力，过滤冗余信息。无锚框检测头模块采用无锚框检测框架，避免了锚框的超参数设置与后处理流程，极大地降低了计算量，并能够实现端到端训练。此外为了保证特征图的语义分辨率的同时提升模型的空间分辨率，本文对 conv5 进行了特征图上采样操作，从而提升了对小目标的检测精度。图1中 conv1 到 conv5 分别表示 VGG16 网络中的第一层至第五层卷积层。

2.2 经典特征融合方法

为了选取最佳的特征融合方式，本文分别测试了3种特征融合方案，如图2所示。

(1) 早期融合

早期融合将可见光和红外图像特征在 conv2 进行融合，conv2 的特征图纹理信息比较丰富且保留了目标更多的位置信息。可见光和红外图像特征融合的方法为：先将两个特征以通道的维度连接，然后利用 NIN (network-in-network)^[12]模块对通道进

行降维，接着分别对 conv4、conv5 特征图做反卷积 (deconvolution) 操作，然后对这3个特征图做连接融合操作得到最终的特征图，具体融合网络结构如图2(a)所示。

(2) 中期融合

中期融合将可见光和红外图像特征在 conv3 进行融合，融合模块依然是 NIN，然后对融合后的特征图和 conv4、conv5 特征图做反卷积操作，最后将这3种特征图连接融合得到最终的特征图，具体融合网络结构如图2(b)所示。

(3) 后期融合

后期融合将可见光和红外图像特征在 conv4、conv5 融合，然后分别做反卷积操作、连接融合操作，进而得到最终的特征图。根据 Liu J J^[7]提出的双流 Faster R-CNN 可知，conv4、conv5 的融合效果是最优的，因此本次实验同样将 conv4、conv5 作为特征融合层，不同的是，本文算法对采集后的特征图做了一次反卷积后再做融合，此外本文算法的后期融合是在 conv4、conv5 同时做融合操作，然后对这两个特征图分别进行检测，具体融合网络结构如图2(c)所示。

2.3 差分模态感知特征融合

为了解决可见光和红外图像特征盲目融合的问题，本文通过差分模态信息来增强双模态特征的表达能力^[13]。双模态特征融合最直接的方法是在不同卷积层上融合特征，然而通过传统的直接串联的方式来捕获跨模态互补信息具有歧义性，这是因为

