引用格式：AN Hao-nan, ZHAO Ming, PAN Sheng-da, *et al*. Infrared Target Detection Algorithm Based on Pseudo Multimodal Images[J]. *Acta Photonica Sinica*, 2020, **49**(8):0810002
安浩南,赵明,潘胜达,等.基于伪模态转换的红外目标融合检测算法[J].光子学报,2020,**49**(8):0810002

基于伪模态转换的红外目标融合检测算法

安浩南[1],赵明[1,2],潘胜达[1],林长青[2]
(1 上海海事大学 信息工程学院,上海 201306)
(2 中国科学院智能红外感知重点实验室,上海 200083)
(3 上海船舶尾气智能监测工程技术研究中心,上海 201306)

摘　要:为提高红外图像目标检测的精度和实时性,提出一种基于伪模态转换的红外目标融合检测算法.首先,利用双循环的生成对抗网络无需训练图像场景匹配的优势,获取红外图像所对应的伪可见光图像;然后,构建残差网络对双模态图像进行特征提取,并采取add叠加方式对特征向量进行融合,利用可见光图像丰富的语义信息来弥补红外图像目标信息的缺失,从而提高检测精度;最后,考虑到目标检测效率问题,采用YOLOv3单阶段检测网络对双模态目标进行三个尺度的预测,并利用逻辑回归模型对目标进行分类.实验结果表明,该算法能够有效地提高目标检测准确率.

关键词:红外图像;目标检测;伪模态;生成对抗网络;残差网络
中图分类号:TP391.4　　**文献标识码**:A　　**doi**:10.3788/gzxb20204908.0810002

Infrared Target Detection Algorithm Based on Pseudo Multimodal Images

AN Hao-nan[1], ZHAO Ming[1,2], PAN Sheng-da[1], LIN Chang-qing[2]
(1 *College of Information Engineering*, *Shanghai Maritime University*, *Shanghai* 201306, *China*)
(2 *Key Laboratory of Intelligent Infrared Perception*, *Chinese Academy of Sciences*, *Shanghai* 200083, *China*)
(3 *Shanghai Engineering Research Center of Ship Exhaust Intelligent Monitoring*, *Shanghai* 201306, *China*)

Abstract: In order to improve the accuracy and real-time performance of infrared image target detection, an infrared target fusion detection algorithm based on pseudo modal transformation is proposed. First, the pseudo visible image corresponding to the infrared image is obtained by using the advantage of dual cycle generation confrontation without training image scene matching; then, the residual network is constructed to extract the features of the dual-mode image, and the feature vector is fused by the add superposition method, and the rich semantic information of the visible image is used to make up for the lack of the target information of the infrared image, so as to improve detection accuracy. Finally, considering the target detection efficiency, three scales of dual-mode targets are predicted by using the YOLOv3 single-stage detection network, and the targets are classified by using the logistic expression model. Experimental results show that the algorithm can effectively improve the accuracy of target detection.

Key words: Infrared image; Target detection; Pseudo mode; Generating countermeasure network; Residual network
OCIS Codes: 100.4990; 110.3080; 040.0040; 120.1880; 150.0150

基金项目:中国科学院智能红外感知重点实验室开放课题基金(No.CAS-IIRP-04)
第一作者:安浩南(1995－),男,硕士研究生,主要研究方向为红外图像处理.Email:anhaonan915@163.com
导　　师:赵明(1984－),女,副教授,博士,主要研究方向为图像处理.Email: mingzhao@shmtu.edu.cn
收稿日期:2020-04-09;**录用日期**:2020-05-26

0 引言

随着人工智能技术的发展，目标检测作为计算机视觉的重要分支引起了广大学者们的关注.目前，基于深度学习的目标检测已经取得了较大进展.然而，大多数深度学习的应用领域是基于可见光条件下的目标检测，有关红外场景下的研究相对较少.红外成像技术以其抗干扰性强、可全天时全天候监测等优势广泛应用于军事制导、武器瞄准定位以及安全监测等领域[1-2].

