引用格式：SUN Ying, HOU Zhiqiang, YANG Chen, et al. Object Detection Algorithm Based on Dual-modal Fusion Network [J]. Acta Photonica Sinica, 2023, 52(1):0110002
孙颖,侯志强,杨晨,等. 基于双模态融合网络的目标检测算法[J].光子学报,2023,52(1):0110002

基于双模态融合网络的目标检测算法

孙颖[1,2],侯志强[1,2],杨晨[1,2],马素刚[1,2],范九伦[1]
(1 西安邮电大学 计算机学院, 西安 710121)
(2 陕西省网络数据分析与智能处理重点实验室, 西安 710121)

摘　要：针对红外图像和可见光图像的融合目标检测问题,提出一种基于双模态融合网络的目标检测算法。在同时输入红外和可见光图像对后,利用设计的红外编码器提取红外图像空间特征信息;通过设计的可见光编码器将可见光图像从垂直和水平两个空间方向聚合特征,通过精确的位置信息对通道关系进行编码;最后,采用提出的门控融合网络自适应调节两路特征的权重分配,实现跨模态特征融合。在KAIST行人数据集上,与基准算法YOLOv5-n单独检测可见光图像和红外图像的结果相比,所提算法检测精度分别提升15.1%和2.8%;与基准算法YOLOv5-s相比,检测精度分别提升14.7%和3%;同时,检测速度在两个不同基准算法模型上分别达到117.6 FPS和102 FPS。在自建的GIR数据集上,所提算法的检测精度和速度也同样具有明显优势。此外,该算法还能对单独输入的可见光或红外图像进行目标检测,且检测性能与基准算法相比有明显提升。

关键词：目标检测;门控网络;早期融合;双模态;编码器
中图分类号：TP391.41　　**文献标识码**：A　　**doi**：10.3788/gzxb20235201.0110002

0 引言

目标检测是计算机视觉领域的基本任务之一,在安防监控、自动驾驶和军事领域等都有重要的应用[1-2]。近年来,基于深度学习的目标检测已经成为主流,根据检测步骤的不同,算法可以分为两类:两阶段检测算法和单阶段检测算法。两阶段检测算法首先对图像提取候选框,然后基于候选区域进行修正得到检测结果,如区域卷积神经网络(Region-based Convolutional Neural Network,R-CNN)[3]、Fast-RCNN[4]和Faster-RCNN[5]等。单阶段检测算法直接对图像生成检测结果,如Single Shot Multibox Detector (SSD)[6]和You Only Look Once (YOLO)系列[7-10]等。其中,YOLOv5在检测精度和速度上具有明显的优势。此外,无锚框(Anchor-free)的检测算法近年来也逐渐兴起,如CornerNet[11]、CenterNet[12]和Fully Convolutional One-Stage (FCOS)[13]等,该类算法不用锚框,将目标检测转化为关键点的定位组合问题,摆脱了使用锚框而带来的计算量,提高检测实时性。

目前,目标检测算法主要用与检测任务相关的单模态图像作为训练数据[14-15],然而仅用单模态图像在实际的复杂场景中有时很难检测到目标。为了解决上述问题,许多研究者提出了用多模态图像作为训练数据的方法[16-17]。多模态图像,例如红外图像和可见光图像,具有互补优势。红外图像的优点是依赖目标物体产生的热源,不受照明条件的影响,但无法捕捉到目标的细节信息。可见光图像的优点是能清晰地捕捉目标的纹理特征和细节信息,但容易受到光照条件的影响。因此,基于多模态目标检测研究已成为当前的研究热点[18-23]。DEVAGUPTAPU C等[18]利用CycleGAN创建合成红外图像的策略,并利用照明感知融合框架融合可见光和红外图像,以提升红外图像中目标检测的性能;YANG L等[19]采用照明子网分配权重,联合可

基金项目：国家自然科学基金(No.62072370)
第一作者：孙颖, aurorasuny@163.com
通讯作者：侯志强, hou-zhq@sohu.com
收稿日期：2022-07-13;**录用日期**：2022-08-02

见光和红外图像检测行人；赵明等[20]提出了一种跨域融合网络结构，采用红外域和伪可见光域双通道的多尺度特征金字塔获取每个模态的特征图，对多尺度特征进行双模态特征融合；WANG Q等[21]设计了一种冗余信息抑制网络，能抑制跨模态冗余信息，有助于融合可见光和红外的互补信息；GENG X等[22]通过信道信息交换实现跨模态人员再识别和利用可见光红外双摄像头跟踪；周涛等[23]设计了三编码器提取多模态医学图像特征，解决单模态图像提取能力不足的问题。

对于流行的卷积神经网络，多模态图像特征融合通常有三种融合方式：早期融合、中期融合和后期融合。早期融合是指在第一个卷积层后，将多个分支特征映射融合，ZHANG Y等[24]提出一种基于注意力的多层融合网络用于多光谱行人检测；CAO Z等[25]利用多光谱通道特征融合模块，根据照明条件融合可见光和红外图像特征。中期融合是指在骨干网络提取特征后进行融合，KONIG D等[26]利用新的多光谱区域建议网络（Region Proposal Networks，RPN）来有效融合多光谱图像中的红外和可见光信息；FU L等[27]提出一种自适应空间像素级特征融合网络，可以自适应地从红外和可见光图像中提取特征进行融合。后期融合即决策级融合，执行结果的融合，WAGNER J等[28]提出两个融合架构并分析其在多光谱数据上的性能；白玉等[29]利用两种检测结果决策出最优结果，提出一种基于决策级融合的目标检测算法。但这些方法没有分别提取多模态图像的特征，且融合多模态特征时并不充分，还需要进一步提高多模态特征信息之间的互补优势。

为此，本文提出一种基于双模态融合网络的目标检测算法，采用早期融合，并嵌入了一个门控融合网络，使模型能够确定两种模态图像在不同场景中对检测的贡献。首先，由于红外图像中的特征信息较少，采用设计的红外编码器中对空间信息进行编码；其次，设计了一个可见光编码器，从垂直和水平两个空间方向聚合特征，通过精确的位置信息对通道关系进行编码；最后，引入多任务学习的思想，提出门控融合网络，自适应调节权重分配，实现跨模态特征融合。并在公开数据集和自建数据集上进行验证。

1 本文算法

提出的基于双模态融合网络的目标检测算法框架如图1所示，整个算法由双模态编码器（Dual Mode Encoder）、门控融合网络（Gated Fusion Network）和检测器（Detector）三部分组成。双模态编码器由两部分构成，分别是红外编码器（Encoder-Infrared，EIR）和可见光编码器（Encoder-Visible，EVS），其中，红外图像X_{IR}通过红外编码器EIR获取特征F_{IR}，可见光图像X_{VS}通过可见光编码器EVS得到特征F_{VS}。选择早期融合方式，设计门控融合网络计算红外特征F_{IR}和可见光特征F_{VS}的权重，自适应加权得到融合后的特征F_D。同时，通过两个最大池化（MaxPool）残差保留单模态图像的细节特征信息，防止因权重分配过低导致单模态部分特征信息损失。F_{IR}和F_{VS}分别通过MaxPool层获取更突出的前景特征，将特征F_D与其按通道维度拼接得到特征F_M，实现双模态特征信息的互补。选用YOLOv5作为基准检测器，将门控融合网络的输出特征F_M作为输入，输入至YOLOv5骨干网络的第二层，后续通过该检测器进行目标分类和定位。