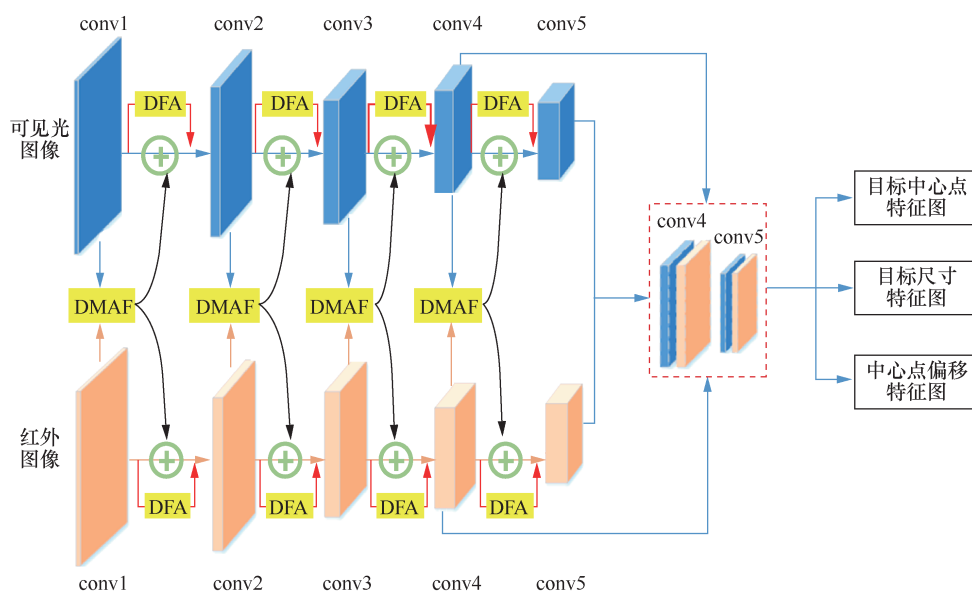


图1 基于差分特征注意力机制的无锚框多光谱行人检测算法总体框架

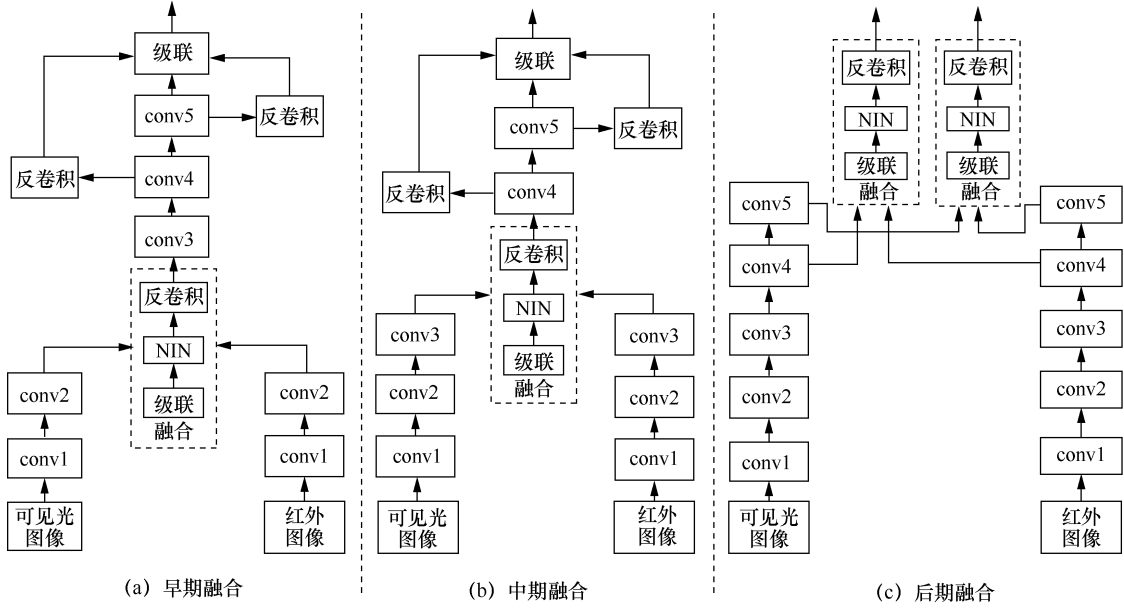


图2 不同特征融合网络结构

两种模态信息都有各自独立的特征，而这些特征都混合了有用信息和噪声。因此简单的特征融合（线性组合或串联）是无法准确得到跨模态特征信息的。本文通过借鉴 DMAF^[5]模块来挖掘这两种模态之间的固有差异。

受到差分放大器的启发，DMAF 模块利用了共模信号被抑制和差模信号被放大的原理。DMAF 模块的主要优点是在保留了原始特征的同时又进行了对应的差异特征补偿，进一步强化了多模态特征之间的联系。可见光 (F_V) 和红外 (F_T) 特征图可以用每个通道上的公共模态和差分模态进行表示，具体如式 (1) 和式 (2) 所示：

$$F_T = \frac{F_T + F_T}{2} + \frac{F_V - F_V}{2} = \frac{F_V + F_T}{2} + \frac{F_T - F_V}{2} \quad (1)$$

$$F_V = \frac{F_V + F_V}{2} + \frac{F_T - F_T}{2} = \frac{F_V + F_T}{2} + \frac{F_V - F_T}{2} \quad (2)$$

其中，公共模态反映了共同的特征，而差异模态反映了两种模态所捕获的独立特征。式 (1) 和式 (2) 分别阐释了差分放大电路和 DMAF 模块背后共同的分裂原理。DMAF 模块的关键思想是通过计算另一个模态的通道差分权重得到当前模态的互补特征，本文算法通过 DMAF 模块建立多模态间的相互依赖性来增强互补特征的学习，从而提高网络对于来自另一模态的特征信息的敏感性。

为了充分利用跨模态的互补差分信息，首先将 DMAF 模块插入 VGG16 的每一层卷积层。直接将双模态特征相减得到差分特征 F_D 。然后对 F_D 做自

适应全局平均池化的空间压缩操作，得到全局差分向量，该池化操作是 PyTorch 独有的，与普通全局平均池化的唯一区别是可以控制输出尺寸。此外全局差分向量可被解释为通道描述符，即可表示可见光与红外模态之间的差异。接着将全局差分向量输入 $\tanh$ 激活函数，以获得融合后的权重向量。再将此权重向量与双模态特征逐通道元素相乘得到差分特征。最后将此差分特征作为互补信息与原特征图相加得到新的特征图。具体如式 (3) 和式 (4) 所示：

$$F'_T = F_T + f(F_T \oplus F_{VD}) = F_T + f(F_T \oplus (\sigma(\text{AGAP}(F_D)) \otimes F_V)) \quad (3)$$

$$F'_V = F_V + f(F_V \oplus F_{TD}) = F_V + f(F_V \oplus (\sigma(\text{AGAP}(F_D)) \otimes F_T)) \quad (4)$$