传统意义上的红外目标检测主要针对目标模板或显著性区域检测开展研究.BERTOZZI M等[3]提出了一种基于概率模板的远红外图像行人检测方法.这种方法基于人比背景红外辐射更强的假设，检测精度容易受到环境辐射的影响.DAVIS J W等[4]提出了一种基于两阶段模板的大视野范围红外图像行人检测方法.为了定位潜在的行人位置，该方法提出了一种快速筛选方法和一种广义模板，然后使用AdaBoost集成分类器来测试行人所在的位置.JUNGLING K等[5]提出了一种在红外图像中基于局部特征的行人检测器，采用多个提示的组合来寻找图像中的兴趣点，使用Surf[6]来描述这些特征点，然后通过构造码本来定位目标中心.这种检测器的挑战主要是在局部特征不明显的情况下能否获得高性能.

近年来，随着神经网络的日益普及，基于深度学习的方法以其强大的特征提取能力，逐渐被应用于红外目标检测中.PENG M等[7]提出了一种用于近红外图像人脸识别的卷积神经网络(Near-Infrared Face Identification Network, NIRFaceNet).该网络只有两个特征提取模块，网络结构更加紧凑.LEE E J等[8]设计了一个由两个卷积层和两个下采样层组成的轻量级卷积神经网络(Convolutional Neural Network, CNN)，然后结合一个增强的随机森林分类器用于监测夜间移动车辆上获取的红外图像中行人的不安全行为.CHEVALIER M等[9]提出一种针对低分辨率图像分类深层结构的卷积神经网络(Low Resolution Convolutional Neural Network, LR-CNN)自动目标识别算法.RODGER I等[10]利用长波红外传感器(Long Wave Infrared Sensor, LWIR)来增强目标识别能力，对包含人、陆上车辆、直升机、飞机、无人驾驶飞行器和假警报等六类目标的高分辨率红外图像进行训练.KRIZHEVSKY A等[11]使用基于You Only Look Once(YOLO)[12]框架的迁移学习方法训练高分辨率红外图像的目标检测网络，以便在低分辨率红外图像中对行人和车辆进行分类.基于YOLO的检测方法将目标检测问题表述为回归问题，其中边界框的坐标和每个边界框的类概率同时生成，从而大幅度提升运算速度.然而，该类目标检测效果依赖于大规模数据集上训练的体系结构和模型.

与可见光像质相比，红外图像的分辨率和信噪比较差.同时红外图像的获取成本也相对较高，目前公开的红外大规模数据集相对缺乏.因此，多模态协同进行目标检测的方法[13]逐渐进入研究者的视野.WAGNER J等[14]应用聚合信道特征(Aggregate Channel Feature, ACF)和增强决策树(Boost Decision Tree, BDT)生成方案，并采用融合了可见光和红外信息的CNN对这些方案进行分类.CHOI H等[15]对可见光图像和红外图像采用两个独立的区域建议网络，并在融合的深层特征上采用支持向量回归对两个网络生成的结果进行评估.KONIG D等[16]和LIU J等[17]提出了一种多模态框架，将可见光图像信息和红外信息结合在一个速度更快的区域卷积神经网络(Regional Convolution Neural Network, RCNN)体系结构当中.然而，上述基于多模态的目标检测方法均是基于多模态的数据集场景一一对应的假设，实际应用受到诸多限制.对此，CHAITANYA D等国外学者提出热成像中的伪多模态目标检测(Pseudo Multi-modal Object Detection in Thermal Imagery, MMTOD)算法[18]，其通过将快速区域卷积神经网络(Faster Regional Convolution Neural Network, Faster-RCNN)[19]改造成双通道图像特征提取，利用可见光信息以达到更好的红外图像目标检测效果，但是该算法对于较小目标特征提取不充分，多尺度预测目标存在劣势，同时该算法运行时间过长，满足不了实时性.