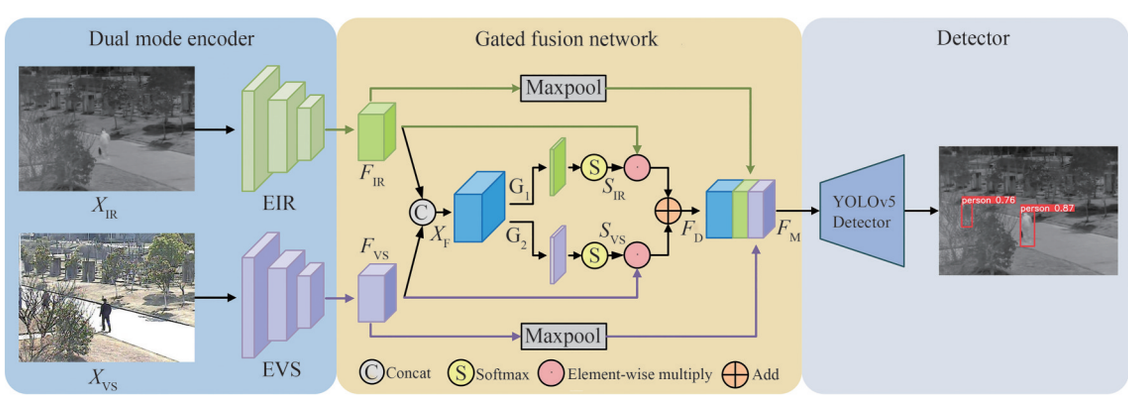

图1 算法整体框架

Fig. 1 Overall algorithm architecture

1.1 双模态编码器

由于红外图像和可见光图像的成像差异及各自的特点，使用不同的网络来提取特征。设计的双模态编码器如图2所示，由红外编码器EIR和可见光编码器EVS两部分构成。

图2左半部分是红外编码器(EIR)，输入为红外图像X_{IR}，整个编码器旨在提取红外图像的空间信息，以弥补复杂场景下可见光图像信息的缺失。迷你残差(Mini Residual)模块由MaxPool层和两个连续的1×1卷积层构成，MaxPool层能减少特征中的无用背景信息，连续两个1×1卷积层将多个特征图线性组合，实现跨通道的信息整合。构建残差，对冗余网络层的特征信息进行恒等映射，能在后续层中补充丢失的特征信息。红外图像通常细节信息不明显，但能提取到空间位置信息，在一些夜间、遮挡等场景下，可以发挥优势对目标定位。故在红外编码器EIR中引入SimAM模块[30]来提取通道中不同的局部空间信息。该模块基于神经科学理论优化了一个能量函数，从而计算出每个神经元的重要性。能量函数为特征层映射计算出特征图的三维注意权重，即考虑了空间和通道权重，由于不在网络层中添加参数，因此对速度没有影响。

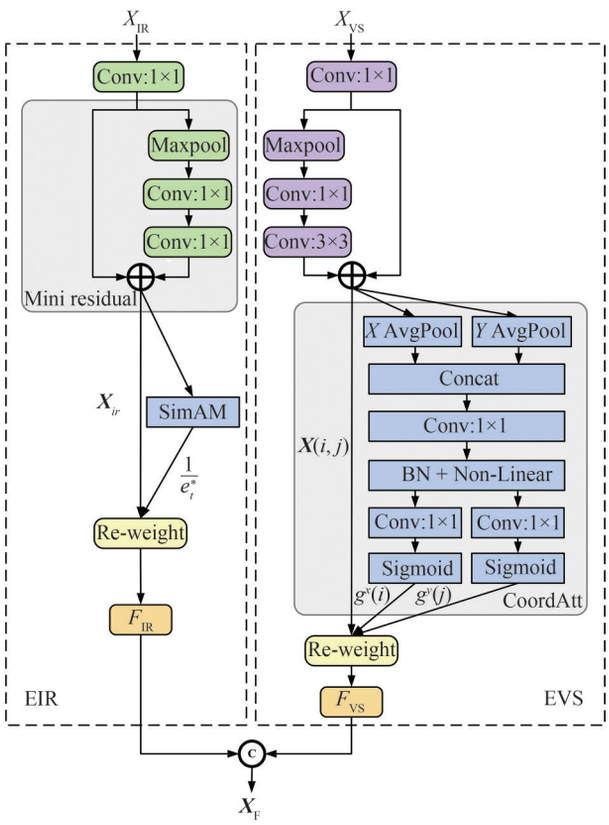

图2 双模态编码器结构

Fig. 2 Dual-mode encoder structure

能量函数e_t^*见式(1)，用于度量神经元之间的区分性，能量越低，当前神经元与周围神经元区别越大。

$$e_t^*=\frac{4(\hat{\delta}^2+\lambda)}{(t-\hat{\mu})^2+2\hat{\delta}^2+2\lambda} \tag{1}$$

式中，λ是超参数，实验中设默认值为10^{-4}，t为目标神经元单一通道输入特征，x_i是其他神经元的单一通道输入特征，$\hat{\delta}^2=\frac{1}{M}\sum_{i=1}^{M}(x_i-\hat{\mu})^2$，$\hat{\mu}=\frac{1}{M}\sum_{i=1}^{M}x_i$，$\hat{\delta}^2$和$\hat{\mu}$为能量函数因子，$M=H\times W$是某个通道上的神经元数量，$H$和$W$分别是通道的长和宽。

每个神经元的权值通过$\frac{1}{e_t^*}$获得，通过计算出的神经元权值，对红外图像特征中的信息$\boldsymbol{X}_{ir}$加权得到红外特征F_{IR}，计算过程为

$$F_{IR}=\boldsymbol{X}_{ir}\times\frac{1}{e_t^*} \tag{2}$$

图2右半部分是可见光编码器(EVS)，输入为可见光图像X_{VS}，整个编码器旨在提取可见光图像丰富的细节特征来弥补红外图像信息的缺失。迷你残差(Mini Residual)模块由MaxPool层、1×1卷积层和3×3卷积层构成，利用3×3卷积增大感受野，提取细节特征信息。然而提取的特征通道总体的平均值信息不足以代表每个单独通道的个体性，会损失丰富的局部信息，导致特征缺乏多样性。因此，在EVS中引入坐标注意力机制(Coordinate Attention，CoordAtt)[31]提取跨通道信息，获得方向和位置信息，使模型更准确地提取到目标特征信息。为了能够获得具有精确位置信息的空间特征信息，从垂直(X)和水平(Y)两个方向利用全局平均池化(AvgPool)操作分别提取特征，得到两个一维向量。沿着两个空间方向提取特征，能够获取通道之间的关系，有助于网络更准确地定位感兴趣的目标。将从X和Y两个方向提取的特征按通道维度进行拼接(Concat)，然后利用1×1卷积压缩通道。为了加快网络训练时的收敛速度，将特征通过批标准化(Batch Normalization，BN)和非线性激活函数(Non-Linear)层。再将特征平分为两部分，通过1×1卷积将通道数调整到与输入特征相同的通道数。处理后的可见光特征F_{VS}计算过程为

$$F_{VS}=\boldsymbol{X}(i,j)\times g^{x}(i)\times g^{y}(j) \tag{3}$$

式中，将原始特征信息记为$\boldsymbol{X}$，i为垂直方向，j为水平方向，g^{x}和g^{y}表示在两个空间方向上的权重。最后通过激活函数Sigmoid归一化后将特征重新赋予权重，与原始的特征逐像素相乘得到特征F_{VS}。