其中， F'_V 和 F'_T 分别代表经注意力优化后的可见光和红外通道特征， $f(x)$ 代表残差函数，AGAP 代表自适应全局平均池化， σ 代表 $\tanh$ 激活函数， $\oplus$ 、 $\otimes$ 分别代表逐元素相加和逐元素相乘操作。此外值得关注的是， F_{VD} （可见光互补特征）和 F_{TD} （红外互补特征）通过 DMAF 模块被植入特征骨干网络中。

因为可见光成像和热辐射成像之间存在特征差异，所以在捕捉行人和背景特征时有一定局限性。随着卷积神经网络的发展，层数越来越深，行人的特征逐渐变得显著，背景特征被重新整合。这就意味着神经网络必须对有用的背景信息进行细

化,尽可能地消除背景里的噪声信息。DMAF 模块的重新组合多模态特征功能有助于背景信息的整合,使行人特征更加显著。因此本文认为 DMAF 模块有助于增强多模态特征之间的联系,减少冗余参数的学习并传递更多的信息。

2.4 差分特征注意力机制

卷积后的特征图在保存了有用信息的同时也包含了许多冗余信息,而通道注意力机制的目的是在特征通道和空间角度上过滤噪声信息。类似于 CBAM 方法^[14],本文提出的差分特征注意力机制整体结构(如图3所示)主要包括混合池化通道注意力和差分特征空间注意力两部分。混合池化通道注意力的具体实现步骤如下。首先,在空间维度上压缩特征图,获得一维矢量。本文对每个通道上的特征图同时采取最大池化、全局平均池化、全连接(1×1 卷积)压缩操作。如果只采用最大池化,会丢失一些特征关联信息,因此用全局平均池化来弥补这一缺点。但是以上两种池化会导致收敛速度减慢,因此将全连接层作为第3种压缩方法。然后,经过 NIN^[12]得到最终的注意力权重,而 NIN 就是将3种压缩方法得到的一维矢量逐元素相加,接着利用 1×1 卷积将上一步相加后得到的一维矢量缩小到原始尺寸的 1/16,再通过 1×1 卷积恢复到原始尺寸。

类似于通道注意机制,网络还需要能够理解特征图的哪些部分在空间上应该具有更高的响应级别。差分特征空间注意力的具体实现步骤如下。对于可见光通道上的注意力机制,首先将可见光通道特征和红外通道特征相减,然后将计算结果在通道维度上做平均池化,再经过 tanh 激活层,得到最终

的基于差分特征的空间注意力权重,将此权重与原可见光通道特征相乘得到优化后的可见光通道特征。对于红外通道上的注意力机制,首先使用红外通道特征减去可见光通道特征,然后将计算结果经池化层和 tanh 激活层后得到相应的红外通道差分特征空间注意力权重,这种计算方法能够使网络更好地拟合空间中的复杂相关性,最后将通道和空间注意模块级联以获得最终的空间和通道混合的注意力机制。

2.5 CenterNet 检测框架

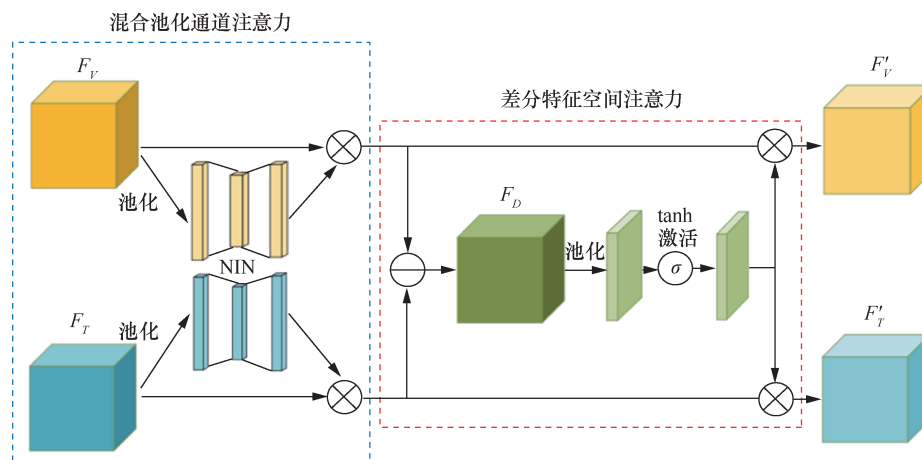
相比于经典的二阶段目标检测方法^[15], CenterNet^[16]是一种基于无锚框机制的通用目标检测方法,该方法具有更快的检测速度和较好的检测精度。与经典的输出检测框坐标方法不同,该方法输出目标中心点、目标框尺寸和中心点偏移量3种参数。