针对以上问题，本文利用循环生成对抗网络[20](Cycle Global Area Network, CycleGAN)作为图像转换(Image-to-Image, I2I)模型得到红外图像所对应的伪可见光图像，两类图像通过残差网络得到各自的特征向量，构建双模态特征向量进行融合，并将该融合特征向量输入到改进的YOLOv3单阶段检测网络部分.

1 方法

图1所示是本文提出的基于伪模态转换的红外目标融合检测算法框架(Pse-model Fused Detection,

PMFD）.整个过程旨在利用可见光图像丰富的语义信息来弥补红外图像信息的缺失，以增强红外目标检测的效果.CycleGAN作为I2I框架，训练过程中无需红外与可见光图像场景一一对应，通过训练循环生成网络模型将给定的红外图像转换为伪可见光图像.为了有效提取红外图像和伪可见光图像特征，将残差网络改进为双模态特征提取结构，将双模态特征叠加成每一维度包含更多信息量的融合特征向量，并将该向量输入到YOLOv3检测网络部分，实现红外目标检测.

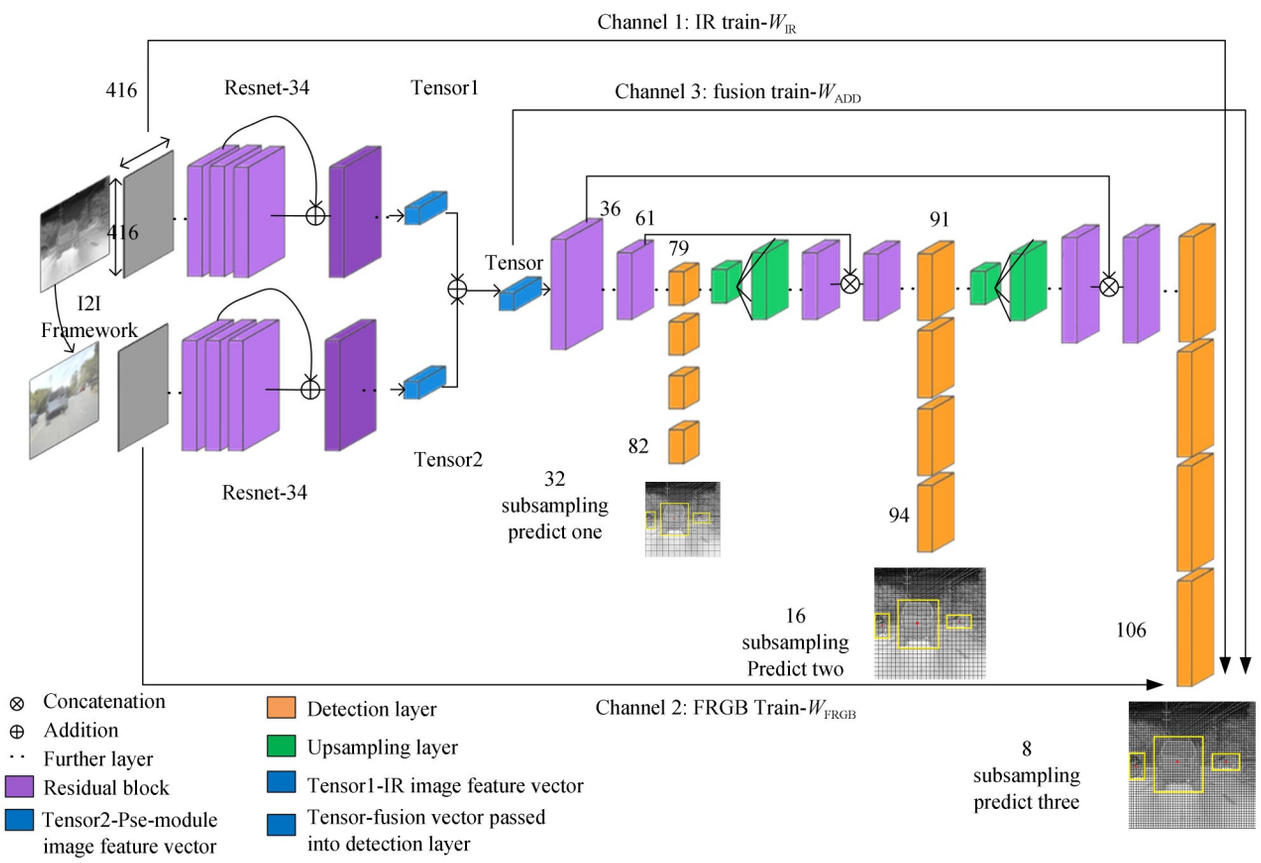


图1　提出算法的结构示意图

Fig. 1　Structure diagram of the proposed algorithm

1.1　伪可见光模态生成

CycleGAN是近年提出的一种无需匹配样本的图像到图像转换框架，旨在通过降低对抗性损失获得映射函数$F: Y \to X$和$G: X \to Y$，其中X和Y分别是源域（红外图像）和目标域（可见光图像）.如图2，函数将图像映射到两个独立的潜在空间，由两个生成器$G_{x\to y}$，$F_{y\to x}$和两个鉴别器D_x，D_y构成.