本文算法在双模态编码器中将两个模态图像的特征信息进行融合，尽管可见光图像对比度高且能反映真实目标环境，具有丰富的图像空间及细节特征信息，但是当有遮挡或者光照不足时，就无法观察到目标。而红外图像虽然不能清晰地观察目标，但能够显示目标的位置信息及粗略的轮廓信息。最后，将EIR和EVS提取的特征F_{IR}和F_{VS}拼接，得到输出特征$\boldsymbol{X}_{F}$。

1.2 门控融合网络

借助多任务学习中的多门混合专家思想[32]（Multi-gate Mixture of Experts，MMoE），提出门控融合网络，结构如图3所示。门控融合网络用于学习单模态特征F_{IR}和F_{VS}对检测的贡献。根据MMoE，将提取红外图像特征和提取可见光特征看作两个任务，即EIR和EVS是专家模块，EIR只用于提取红外图像的特征F_{IR}，同理，EVS仅用于提取可见光图像的特征F_{VS}，将两个专家模块的结果融合。在融合过程中，需要对某个任务有一定的偏向性，即将专家模块EIR和EVS的输出结果F_{IR}和F_{VS}映射到概率。

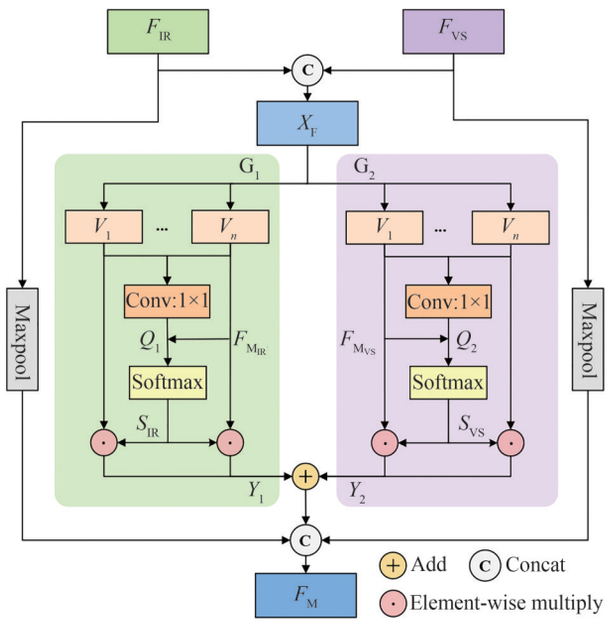

图3 门控融合网络结构

Fig. 3 Gated fusion network structure

为了充分利用红外和可见光图像特征的互补性，考虑计算两种模态特征对检测的贡献，为红外特征F_{IR}和可见光特征F_{VS}生成计算概率的门G_1和G_2。首先将F_{IR}和F_{VS}两个特征按通道维度拼接得到$\boldsymbol{X}_{F}$，以在空间方向组合它们的特征。利用两个门控模块G_1和G_2产生权重，实现自适应融合。门控网络是一个两层模块，即1×1卷积层和Softmax层。将$\boldsymbol{X}_{F}$分为两组向量，记为$F_{M_{IR}}$和$F_{M_{VS}}$，每组向量由1到n，记作$\boldsymbol{V}_{I}$到$\boldsymbol{V}_{n}$。G_1中$\boldsymbol{V}_{I}$到$\boldsymbol{V}_{n}$通过1×1卷积层削减通道数得到权重$\boldsymbol{Q}_1$，G_2中$\boldsymbol{V}_1$到$\boldsymbol{V}_{n}$通过1×1卷积层得到权重$\boldsymbol{Q}_2$。将原始向量和$\boldsymbol{Q}_{I}$、$\boldsymbol{Q}_{2}$分别逐像素相乘加权，输出由后续的Softmax层归一化得到$\boldsymbol{S}_{IR}$和$\boldsymbol{S}_{VS}$，Softmax函数在G_1和G_2两个门中的计算过程分别为

$$\boldsymbol{S}_{IR}=\frac{e^{F_{M_{IR}}\times\boldsymbol{Q}_1}}{e^{F_{M_{IR}}\times\boldsymbol{Q}_1}+e^{F_{M_{VS}}\times\boldsymbol{Q}_2}} \tag{4}$$

$$\boldsymbol{S}_{VS}=\frac{e^{F_{M_{VS}}\times\boldsymbol{Q}_2}}{e^{F_{M_{IR}}\times\boldsymbol{Q}_1}+e^{F_{M_{VS}}\times\boldsymbol{Q}_2}} \tag{5}$$

式中，$\boldsymbol{Q}_1\in\mathrm{R}^{1\times H\times W}$是分配给红外特征的权重，$\boldsymbol{Q}_2\in\mathrm{R}^{1\times H\times W}$是分配给可见光特征的权重。

归一化后的权值$\boldsymbol{S}_{IR}$和$\boldsymbol{S}_{VS}$与F_{M_I}，$i\in$IR，VS，逐元素相乘自适应加权，得到输出特征$\boldsymbol{Y}_1$和$\boldsymbol{Y}_2$，即

$$Y_1=\boldsymbol{S}_{IR}\odot F_{M_{IR}} \tag{6}$$

$$Y_2=\boldsymbol{S}_{VS}\odot F_{M_{VS}} \tag{7}$$

对红外图像和可见光图像的特征重新加权，逐像素相加得到合并的特征$\boldsymbol{Y}_1+\boldsymbol{Y}_2$，实现多模态特征融合。Softmax的可分性会使权重向量类内紧凑，类间分离，类内即同一个编码器提取的特征，类间即不同编码器提取的特征。因此，门控融合网络中的Softmax层会丢失同等重要但不同类的特征信息。为了解决这个问题，利用两个MaxPool层来弥补丢失的特征信息。MaxPool操作能够提取特征中的边缘和纹理信息并抑制背景信息，将F_{IR}和F_{VS}通过MaxPool跳连至输出处，补充丢失的特征信息。原始输入特征F_{IR}和F_{VS}通过MaxPool后，产生新的特征，与门控融合后的加权特征$\boldsymbol{Y}_1+\boldsymbol{Y}_2$按通道维度拼接（Concat）得到最终特征$F_{M}$，即

$$F_{M}=\mathrm{Concat}(Y_1+Y_2,\mathrm{MaxPool}(F_{IR}),\mathrm{MaxPool}(F_{VS})) \tag{8}$$