目标中心点通过高斯核函数输出一个关键点 $\hat{Y} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$, 其中 W 、 H 分别为输入图像的宽和高, R 是步长, C 是目标检测中的类别个数。如果 $\hat{Y}=1$, 则该点为中心点; 如果 $\hat{Y}=0$, 则该点为背景。本文采用 focal loss^[17]表示中心点预测的损失函数, 如式(5)、式(6)所示:

$$L_k = \frac{-1}{N} \sum_{yyc} \begin{cases} (1 - \hat{Y}_{yye})^\alpha \log(\hat{Y}_{yye}), Y_{yye} = 1 \\ (1 - Y_{yye})^\beta (\hat{Y}_{yye})^\alpha \log(1 - \hat{Y}_{yye}), \text{其他} \end{cases} \quad (5)$$

$$Y = \exp\left(-\frac{(x - \hat{P}_x)^2 + (y - \hat{P}_y)^2}{2\sigma_p^2}\right) \quad (6)$$

其中, σ 是目标尺寸的自适应标准差, $(\hat{P}_x, \hat{P}_y)$ 表示



⊖和⊗分别代表逐元素相减和逐元素相乘, σ代表tanh激活层

图3 差分特征注意力机制整体结构

热力图中目标中心点的坐标, α 、 β 是 focal loss 的超参数, N 是每幅图像关键点的个数, $\hat{Y}_{\text{yyc}}$ 表示当前中心点对应类别的得分。本文在实验中设置 $\alpha=2$ 和 $\beta=4$ 。

由于池化和下采样会造成中心点的偏移, 因此针对中心点额外预测了一个偏移量 $\hat{O}_{\hat{p}} \in R^{\frac{W}{R} \times \frac{H}{R} \times 2}$, 类别 C 中每一个关键点的真实值 P 都对应一个低分辨率等价的 $\hat{P} = \left\lceil \frac{P}{R} \right\rceil$, 其中 $\lceil \cdot \rceil$ 代表向上取整, 表示的是标注信息从输入图像映射到输出特征图时由于取整操作带来的坐标误差。 L_{off} 将 L1 loss 作为中心点偏移量的损失函数, 如式 (7) 所示:

$$L_{\text{off}} = \frac{1}{N} \sum_p \left| \hat{O}_{\hat{p}} - \left(\frac{P}{R} - \hat{P} \right) \right| \quad (7)$$

假设目标 k 的坐标为 $(x_1^{(k)}, y_1^{(k)}, x_2^{(k)}, y_2^{(k)})$, 则目标尺寸为 $s_k = (x_2^{(k)} - x_1^{(k)}, y_2^{(k)} - y_1^{(k)})$ 。该网络的坐标回归将 L1 loss 作为损失函数 L_{size} , 如式 (8) 所示:

$$L_{\text{size}} = \frac{1}{N} \sum_{k=1}^N |S_{pk} - s_k| \quad (8)$$

其中, S_{pk} 代表预测尺寸, 整个网络的损失函数 L_{det} 如式 (9) 所示:

$$L_{\text{det}} = L_k + \lambda_{\text{size}} L_{\text{size}} + \lambda_{\text{off}} L_{\text{off}} \quad (9)$$

在实验中, 设置 $\lambda_{\text{size}} = 0.1$ 和 $\lambda_{\text{off}} = 1$ 。整个网络主要用来预测关键点 $\hat{Y}$ 、偏移量 $\hat{O}$ 以及尺寸 $\hat{S}$ 。在特征图中的每一个位置, 网络共预测 $C+4$ (2 个尺寸变量和 2 个关键点偏移量) 个值。

在模型推演环节中, 分别提取每个通道热力图上的峰值点。峰值点的提取方法采用 3×3 的最大池化, 即筛选出邻域里的 9 个点的最大值。假设此最大值作为预测到的关键点对应类别 C , 本文用 $\hat{P}_c = (\hat{x}_i, \hat{y}_i)_{i=1}^n$ 表示预测到的 n 个中心点的集合, 假设每个真实关键点的位置坐标是 (x_i, y_i) 。用关键点 $\hat{Y}_{x,y,c}$ 的值来评估检测点的置信度, 最终目标框的表达式如下:

$$\begin{aligned} & (\hat{x}_i + \delta \hat{x}_i - \hat{w}_i / 2, \hat{y}_i + \delta \hat{y}_i - \hat{h}_i / 2, \hat{x}_i + \\ & \delta \hat{x}_i + \hat{w}_i / 2, \hat{y}_i + \delta \hat{y}_i + \hat{h}_i / 2) \end{aligned} \quad (10)$$

其中, $(\hat{x}_i, \hat{y}_i)$ 是预测的中心点坐标, $\delta \hat{x}_i$ 、 $\delta \hat{y}_i$ 是预测的中心点偏移, $\hat{w}_i$ 、 $\hat{h}_i$ 是预测目标的尺寸。整个网络的输出直接通过关键点预测来完成, 不需要计