生成器$G_{x\to y}$尝试生成与域Y相似的图像$\hat{y}_i$，而D_y用于区分转换的样本$\hat{y}_i$和真实样本y_i.为了减少可能的映射函数的空间，确保源

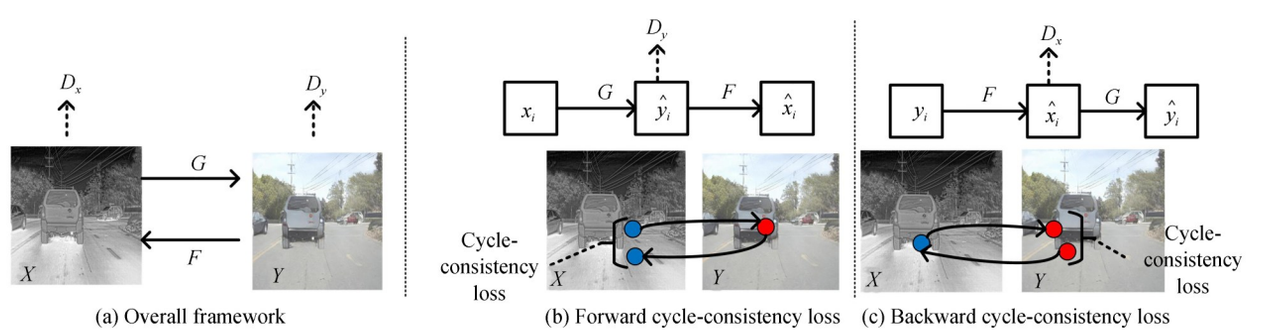


图2　CycleGAN双循环对抗生成伪可见光图像过程

Fig. 2　Process of generating pseudo visible light image using CycleGAN dual cycle countermeasure

域图像x_i在被转换成目标域($\hat{y}_i$)并重新变换回到源域($\hat{x}_i$)时$\hat{x}_i$和x_i将属于相同的分布,实施周期一致性约束.

图2(a)表示该模型包含两个映射函数$G: X \rightarrow Y$和$F: Y \rightarrow X$,以及相关的对抗鉴别器D_x和D_y.D_y支持G将X转换为与域Y不可区分的输出;同理,D_x支持F将Y转换为与域X不可区分的输出.为了进一步正则化映射关系,引入循环一致性损失必须符合以下原则:如果从一个域转换到另一个域,然后再返回,应该到达源域.相应地,图2(b)表示前循环一致性损失$x_i \rightarrow G(x_i) \rightarrow F[G(x_i)] \approx \hat{x}_i$,图2(c)则表示后循环一致性损失:$y_i \rightarrow F(y_i) \rightarrow G[F(y_i)] \approx \hat{y}_i$.