2 实验结果及分析

2.1 实验条件

实验的操作系统为Ubuntu 16.04，CPU为i5-8400，GPU为NVIDIA GeForce GTX 1080Ti（显存11 GB），CUDA以及CUDNN的版本为10.0和CUDNN7.4.2。实验采用Python和PyTorch平台。训练过程中使用SGD优化器对网络参数进行迭代更新，动量参数设为0.937，BatchSize设为8，共训练150个Epoch，在加载数据时将所有图像的分辨率统一调整到640×640，再对整体网络进行端到端训练。

2.2 数据集与评价指标

用公开的KAIST行人数据集[33]和自建的GIR数据集对算法进行评估。

KAIST行人数据集共有95 328张图片，每张图片都包含红外和可见光图像两个版本，图像尺寸为640×512，包括校园、街道以及乡下等常规交通场景。由于原数据集标注较差且数据集是取自视频连续帧图片，相邻图片相差不大，故进行一定程度的数据集清洗。清洗规则为：训练集每隔2张图片取一张，并去掉所有不包含任何行人的图片，可得到7 601张训练集[34]图片，4 755张白天图片，2 846张夜晚图片。测试集每隔19张取一张，保留负样本，可得到2 252张测试集[35]图片，1 455张白天图片，797张夜晚图片。原数据集的标签中包含person、people和cyclist三个类别，当照明条件差或分辨率低时，很难区分这三类，故将标签类别仅标注为person一类。

GIR数据集是自行创建的通用目标数据集。图像来源于李成龙团队建立的RGBT210数据集[36]，每张图片包含红外和可见光彩色图像两个版本，图像尺寸为630×460。从该数据集中选取5 105张图片，划分为训练图像4 084张，测试图像1 021张。对图片进行标注，确定10类目标为person、dog、car、bicycle、plant、motorcycle、umbrella、kite、toy和ball。

两种数据集每张图片均包含红外和可见光两个版本。通过对成像硬件设备捕捉的图片进行高度对齐裁剪，每个图像对是已经配准好的两张图像。对红外图像、可见光图像和两者组成的图像对分别进行训练和测试，所有类型图像共享一套标签。

实验使用的算法评价指标是平均检测精度（Average Precision，AP）、每秒帧数（Frames Per Second，FPS）、精确度（Precision，P）和召回率（Recall，R）。其中$AP_{0.50:0.95}$指IoU从0.50到0.95每隔0.05计算的所有类别的AP平均值。

2.3 消融实验与定性分析

本文算法在KAIST数据集和GIR数据集上分别进行消融实验，并与基准算法的n和s两个模型进行对比，评估提出模块对检测器的贡献，以验证所提出方法的有效性。

2.3.1 KAIST数据集

在训练之前，使用k-means++聚类标签框，得到合适的预设框尺寸。常见的输入图像尺寸有416×416、512×512、608×608、640×640，考虑到输入不同尺寸的图像对可能会影响检测器性能，使用本文算法的n和s两个模型在KAIST数据集上进行实验，结果见表1和2。从表中可以看出，输入图片的尺寸对检测模型的性能影响相当明显。在基础网络部分常常会生成比原图小数十倍的特征图，导致小目标的特征不容易被检测网络捕捉。通过输入大尺寸但不超过数据集原图大小的图片进行训练，能够在一定程度上提高检测模型对目标大小的鲁棒性。网络结构不变，网络的感受野是一定的。输入图像的分辨率提高，即尺寸小，感受野在图像中的占比会下降，导致网络提取的局部信息无法有效预测所有尺度的前景物体，从而造成检测准确率下降。因此，实验中输入图像对尺寸均为640×640。

表1 n模型上不同输入图像对尺寸的检测器性能

Table 1 Detector performance for different input image pairs sizes on n-model

Algorithm	Resolution	$AP_{0.5:0.95}$	$AP_{0.5}$
Ours-n	416×416	30.5	70
Ours-n	512×512	32.5	73.1
Ours-n	608×608	32.9	73.3
Ours-n	640×640	33.3	73.8

表 2 s模型上不同输入图像对尺寸的检测器性能

Table 2 Detector performance for different input image pairs sizes on s-model

Algorithm	Resolution	$AP_{0.5:0.95}$	$AP_{0.5}$
Ours-s	416×416	31.1	71
Ours-s	512×512	31.9	72.7
Ours-s	608×608	34.3	73.9
Ours-s	640×640	35.2	74.5

为了更好地分析编码器和门控融合网络对检测器性能的贡献，在KAIST数据集上进行消融实验，结果见表3。可以看出，在YOLOv5-n模型上，仅输入可见光图像(VS)时，检测精度为58.7%；仅输入红外图像(IR)时，检测精度为71%。用可见光编码器取代YOLOv5-n的第一层卷积，检测精度由58.7%提升到59.1%；同理，利用红外编码器替换后，检测精度由71%提升为71.3%。当输入为双模态图像时，利用可见光和红外编码器提取特征后，使用门控融合网络分配权重，检测精度达到73.8%，较基准模型单独检测可见光和红外图像分别提升了15.1%和2.8%。在YOLOv5-s模型上，仅输入可见光图像或红外图像时检测精度分别为59.8%和71.5%。加入可见光编码器检测精度提升到60.2%；加入红外编码器后，检测精度提升为71.9%。当输入为双模态图像时，检测精度达到74.5%，较基准模型单独检测可见光和红外图像分别提升了14.7%和3%。虽然所提算法较基准算法的速度稍微有所下降，但在检测精度有显著提高。

表 3 不同模型在KAIST数据集上的消融实验结果

Table 3 Ablation experimental results of different models on the KAIST dataset

Method	Encoder-VS	Encoder-IR	Gated Fusion	Input	$AP_{0.5:0.95}$	$AP_{0.5}$	FPS
YOLOv5-n				VS	24.8	58.7	158.7
YOLOv5-n				IR	31.6	71	158.7
YOLOv5-n-EVS	√			VS	25	59.1	125
YOLOv5-n-EIR		√		IR	31.8	71.3	125
Ours-n	√	√	√	VS+IR	**33.3**	**73.8**	**117.6**
YOLOv5-s				VS	26.7	59.8	112.4
YOLOv5-s				IR	32	71.5	112.4
YOLOv5-s-EVS	√			VS	26.9	60.2	107.5
YOLOv5-s-EIR		√		IR	32.2	71.9	107.5
Ours-s	√	√	√	VS+IR	**35.2**	**74.5**	**102**