算交并比 (intersection over union, IoU) 进行非极大值抑制或其他后处理。

3 实验结果与分析

3.1 实验参数设置

本文算法利用 PyTorch 实现, 软件环境: Python 3.5, PyTorch 1.3, CUDA 10.1, CuDNN 7.603, Ubuntu 16.04, OpenCV 3.1.0。硬件环境: 显卡 GTX 1080TI, CPU i7 9700。在训练时, conv1 和 conv2 参数固定不变, 其余卷积层的参数需要根据反向传播算法做对应的调整。本网络训练数据集共包含 50 172 对 (红外和可见光图像), 模型共迭代了 20 个周期 (epoch), 批大小 (batch size) 设置为 10, 初始学习率设为 1.25×10^{-3} , 第 10 个周期的学习率降为 1.25×10^{-4} 。为了扩展数据集以增强模型的检测性能, 本文采用随机裁剪、翻转方法对双模态数据集进行增强。

本文采取的性能评价指标为平均缺失率 (MR^{-2}) [4], 表示 FPPI (false positive per image) 在 [0.01, 1] 范围内 log 空间均匀取 9 个点时的缺失率的平均值, 即 FPPI 分别为 [0.010 0, 0.017 8, 0.031 6, 0.056 2, 0.100 0, 0.177 8, 0.316 2, 0.562 3, 1.000] 时的平均缺失率 (以下简称缺失率)。 MR^{-2} 值越低, 模型性能越好。

3.2 消融实验分析

(1) 特征融合方法的对比实验

特征融合消融实验见表 1, 前两组实验为采用单通道的检测结果, 即不采取任何融合方案, 通过结果对比可知: 采用任意一种双通道特征融合的方法性能均大幅度优于不采取融合方案的模型性能, 这说明对于行人检测任务来说, 采取双通道特征融合进行检测更有利于模型性能的提升。进一步观察表 1 可以发现, 后期融合模型在全天、白天和夜间场景下的缺失率都低于其他融合方案的模型。早期融合由于其特征图语义性较弱, 会包含较多噪声, 因此其特征表达能力也较弱。而后期融合具有较高的语义分辨率, 经过多次卷积和池化操作可以有效增强特征图的表达能力, 且后期融合的所有模型的缺失率均比早期融合低 7.96%~9.07%。此外, 在后期融合方案下, 全天、白天和夜间场景下的缺失率均在 14% 左右, 而白天场景下的行人检测缺失率为 14.39%, 比全天条件下的缺失率低 0.12%。由此可以得出, 后期融合方案对模型整体性能有较大的提升。

表1 特征融合消融实验结果

融合方案	全天	白天	夜间
仅用可见光	30.37%	29.62%	33.76%
仅用红外	28.21%	30.56%	26.91%
早期融合	23.58%	22.35%	23.12%
中期融合	19.41%	17.32%	21.28%
后期融合	14.51%	14.39%	14.84%

(2) 上采样方法的对比实验

对特征图进行上采样可以有效提高特征的空间分辨率，从而提升对小目标的检测精度。在后期融合的基础上，笔者分别采用了3种不同的上采样方法进行性能比较，实验结果见表2。从表2可以发现，反卷积方法检测精度最高，比基础算法的缺失率低2%；还可以发现上池化方法性能提升效果不明显（缺失率仅降低了0.29%），原因在于上池化方法采用空位补零的方式来扩大特征图的分辨率，不增加额外信息，因此性能差异不明显。而双线性插值方法在相邻的像素点上进行了插值计算，保留了原始特征图的信息并进行了扩展，因此性能要优于上池化方法，且双线性插值方法的性能仅次于反卷积方法。因此本文将后期融合和反卷积方法作为基础算法与其他方法进行对比分析。

表2 上采样方法消融实验结果

上采样方法	MR ⁻²
基础算法	14.51%
双线性插值	13.43%
反卷积	12.51%
上池化	14.22%

(3) DMAF 与 DFA 消融实验

使用 DMAF 与 DFA 的消融实验结果见表3。从表3可以发现，相比基础算法，单独采用 DMAF 可以使缺失率降低2.01%，而单独使用 DFA 可以使缺失率降低1.39%，实验结果证明了这两个模块的有效性。另外，同时使用 DMAF 和 DFA 可以使缺失率降低3.03%，说明这两种方法具有正交互补性。