1.2 双模态特征提取与融合

残差模块由两个卷积核大小分别为1×1和3×3的卷积模块和一条捷径连接(Shortcut Connection)[21-23]构成.如图3所示,1×1卷积基于network in network[24]的思想对输入进行降维,减少参数量和计算量;3×3卷积层提取特征、恢复特征维度;捷径连接构建残差,对冗余网络层进行恒等映射,从而有效地解决随着网络深度加深时可能出现的梯度消失、梯度爆炸和网络退化等问题.

期望的底层映射用$H(\boldsymbol{x})$表示,用来叠加的非线性层为$F(\boldsymbol{x})=H(\boldsymbol{x})-\boldsymbol{x}$,则原始映射重新转换为$F(\boldsymbol{x})+\boldsymbol{x}$,那么优化残差映射要比优化原始的未参考映射更容易.$F(\boldsymbol{x})+\boldsymbol{x}$可通过前馈神经网络

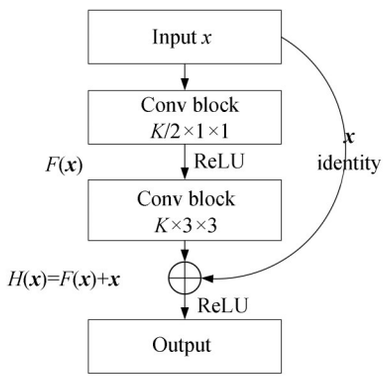


图3 双模态特征提取网络中的残差模块

Fig. 3 Residual module in bimodal feature extraction network

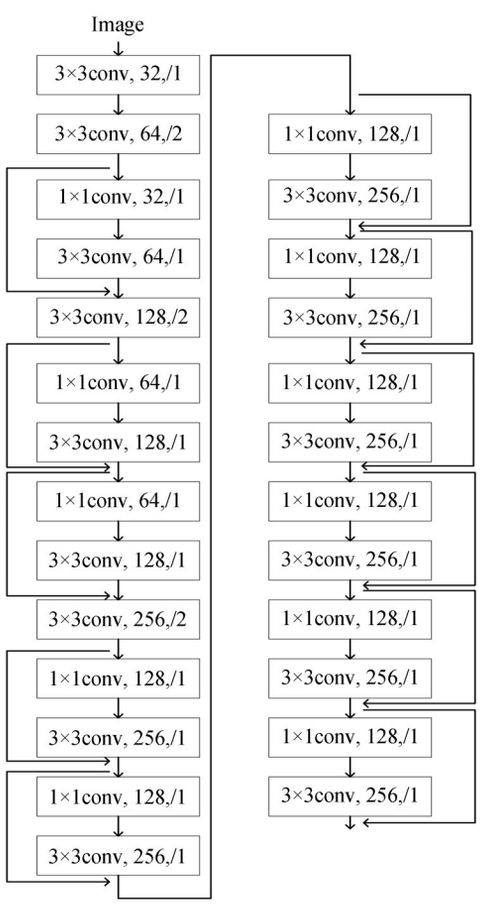


图4 改进的双模态特征提取残差网络

Fig. 4 Improved residual network of bimodal feature extraction

Shortcut Connections实现，Shortcut Connections可跨过一个或多个层的连接.重新将残差单元定义为

$$\boldsymbol{y}=F(\boldsymbol{x},\{W_i\})+\boldsymbol{x} \tag{1}$$

式中，$\boldsymbol{x},\boldsymbol{y}$代表每一残差单元的输入输出向量，具有同一维度；$\{W_i\}$代表每一网络层权重值构成的集合；$F(\boldsymbol{x},\{W_i\})$代表所学习到的残差映射.图3所示非线性层输出为$F=W_2(\delta(W_1\boldsymbol{x}))$，其中$\delta$代表激活函数ReLU[25]，此处偏差省略，然后将$\delta(\boldsymbol{y})$输入到下一残差单元执行相同操作.在式(1)中，$\boldsymbol{x}$与F尺寸必须相等，在更改输入和输出通道时，可通过残差连接执行线性W_s以匹配尺寸，其公式表示为

$$\boldsymbol{y}=F(\boldsymbol{x},\{W_i\})+W_s\boldsymbol{x} \tag{2}$$

采用图3所示的残差模块构建深度残差网络，将其作为图像特征提取的主干网，其结构如图4所示.为有效地得到红外图像及其对应的伪可见光图像的特征向量，经过训练得到双通道残差网络权重模型W_i、W_s.