图4为目标person在不同模型上的精准率与召回率(Precision-Recall，P-R)曲线。可以看出利用双模态图像互补后的特征信息算法，AP有大幅提升。

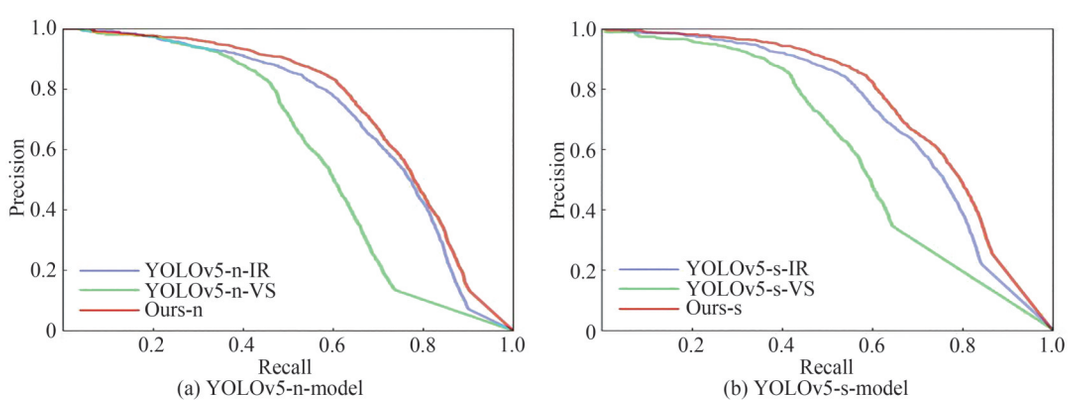

图4 不同模态输入在两种模型的P-R曲线

Fig. 4 P-R curves of the two models with different modal inputs

为更直观地比较与基准算法的检测结果，部分可视化结果如图5所示。图5(a)和(b)是在双模态图像上的真实框(Ground Truth，GT)，图5(c)是基准算法在可见光图片训练后结果，图5(d)是在红外图像上的结果，图5(e)是本文算法的可视化结果。考虑到该数据集可见光图像的背景较清晰，因此将可视化结果展示在可见光图像上。

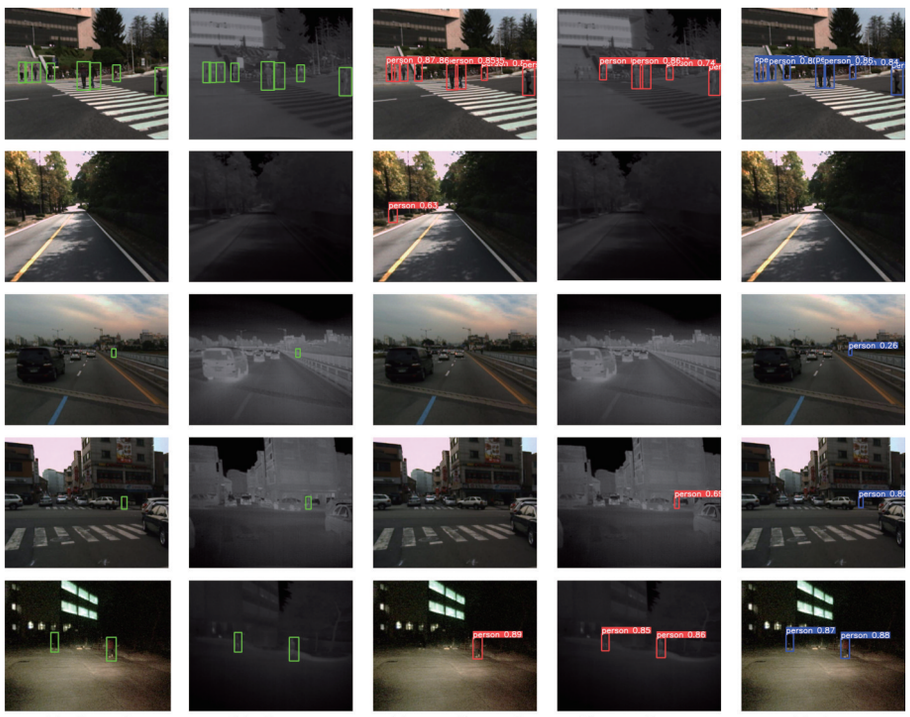

图5　在KAIST数据集上的检测结果

Fig. 5　Detection results on the KAIST dataset

第一行是白天场景，可见光图像光照充足，目标明显，相反，红外图像目标不明显，本文算法将可见光特征补充后能准确检测到目标且置信度分数较高。第二行在可见光图像上错检，通过与红外特征互补后，本文算法能够改善错检问题。第三行基准算法YOLOV5-s在两种图像上均未检测到小目标，而本文算法的s模型能准确地检测。第四行可见光图像光照不足，很难检测到目标，但红外图像可以解决此类场景的漏检问题，本文算法将红外特征补充后，能够准确地检测到目标。第五行是夜晚场景，此类场景下可见光图像处于劣势，相反，红外图像很有优势，因此，结合双模态图像的特征后，能够检测到目标且置信度分数比单模态图像高。

2.3.2　GIR数据集

同理使用k-means++聚类标签框设置预设框尺寸。在GIR数据集上进行消融实验，结果见表4。可以看出，在YOLOv5-n模型上，仅输入可见光图像时，检测精度为88.8%；仅输入红外图像时，检测精度为75.5%。用可见光编码器取代YOLOv5-n的第一层卷积，检测精度由提升到89.1%；用红外编码器取代YOLOv5-n的第一层卷积，检测精度提升为76.3%。当输入为双模态图像时，利用可见光和红外编码器提取特征后，使用门控融合网络分配权重，检测精度达到89.8%，较基准模型单独检测可见光和红外图像提升了1%和14.3%。在YOLOv5-s模型上，仅输入可见光图像或红外图像时检测精度分别为89.9%和76.8%。加入可见光编码器检测精度提升到90.1%；加入红外编码器后，检测精度提升为77%。当输入为双模态图像时，检测精度达到90.5%，较基准模型单独检测可见光和红外图像提升了0.7%和13.7%。

表5为在GIR数据集上10类目标在不同模型上的检测精度结果。可以看出在两个模型上利用双模态图像的特征信息后，算法大部分类别的检测精度有一定程度的提升，尤其是目标ball和kite的精度有大幅提升。

表 4 不同模型在GIR数据集上的消融实验结果

Table 4 Ablation experimental results of different models on the GIR dataset

Method	Encoder-VS	Encoder-IR	Gating Fusion	Input	$AP_{0.5:0.95}$	$AP_{0.5}$	FPS
YOLOv5-n				VS	48.4	88.8	158.7
YOLOv5-n				IR	36.3	75.5	158.7
YOLOv5-n-EVS	√			VS	49.4	89.1	105.3
YOLOv5-n-EIR		√		IR	36.4	76.3	105.3
Ours-n	√	√	√	VS+IR	**49.7**	**89.8**	**101**
YOLOv5-s				VS	51.4	89.9	111.1
YOLOv5-s				IR	36.6	76.8	111.1
YOLOv5-s-EVS	√			VS	51.9	90.1	91.7
YOLOv5-s-EIR		√		IR	36.7	77	91.7
Ours-s	√	√	√	VS+IR	**52.2**	**90.5**	**85.5**

表 5 所提算法和基准算法各类的检测精度($AP_{0.5}$%)

Table 5 The detection accuracy of the proposed algorithm and the baseline algorithm ($AP_{0.5}$%)

Class	Ours-n	YOLOv5-n-VS	YOLOv5-n-IR	Ours-s	YOLOv5-s-VS	YOLOv5-s-IR
Person	90.7	**91.2**	84.0	**91.7**	91.7	85.4
Dog	**99.5**	99.5	99.5	**99.5**	99.5	91.6
Car	**95.4**	95.2	94.3	**95.8**	95.1	94.7
Bicycle	80.4	**83.7**	70.7	80.8	**84.7**	72.8
Plant	**85.6**	84.5	79.1	86.4	**87.0**	76.0
Motorcycle	**82.8**	82.0	76.1	**83.9**	82.4	77.7
Umbrella	86.0	**87.8**	70.5	85.7	**86.6**	76.1
Kite	**93.6**	82.9	64.6	**94.4**	89.2	67.6
Toy	95.6	**96.3**	86.7	96.4	**97.0**	83.7
Ball	**88.5**	84.7	29.5	**90.7**	85.5	42.1