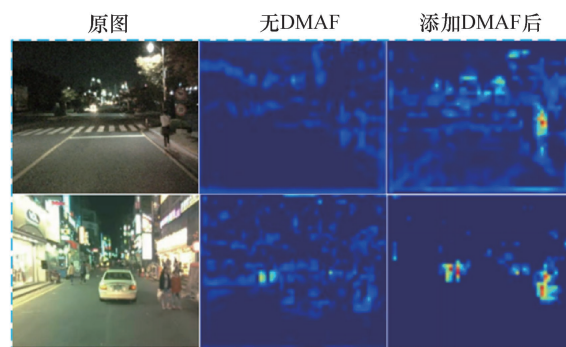
表3 使用 DMAF 与 DFA 的消融实验结果

方法	MR ⁻² （全天）
基础算法	12.51%
基础算法+DMAF	10.50%
基础算法+DFA	11.12%
基础算法+DMAF+DFA	9.48%

为了进一步定性分析 DMAF 模块，本文在两组白天和夜间图像的深度特征 conv5 的特征图上比较了基础算法（无 DMAF 模块）和添加 DMAF 模块后的可视化结果，如图4所示。从图4可以明显发现，添加 DMAF 模块后的特征图中目标区域特征响应更加显著，这也证明本文引用的 DMAF 模块的有效性。



(a) 白天场景下的可视化结果



(b) 夜间场景下的可视化结果

图4 特征图 conv5 可视化结果示例

3.3 与其他先进算法的对比实验

(1) 在 KAIST 数据集上的对比实验

本文算法与其他先进算法在 KAIST 数据集上的实验结果见表4，其中“近距离”代表高度大于115像素的行人，“中距离”代表高度为45~115像素的行人，“远距离”代表高度小于45像素的行人；对于行人是否被遮挡则分为无遮挡、部分遮挡和严重遮挡3种等级。为了进一步提升模型性能，笔者在使用了 DMAF 模块和 DFA 机制的基础上对网络添加了可偏移卷积^[18]。普通卷积因为感受野均为方形而不能很好地覆盖实际识别过程中复杂的物体形状信息，而可偏移卷积在普通卷积的基础上额外学习了采样位置（感受野）的偏移，使得学习到的特征图感受野可以自动拟合物体的形状信息，进而提升模型性能。通过观察表4可知，添加可偏移卷

表4 本文算法与其他先进算法在 KAIST 数据集上的实验结果

方法	全天	白天	夜间	近距离	中距离	远距离	无遮挡	部分遮挡	严重遮挡
ACF+C+T ^[4]	47.32%	42.47%	56.17%	28.74%	53.67%	88.20%	62.94%	81.40%	88.08%
RPN+BF ^[20]	18.29%	19.57%	16.27%	0.04%	30.87%	88.86%	47.45%	56.10%	72.20%
Halfway Fusion ^[7]	25.75%	24.88%	26.59%	8.13%	30.34%	75.70%	43.13%	65.21%	74.36%
IAF ^[21]	15.73%	14.55%	18.26%	0.96%	25.54%	77.84%	40.17%	48.40%	69.70%
IATDNN ^[22]	14.95%	14.67%	15.72%	0.04%	28.55%	83.42%	45.43%	46.25%	64.57%
AR-CNN ^[19]	10.43%	11.34%	8.85%	0.79%	13.15%	37.41%	20.61%	33.29%	61.42%
MBNet ^[5]	8.01%	8.34%	7.89%	0.12%	14.09%	53.75%	25.71%	35.24%	58.16%
本文算法	8.76%	8.76%	8.71%	0.01%	15.71%	51.14%	27.02%	33.93%	61.74%

积后的本文算法在各种条件下都能够表现出优异的性能，特别是在近距离条件下缺失率已达到0.01%，相比 ACF+C+T 降低了 28.73%。在无遮挡、部分遮挡和严重遮挡条件下，本文算法的缺失率比 Halfway Fusion 降低了 16.11%、31.28%和 12.62%。在部分遮挡条件下，本文算法的缺失率比 MBNet 降低了 1.31%；在严重遮挡条件下，本文算法的缺失率仅与 AR-CNN^[19]相差 0.32%。

本文算法与其他先进算法在 KAIST 数据集上的 FPPI-MR 性能如图 5 所示。观察曲线可知，本文算法在全天、白天、夜间的情况下的缺失率分别达到了 8.76%、8.76%、8.71%，均优于绝大部分主流算法，充分证明了本文算法的特征融合机制有效获取了双通道模态的互补信息，使得网络对于白天行人和夜间行人检测的适应性强，网络在不同光照条件下也有较强的鲁棒性。此外，在 DMAF 模块和 DFA 机制的共同作用下，本文算法对行人特征的捕捉能力较强，在白天和夜间光照环境变化较大的情况下，算法依然能够比较全面、有效地提取行人特征。