基于残差网络每一层结构计算出每一卷积层输出大小，即

$$N=(W-F+2P)/S+1 \tag{3}$$

式中，输入图片大小为$W\times W$，卷积核大小为$F\times F$，步长为S，padding的像素数P，输出图片大小为$N\times N$.当红外图像与其对应生成的伪可见光图像经过图4网络后，特征输出大小如图5所示.其优势在于利用双通道的残差网络提取双模态图像特征进行融合，能够利用伪可见光图像信息来丰富红外图像信息.当采用红外通道网络结构时，仅获得红外图像预训练权重W_{IR}；当采用伪可见光通道网络结构进行训练时，会获得伪模态图像预训练权重W_{FRGB}；当采用双模态残差网络通道进行训练时，获得的网络模型W_{ADD}将同时具备红外和可见光这两种图像共同特性.

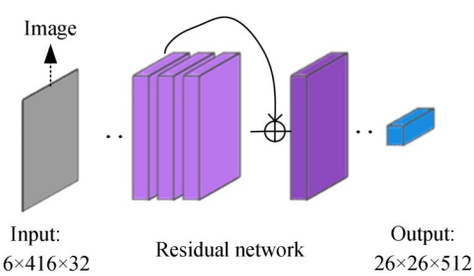


图5　图像经过残差网获得的特征向量

Fig. 5　Feature vector of image obtained by residual network

1.3　单阶段目标检测

如图1所示，本文的目标检测框架采用两种模态图像同时处理的结构，由两分支组成，一支用于红外图像输入，另一支用于伪模态图像(伪可见光图像)输入.

首先，通过每一分支上对应模态的图像进行预训练，对网络进行初始化.为避免两种模态成对训练，同时能够使用两种模态图像，在框架中使用CycleGAN循环生成对抗网络.在训练过程中，对于每幅红外图像的输入，都会生成一个伪RGB模态图像，并将伪RGB和红外图像传递给各输入分支.利用双通道残差网络对双模态图像提取各自特征向量，并利用add方式对特征向量值进行叠加，其通道数不变，对于描述图像特征映射的信息量增多，即描述图像的维度并无增加，而是每一维度的信息量在增加，从而有益于最终的分类.故而，针对红外目标检测的特性学习，采用各分支经过残差网络提取出的特征向量进行叠加，将融合的特征向量直接传递给单阶段检测网络进行训练.引入YOLOv3的损失函数，预测框的损失函数采用总方误差，其他部分的损失函数采用二值交叉熵.

1)　中心点坐标(x,y)损失的计算公式为

$$l_{\mathrm{xy}}=c\times(2-a)\times e \tag{4}$$

式中，c是置信度；e是x,y值的二值交叉熵，此值越小，整个网络损失值越小.a是归一化后预测框大小，$a=w\times h$越大，则$c\times(2-w\times h)$值越小.因此，这部分损失主要优化x,y的预测值和置信度以及$w\times h$回归值.

2） l_{wh}(anchor长宽回归值)损失的计算公式为

$$l_{wh}=c\times(2-a)\times(T_{twh}-P_{pwh})^2 \tag{5}$$

在确保置信度c为某一值时，为使损失值越小，需a越大，预测框大小P_{pwh}需要尽可能靠近真实框大小T_{twh}.这部分损失函数主要优化置信度以及l_{wh}回归值(a, T_{twh}).