为更直观地比较与基准算法的检测结果，部分可视化结果如图6所示。本文算法具有灵活性，在该数据集上的可视化结果将展示在红外图像上。第一行，可见光图像误检，相反，红外图像检测准确，本文算法也能准确检测到目标且置信度分数比红外图像高。第二行可见光图像在夜间场景错检，通过与红外特征互补后，本文算法能够改善错检问题。第三行可见光能在光线充足时检测到球，但红外图像未能检测到球，双模

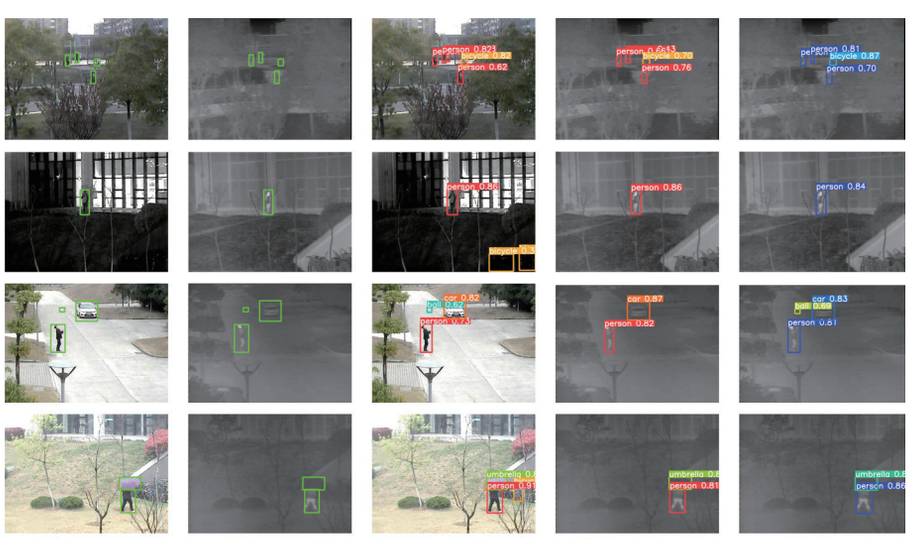

图 6 在GIR数据集上的检测结果

Fig. 6 Detection results on the GIR dataset

态图像特征互补后，本文算法能准确地检测到目标球。第四行可见光错检，但红外图像可以解决此类场景的错检问题，本文算法能够准确地检测到目标且置信度较高。

2.4 对比实验与定量分析

将本文算法运用于KAIST数据集和GIR数据集，并利用目标检测常用指标，如$AP_{0.5:0.95}$、$AP_{0.5}$等，评估所提网络。最后与近几年具有代表性的目标检测算法比较，验证本文方法的有效性。

2.4.1 KAIST数据集

表6给出本文算法与ATSS、YOLOX等部分主流通用目标检测算法的对比结果。与原基准算法相比，本文算法在检测精度上取得一定优势。与部分经典的双模态检测算法对比，如MMTOD[18]、CMDet[37]、RISNet[38]，本文算法有较高的检测精度和较快的检测速度。改进后的算法既可以单独训练红外或可见光图像，也能同时训练红外和可见光图像对，且效果均优于原算法。

表6 在KSIAT数据集上的对比实验结果

Table 6 Comparative experimental results on the KSIAT dataset

Input	Algorithm	Backbone	Resolution	$AP_{0.5:0.95}$	$AP_{0.5}$	FPS
VS	Faster R-CNN (2015)	ResNet-50	1 000×600	24.2	58.3	15.2
	SSD (2016)	VGG-16	512×512	18.1	48.2	38.1
	RetinaNet (2017)	ResNet-50	1 333×800	22.5	57.7	16.6
	YOLOv3 (2018)	DarkNet-53	416×416	18.3	46.7	56.2
	FCOS (2019)	ResNet-50	1 333×800	22.7	56.7	18.3
	ATSS (2020)	ResNet-50	1 333×800	24.3	57.8	17
	YOLOv4 (2020)	CSPDarkNet-53	416×416	23.7	57.4	55
	YOLOX-s (2021)	Modified CSP v5	416×416	27	61.1	48.4
	YOLOX-m (2021)	Modified CSP v5	416×416	27.7	61.8	40.3
	YOLOF (2021)	ResNet-50	1 333×800	22.2	54.1	25.7
	YOLOv5-n (2020)	Modified CSP v5	640×640	24.8	58.7	158.7
	YOLOv5-s (2020)	Modified CSP v5	640×640	26.4	59.8	112.4
	YOLOv5-n-EVS	**Modified CSP v5**	**640×640**	**25**	**59.1**	**125**
	YOLOv5-s-EVS	**Modified CSP v5**	**640×640**	**26.9**	**60.2**	**107.5**
IR	Faster R-CNN (2015)	ResNet-50	1 000×600	28.8	68.6	12
	SSD (2016)	VGG-16	512×512	23.2	60.9	34
	RetinaNet (2017)	ResNet-50	1 333×800	27.8	68.2	14.1
	YOLOv3 (2018)	DarkNet-53	416×416	25.3	63.6	37
	FCOS (2019)	ResNet-50	1 333×800	29.6	69.4	14
	ATSS (2020)	ResNet-50	1 333×800	29	69	13.8
	YOLOv4 (2020)	CSPDarkNet-53	416×416	27.4	68.5	52.6
	YOLOX-s (2021)	Modified CSP v5	416×416	32.8	72.1	45
	YOLOX-m (2021)	Modified CSP v5	416×416	33.5	73.1	40
	YOLOF (2021)	ResNet-50	1 333×800	27.3	65.6	25
	YOLOv5-n (2020)	Modified CSP v5	640×640	31.6	71	158.7
	YOLOv5-s (2020)	Modified CSP v5	640×640	32	71.5	112.4
	YOLOv5-n-EIR	**Modified CSP v5**	**640×640**	**31.8**	**71.3**	**125**
	YOLOv5-s-EIR	**Modified CSP v5**	**640×640**	**32.2**	**71.9**	**107.5**
VS+IR	MMTOD(2019)[18]	ResNet-101	1 000×600	31.1	70.7	13.2
	CMDet(2021)[37]	ResNet-101	640×512	28.3	68.4	25.3
	RISNet(2022)[38]	DarkNet-53	416×416	33.1	72.7	23
	Ours-n	**Modified CSP v5**	**640×640**	**33.3**	**73.8**	**117.6**
	Ours-s	**Modified CSP v5**	**640×640**	**35.2**	**74.5**	**102**

2.4.2 GIR数据集

表7给出本文算法与YOLOv3、YOLOv4、YOLOv5、YOLOX等部分主流目标检测算法的对比结果。与部分经典的双模态检测算法相比，本文算法在检测精度和速度上均有一定优势。