(2) 检测速度对比实验

将本文算法与 Faster R-CNN 和 MBNet 算法在相同平台进行检测速度比较，实验结果见表 5。由表 5 可知，本文算法的检测速度是 Faster R-CNN^[15]的 2 倍多，其根本区别在于 Faster R-CNN 是基于锚框的检测机制，而本文算法抛弃了锚框，采用关键点预测的方法，舍去了非极大值抑制^[23]等后处理操作，在保证精度的同时也极大地提高了模型的检测速度。此外，本文算法的检测速度比 MBNet 提升了 57.14%，主要原因在于其抛弃了 MBNet 的 IAF 模块和基于锚框的检测方法，在权衡检测精度的基础上大大提升了检测速度。

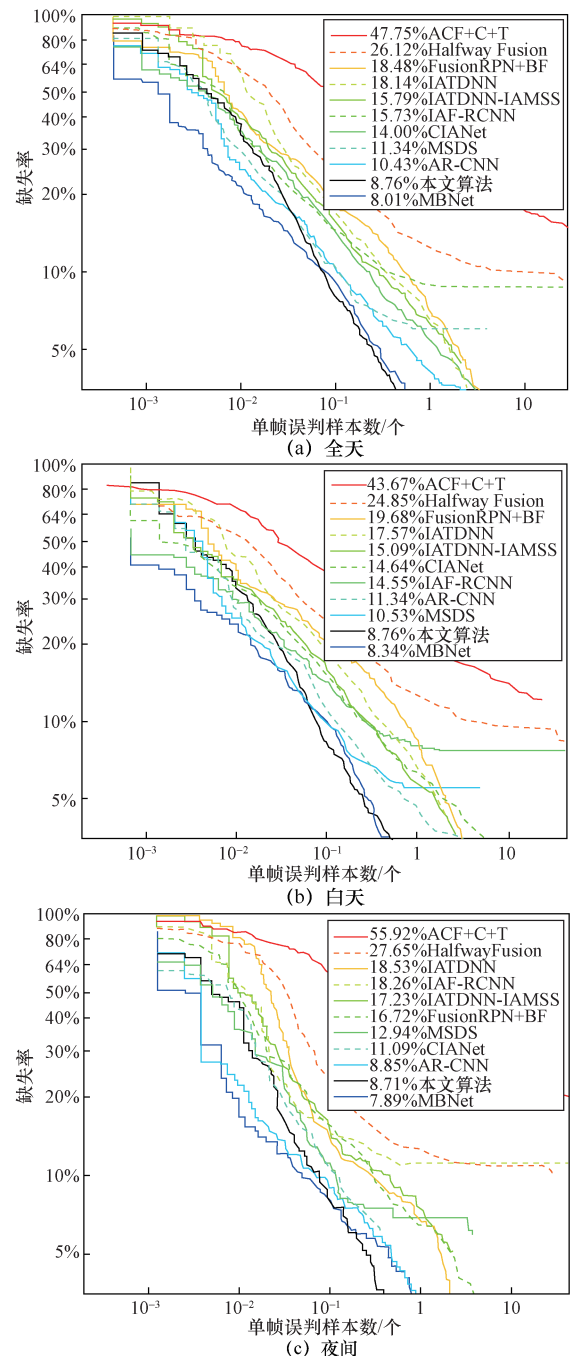


图5 本文算法与其他先进算法 KAIST 数据集上的 FPPI-MR 性能

表5 检测速度对比实验

方法	检测速度/帧每秒
Faster R-CNN ^[15]	10
MBNet ^[5]	14
本文算法	22

(3) 在 CVC-14 和 FLIR 数据集上的实验

本文算法在 CVC-14 和 FLIR 数据集上的检测结果见表 6。在 CVC-14 数据集严重不对齐和 FLIR 训练样本较少的情况下, 本文算法仍然达到了 25.54% 和 28.43% 的缺失率, 这充分说明了本文算法的鲁棒性和泛化能力较强, 具有较好的实际应用价值。

表6 本文算法在 CVC-14 和 FLIR 数据集上的检测结果

数据集	MR ⁻²
CVC-14	25.54%
FLIR	28.43%

(4) 检测结果可视化

本文算法在 KAIST、CVC-14、FLIR 数据集上的部分检测结果如图 6 所示。其中绿色框是真实目标框, 红色框是检测框。第一列和第二列分别是夜间可见光和红外图像, 第三列和第四列分别是白天可见光与红外图像。由图 6 可知, 无论是小尺寸行

人还是大尺寸行人都能被本文算法正确检测, 本文算法在这 3 种数据集上都取得了比较不错的检测效果, 同时也证明了本文提出的算法具有良好的泛化能力, 可被应用在实际交通检测领域。