3） 置信度损失(前背景)损失的计算公式为

$$l_{conf}=c\times e+(1-c)\times e\times i \tag{6}$$

式中，i表示IoU低于一定阈值但确实存在的目标(区域)，相当于Faster-RCNN中的中性点位，既不是前景也不是背景，可忽略不计.在确保置信度c的情况下，预测值需要尽可能靠近真实值，没有目标的部分需要尽可能靠近背景真实值，同时乘以相应的需要忽略的点位.这部分损失函数主要优化置信度，减小检测的目标量级.

4） 网络总损失函数表述为

$$L=l_{xy}+l_{wh}+l_{conf}+l_{class} \tag{7}$$

式中，l_{class}指置信度乘以多分类的交叉熵.为提高检测网络对场景中不同尺度目标的检测效果，采用三个不同尺度的特征图进行目标检测.结合图1，卷积神经网络在79层后，经过下方黄色的卷积层得到一种尺度的检测结果.相比输入图像，用于检测的特征图有32倍的下采样.如，当输入尺寸是416×416时，所得特征图尺寸是13×13.由于下采样倍数高，这里特征图的感受视野比较大，因此适合检测图像中尺寸比较大的目标.为了实现细粒度的检测，第79层的特征图又开始上采样(从79层往右开始上采样卷积)，随后与第61层特征图融合(Concatenation)，可得到第91层较细粒度的特征图.同理，经过多个卷积层后得到相对输入图像16倍下采样的特征图.它具有中等尺度的感受野，适合检测中等尺度的目标.最后，第91层特征图再次上采样，并与第36层特征图融合(Concatenation)，得到相对输入图像8倍下采样的特征图.它的感受视野最小，适合检测小尺寸的目标.预测对象类别时不使用softmax，而是改用logistic的输出进行预测，从而能够使检测网络支持多类别目标.

算法流程见表1.

表1 算法详细流程

Table 1 Algorithm detailed process

Algorithm：**PMFD**：Pse-model fused detection
Input：(1)Infrared image training set：$\{(x_i, y_i)\}_{i=1}^{n}$
(2)Generator of I2I framework：W_{I2R}
(3)stage1：IR Pre-trained：W_{IR}
(4)stage2：FRGB Pre-trained：W_{FRGB}
(5)stage3：Fusion Pre-train：W_{ADD}
Output：Trained PMFD model，$F(g)$
for num_epoches do
for x_i, $i=1, \cdots, n$ do
Through I2I framework generate a pseudo RGB $\hat{x}_i$ using W_{I2R}
Then the infrared image x_i and its corresponding pseudo RGB image $\hat{x}_i$ are input into the respective training channels using W_{IR} and W_{FRGB}, generate fusion vector (Tensor in Fig.1)
Pass the fusion vector to Fusion Pre-train network using W_{ADD}
Update W_{I2R}, W_{IR}, W_{FRGB}, W_{ADD} by minimizing Loss function of the PMFD model
end
end

2 实验结果与分析

2.1 数据集

实验部分采用2018年FLIR公司发布的自动驾驶数据集以及SODA数据集.FLIR ADAS[26]共包含12 886幅带标签的图像，每幅图像的分辨率为640×512.SODA[27]选取可用数据集15 00幅，每幅图像分辨率为640×480.该数据集无标签，首先将图片上采样到640×512，然后采用开源工具LabelImg[28]对图像进行打标

签.所有数据集均涵盖昼夜拍摄场景.如表2所示，按照标准(7∶2∶1)对数据集分割成训练集、验证集以及测试集.总的数据集包含person(31 242个实例)和car(61 763个实例)类别.图6显示了来自数据集的一些示例图像.