表7 在GIR数据集上的对比实验结果

Table 7 Comparative experimental results on the GIR dataset

Input	Algorithm	Backbone	Resolution	$AP_{0.5:0.95}$	$AP_{0.5}$	FPS
VS	YOLOv3 (2018)	DarkNet-53	416×416	41.2	85.7	50
	FCOS (2019)	ResNet-50	1 333×800	40.4	84	16
	ATSS (2020)	ResNet-50	1 333×800	47.1	87.1	14
	YOLOv4 (2020)	CSPDarkNet-53	416×416	44.5	87.9	53
	YOLOX-s (2021)	Modified CSP v5	416×416	51.7	90.3	52
	YOLOv5-n (2020)	Modified CSP v5	640×640	48.4	88.8	158.7
	YOLOv5-s (2020)	Modified CSP v5	640×640	51.4	89.8	111.1
	YOLOv5-n-EVS	**Modified CSP v5**	**640×640**	**49.4**	**89.1**	**105.3**
	YOLOv5-s-EVS	**Modified CSP v5**	**640×640**	**51.9**	**90.1**	**91.7**
IR	YOLOv3 (2018)	DarkNet-53	416×416	35.6	74.2	48.4
	FCOS (2019)	ResNet-50	1 333×800	34.5	72.3	12
	ATSS (2020)	ResNet-50	1 333×800	35.2	73.4	11.7
	YOLOv4 (2020)	CSPDarkNet-53	416×416	35.8	74.7	49
	YOLOX-s (2021)	Modified CSP v5	416×416	36.9	76.3	53
	YOLOv5-n (2020)	Modified CSP v5	640×640	36.3	75.5	158.7
	YOLOv5-s (2020)	Modified CSP v5	640×640	36.6	76.8	111.1
	YOLOv5-n-EIR	**Modified CSP v5**	**640×640**	**36.4**	**76.3**	**105.3**
	YOLOv5-s-EIR	**Modified CSP v5**	**640×640**	**36.7**	**77**	**91.7**
VS+IR	MMTOD (2019)[18]	ResNet-101	1 000×600	40.7	84.3	11.2
	CMDet (2021)[37]	ResNet-101	640×512	48.6	88.9	22.7
	RISNet (2022)[38]	DarkNet-53	416×416	49.3	89.2	23.3
	Ours-n	**Modified CSP v5**	**640×640**	**49.7**	**89.8**	**101**
	Ours-s	**Modified CSP v5**	**640×640**	**52.2**	**90.5**	**85.5**

3 结论

本文提出了一种基于双模态融合网络的目标检测算法，有效结合了红外和可见光图像特征互补的优势，在白天或夜间条件下都能较准确地检测到目标。

算法能对同时输入的红外和可见光图像对进行目标检测，其中，由于红外图像中的特征信息较少，对红外图像采用红外编码器中的迷你残差块对空间信息进行编码，增强了目标在复杂背景中的空间信息；对可见光图像采用设计的可见光编码器从垂直和水平两个空间方向聚合特征，通过精确的位置信息对通道关系进行建模。引入多任务学习的思想，提出门控融合网络，计算不同模态的特征对检测的贡献概率，自适应调节两路特征的权重分配，实现跨模态特征融合。

该算法在KAIST和GIR数据集上的检测效果显著。在KAIST行人数据集上，与基准算法YOLOv5-n单独检测可见光图像和红外图像的结果相比，所提算法检测精度分别提升15.1%和2.8%；与基准算法YOLOv5-s相比，检测精度分别提升14.7%和3%；同时，检测速度在两个不同基准算法模型上分别达到117.6 FPS和102 FPS。在自建的GIR数据集上，所提算法的检测精度和速度也同样具有明显优势。与多个经典的双模态目标检测算法实验对比结果验证了本文的方法具有较好的检测性能，同时具有良好的实时性和鲁棒性。此外，所提算法还能对单独输入的可见光或红外图像进行目标检测，且检测性能与基准算法相比有明显提升。

参考文献

[1] ZHOU Y, TUZEL O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 4490-4499.

[2] KIM S, SONG W J, KIM S H. Infrared variation optimized deep convolutional neural network for robust automatic ground target recognition[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1-8.

[3] GIRSHICK R, DONAHUE J, DARRELL T. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014: 580-587.

[4] GIRSHICK R. Fast R-CNN[C]. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 1440-1448.

[5] REN S, HE K, GIRSHICK R. Faster R-CNN: towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 91-99.

[6] LIU W, ANGUELOV D, ERHAN D. Ssd: single shot multibox detector[C]. 2016 European Conference on Computer Vision (ECCV), 2016: 21-37.

[7] REDMON J, DIVVALA S, GIRSHICK R. You only look once: unified, real-time object detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 779-788.

[8] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 7263-7271.

[9] REDMON J, FARHADI A. Yolov3: an incremental improvement[J]. arXiv preprint arXiv: 1804.02767, 2018.

[10] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of object detection[J]. arXiv preprint arXiv: 2004.10934, 2020.

[11] LAW H, DENG J. Cornernet: detecting objects as paired keypoints[C]. 2018 European Conference on Computer Vision (ECCV), 2018: 734-750.

[12] ZHOU X, WANG D, KRÄHENBÜHL P. Objects as points[J]. arXiv preprint arXiv: 1904.07850, 2019.

[13] TIAN Z, SHEN C, CHEN H. Fcos: fully convolutional one-stage object detection[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9627-9636.

[14] ZHAO F, WEI R, CHAO Y, et al. Infrared bird target detection based on temporal variation filtering and a gaussian heatmap perception network[J]. Applied Sciences, 2022, 12(11): 5679-5694.

[15] ZHU K, XU C, WEI Y, et al. Fast-PLDN: fast power line detection network[J]. Journal of Real-Time Image Processing, 2022, 19(1): 3-13.

[16] XU H, WANG X, MA J. DRF: Disentangled representation for visible and infrared image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13.

[17] YAO X, ZHAO S, XU P, et al. Multi-source domain adaptation for object detection[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 3273-3282.

[18] DEVAGUPTAPU C, AKOLEKAR N, SHARMA MM, et al. Borrow from anywhere: pseudo multi-modal object detection in thermal imagery[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019: 1029-1038.

[19] YANG L, MA R, ZAKHOR A. Drone object detection using RGB/IR fusion[J]. arXiv preprint arXiv: 2201.03786, 2022.

[20] ZHAO Ming, ZHANG Haoran. An infrared object detection method based on cross-domain fusion network[J]. Acta Photonica Sinica, 2021, 50(11): 1110001.
赵明,张浩然. 一种基于跨域融合网络的红外目标检测方法[J]. 光子学报, 2021, 50(11): 1110001.

[21] WANG Q, CHI Y, SHEN T, et al. Improving RGB-infrared object detection by reducing cross-modality redundancy[J]. Remote Sensing, 2022, 14(9): 2020.

[22] GENG X, LI M, LIU W, et al. Person tracking by detection using dual visible-infrared cameras[J]. IEEE Internet of Things Journal, 2022, 9(22): 23241-23251.