4 结束语

当前主流的多光谱行人检测算法采用暴力特征融合和锚框处理机制, 存在特征融合质量较差、超参数众多和后处理步骤复杂等问题, 极大地增加了计算复杂度, 从而影响实时检测性能。本文提出的基于差分特征注意力机制的无锚框多光谱行人检测算法可以有效解决以上问题。利用 CenterNet 检测模型可以有效改善基于锚框算法的缺点, 提高检测速度。基于差分模态感知融合模块的特征融合与基于差分特征注意力机制的特征优化模块都可以明显提升算法的检测精度。本文算法在保证检测精度的同时极大地提升了检测速度, 具有较好的应用前景。

参考文献:

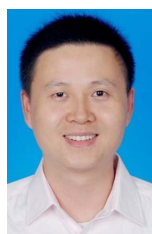
- [1] 郑南宁. 人工智能新时代[J]. 智能科学与技术学报, 2019, 1(1): 1-3.
ZHENG N N. The new era of artificial intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 1-3.
- [2] 张钊. 人工智能进入后深度学习时代[J]. 智能科学与技术学报, 2019, 1(1): 4-6.
ZHANG B. Artificial intelligence is entering the post deep-learning



图6 本文算法在 KAIST、CVC-14、FLIR 数据集上的部分检测结果

- era[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 4-6.
- [3] GRASSI A P, FROLOV V, LEÓN F P. Information fusion to detect and classify pedestrians using invariant features[J]. Information Fusion, 2011, 12(4): 284-292.
 - [4] 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. Piscataway: IEEE Press, 2015: 1037-1045.
 - [5] ZHOU K L, CHEN L S, CAO X. Improving multispectral pedestrian detection by addressing modality imbalance problems[M]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 787-803.
 - [6] JÖRG W, VOLKER F, MICHAEL H, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks[C]// Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. [S.l.:s.n.], 2016.
 - [7] LIU J J, ZHANG S T, WANG S, et al. Multispectral deep neural networks for pedestrian detection[C]//Proceedings of the British Machine Vision Conference. [S.l.:s.n.], 2016: 731-733.
 - [8] VANDERSTEEGEN M, BEECK K, GOEDEME T. Real-time multispectral pedestrian detection with a single-pass deep neural network[C]//Proceedings of the International Conference on Image Analysis and Recognition. [S.l.:s.n.], 2018: 419-426.
 - [9] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6517-6525.
 - [10] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.
 - [11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014: 113-123.
 - [12] LIN M, CHEN Q, YAN S C. Network in network[J]. Computer Science, 2013: 211-223.
 - [13] SHEN J F, ZUO X, YANG W K, et al. Differential features for pedestrian detection: a Taylor series perspective[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(8): 2913-2922.
 - [14] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[M]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.
 - [15] REN S Q, HE K M, 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, 2017, 39(6): 1137-1149.
 - [16] ZHOU X Y, WANG D Q, PHILIPP K. Objects as points[J]. Communications and Computer Sciences, 2019: 232-239.
 - [17] 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. Piscataway: IEEE Press, 2017: 2999-3007.
 - [18] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 764-773.
 - [19] ZHANG L, ZHU X Y, CHEN X Y, et al. Weakly aligned cross-modal learning for multispectral pedestrian detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 5126-5136.
 - [20] KÖNIG D, ADAM M, JARVERS C, et al. Fully convolutional region proposal networks for multispectral person detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2017: 243-250.
 - [21] LI C Y, SONG D, TONG R F, et al. Illumination-aware Faster R-CNN for robust multispectral pedestrian detection[J]. Pattern Recognition, 2019, 85: 161-171.
 - [22] GUAN D Y, CAO Y P, YANG J X, et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection[J]. Information Fusion, 2019, 50: 148-157.
 - [23] LIU S T, HUANG D, WANG Y H. Adaptive NMS: refining pedestrian detection in a crowd[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 6452-6461.

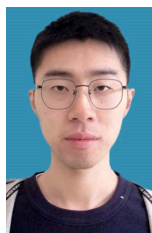
[作者简介]



沈继锋(1980—), 男, 博士, 江苏大学电气信息工程学院副教授, 主要研究方向为计算机视觉与模式识别。



刘岳(1994—), 男, 江苏大学电气信息工程学院硕士生, 主要研究方向为计算机视觉。



韦浩(1995—), 男, 就职于中电海康集团有限公司, 主要研究方向为计算机视觉。



左欣(1980—), 女, 博士, 江苏科技大学计算机学院副教授, 主要研究方向为计算机视觉与模式识别。



杨万扣(1979—), 男, 博士, 东南大学自动化学院副研究员、博士生导师, 主要研究方向为计算机视觉、模式识别与机器学习。