[23] ZHOU Tao, DONG Yali, LIU Shan, et al. Cross-modality multi-encoder hybrid attention U-net for lung tumors images segmentation[J]. Acta Photonica Sinica, 2022, 51(4): 0410006.
周涛,董雅丽,刘珊,等. 用于肺部肿瘤图像分割的跨模态多编码混合注意力U-Net[J]. 光子学报, 2022, 51(4): 0410006.

[24] ZHANG Y, YIN Z, NIE L, et al. Attention based multi-layer fusion of multispectral images for pedestrian detection[J]. IEEE Access, 2020, 8: 165071-165084.

[25] CAO Z, YANG H, ZHAO J, et al. Attention fusion for one-stage multispectral pedestrian detection[J]. Sensors, 2021, 21(12): 4184-4198.

[26] KONIG D, ADAM M, JARVERS C, et al. Fully convolutional region proposal networks for multispectral person detection[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017: 49-56.

[27] FU L, GU W, AI Y, et al. Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection[J].

Infrared Physics & Technology, 2021, 116: 103770.
[28] WAGNER J, FISCHER V, HERMAN M, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks[C]. ESANN, 2016, 587: 509-514.
[29] BAI Yu, HOU Zhiqiang, LIU Xiaoyi, et al. Target detection algorithm based on decision-level fusion of visible light image and infrared image[J]. Journal of Air Force Engineering University (Natural Science Edition), 2020, 21(6): 53-59.
白玉,侯志强,刘晓义,等. 基于可见光图像和红外图像决策级融合的目标检测算法[J].空军工程大学学报(自然科学版),2020,21(6):53-59.
[30] YANG L, ZHANG R Y, LI L. Simam: a simple, parameter-free attention module for convolutional neural networks[C]. International Conference on Machine Learning, PMLR, 2021: 11863-11874.
[31] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 13713-13722.
[32] MA J, ZHAO Z, YI X, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts[C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 1930-1939.
[33] HWANG S, PARK J, KIM N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1037-1045.
[34] LI C, SONG D, TONG R. Multispectral pedestrian detection via simultaneous detection and segmentation[J]. arXiv preprint arXiv:1808.04818, 2018.
[35] LIU J, ZHANG S, WANG S, et al. Multispectral deep neural networks for pedestrian detection[J]. arXiv preprint arXiv: 1611.02644, 2016.
[36] LI C, ZHAO N, LU Y. Weighted sparse representation regularized graph learning for RGB-T object tracking[C]. Proceedings of the 25th ACM International Conference on Multimedia, 2017: 1856-1864.
[37] SUN Y, CAO B, ZHU P, et al. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning[J]. arXiv:2003.02437v2, 2021.
[38] WANG Q, CHI Y, SHEN T, et al. Improving RGB-infrared object detection by reducing cross-modality redundancy[J]. Remote Sensing, 2022, 14(9): 2020-2031.

Object Detection Algorithm Based on Dual-modal Fusion Network

SUN Ying[1,2], HOU Zhiqiang[1,2], YANG Chen[1,2], MA Sugang[1,2], FAN Jiulun[1]
(1 *School of Computer Science and Technology, Xi'an University of Posts & Telecommunications, Xi'an* 710121, *China*)
(2 *Shaanxi Provincial Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an* 710121, *China*)

Abstract: In object detection, unimodal images related to the detection task are mainly used as training data, but it is difficult to detect targets in actual complex scenes using only unimodal images. Many researchers have proposed methods using multimodal images as training data to address the above problems. Multimodal images, such as Infrared (IR) images and Visible (VS) images, have complementary advantages. The advantage of IR images is that they rely on the heat source generated by targets and are not affected by lighting conditions but cannot capture the detailed information of targets. The advantage of VS image is that it can clearly capture the texture features and details of targets, but it is easily affected by lighting conditions. Therefore, for the object detection problem of IR and VS image fusion, an object detection algorithm based on the dual-modal fusion network is proposed. The algorithm can input IR images and VS images at the same time. Due to the imaging differences and their respective characteristics of IE images and VS images, different networks are used to extract features. Among them, IR images use the designed infrared encoder, Encoder-Infrared (EIR), and the SimAM module is introduced into the EIR to extract different local spatial information in the channel. This module optimizes an energy function based on neuroscience theory, thereby calculating the importance of each neuron and extracting spatial feature information by weighting. VS images adopt the designed visible encoder, Encoder-Visible (EVS), and introduce the Coordinate Attention (CoordAtt) mechanism into the EVS to extract cross-channel information, obtain orientation and position information, and enable the model to more accurately extract feature information of the target. To obtain spatial feature information with precise location information,

the global average pooling (AvgPool) operation is used to extract features from both vertical and horizontal directions, respectively, and aggregate features from both vertical and horizontal spatial directions. The precise location information encodes the channel relationship. Finally, this paper proposes a gated fusion network with the help of Multi-gate Mixture of Experts (MMoE) in multi-task learning. A gated fusion network is used to learn unimodal features and contributions to detection. According to MMoE, the extraction of infrared image features and the extraction of visible light features are regarded as two tasks, that is, EIR and EVS are expert modules, and EIR is only used to extract features of IR images. Similarly, EVS is only used to extract features of VS images. The results of the two expert modules are integrated to achieve the purpose of specialization in the surgical industry. In the process of fusion, it is necessary to have a certain bias for a certain task, that is, the output results of the expert modules EIR and EVS are mapped to the probability. The proposed gated fusion network adapts the weight distribution of the two-way features to achieve cross-modal feature fusion. This paper uses two datasets to evaluate the algorithm, the first is the public KAIST pedestrian dataset, and the second is the self-built GIR dataset. Each image in the KAIST pedestrian dataset contains both VS and IR versions. The dataset captures routine traffic scenes, including campus, street, and countryside during daytime and nighttime. The GIR dataset is general targets dataset created in this paper. These images are from the RGBT210 dataset established by Chenglong Li's team, and each image contains two versions of the VS image and the IR image. Among them, all types of images share a set of labels. On the KAIST pedestrian dataset, we validate our algorithm on two models of YOLOv5. Compared with YOLOv5-n, the detection accuracy of the proposed algorithm is improved by 15.1% and 2.8% respectively for VS and IR images; compared with YOLOv5-s, the detection accuracy is improved by 14.7% and 3%; at the same time, the detection speed reaches 117.6 FPS and 102 FPS respectively on two different models. On the self-built GIR dataset, compared with YOLOv5-n, the detection accuracy of the proposed algorithm is improved by 14.3% and 1% respectively for IR and VS images; compared with YOLOv5-s, the detection accuracy is improved by 13.7% and 0.6%. At the same time, the detection speed reaches 101 FPS and 85.5 FPS respectively on two different models. The algorithm in this paper is compared with the current classical object detection algorithms, and the effect has obvious advantages. In addition, the proposed algorithm can also perform object detection on separately input VS or IR images, and the detection performance is significantly improved compared with the baseline algorithm. In the process of visualization, the algorithm in this paper can flexibly display the visualization results of detection on VS images or IR images.

Key words: Object detection; Gating network; Early fusion; Dual-model; Encoder

OCIS Codes: 100.4996; 040.1880; 040.3060

Foundation item: National Natural Science Foundation of China (No.62072370)