表2　数据集分布

Table 2　Dataset distribution

Data sets	Number of samples
Training set	8 140
Validation set	2 326
Testing set	1 163
Total	11 629

(a) FLIR′s autopilot sample image

(b) Infrared street view image data from soda

图6　数据集示例图像

Fig. 6　Dataset sample images

2.2　评价标准

为了评估网络模型性能，采用通用技术指标：交并比(IoU)、F1-score和平均准确率(mAP).IoU是模型生成的预测边界框与ground truth的重叠率.当IoU超过阈值时，边界框被认为是正确，即

$$a=\frac{B_{\text{pred}}\cap B_{\text{truth}}}{B_{\text{pred}}\cup B_{\text{truth}}}\geqslant a_0 \tag{8}$$

该标准用于测量ground truth和预测之间的相关性，相关性越高，IoU值越高.后续将使用IoU计算检测模型的mAP.将输入图像放入模型中进行预测，得到红外图像中目标预测边界框B_{pred}.当B_{pred}和B_{truth}的IoU大于阈值a_0，并且满足式(8)时，则认为预测正确.采用该指标的目的是计算预测边界框和ground truth之间的IoU，当IoU大于50%阈值时，测试结果为真阳性(N_{TP})；小于阈值时，称为假阳性(N_{FP}).假阴性(N_{FN})表示模型预测图像中没有目标，但实际图像包含目标的情况.将模型检测性能用精确度和召回率两个指标来衡量，公式分别为

$$P_{\text{precision}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}} \tag{9}$$

$$R_{\text{recall}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}} \tag{10}$$

$$F_{\text{F1-score}}=\frac{2\times P_{\text{precision}}\times R_{\text{recall}}}{P_{\text{precision}}+R_{\text{recall}}} \tag{11}$$

第三种评估指标mAP是描述召回率和精度交并的区域，此值介于0和1之间，mAP值越大，检测效果越好.红外图像以及利用图像转换模型CycleGAN生成其对应场景下的伪可见光图像，如图7所示.

(a) Infrared image data

(b) Pseudo visible image generated by CycleGAN

图7 红外图像以及其对应伪可见光图像

Fig. 7 Infrared image and its corresponding pseudo visible image

2.3 测试结果

本文算法基于pytorch框架实现.训练软硬件平台配置为CPU：Intel i7-9700F、GPU：配置CUDA10.0和cuDNN7.6.5的Nvidia GTX2080Ti、内存为16G、显存为11G.训练过程中采用GPU加速.

单个深层神经网络检测目标算法(Single Shot Multi-box Detector, SSD)[29]采用VGG-16作为图像特征提取的主干网络，由于使用conv4_3提取低级特征去检测目标，并且该卷积层的数量较少，存在严重的特征提取不充分.Faster-RCNN通过Resnet提取红外图像特征，然后将得到的特征映射送入RPN(Region Proposal Network)，RPN生成待检测框(称指定RoI的位置)，并对RoI进行第一次修正.RoI Pooling Layer根据RPN的输出，在特征映射上选取每个RoI对应的特征.最后，使用全连接层对目标框分类，并对RoI进行第二次修正.Faster-RCNN属于双阶段网络，检测红外图像目标运行时间过长，满足不了实时性.同时，该网络采用浅层Resnet网络提取红外图像主要特征，对较小目标信息提取不充分，存在红外图像中目标信息丢失的问题.MMTOD通过将Faster-RCNN改造成双通道提取红外图像与可见光图像特征，但由于Resnet对图像中较小目标特征提取不明显，丢失目标信息较多，且结构冗余，既是双通道又是双阶段检测网络，该算法运行时间较Faster-RCNN更长.

针对上述效率与准确率问题，利用本文提出优化后的深度学习模型PMFD训练已有数据集，同时采用Baseline(基于Yolov3网络仅对单模态红外图像进行训练)作对比.基于Baseline改造原因，该网络将Resnet-34深度残差网络作为主干网络，采用跳跃式残差连接，能够充分提取图像信息，且在提取目标特征效率方面较前三者有很大优势.由于红外图像自身特点，Baseline对其包含的目标内容进行检测仍存在较大难度.本文算法借助可见光图像信息弥补红外信息缺失，由于Baseline中Resnet-34对于特征提取优势显著，将其改造成双通道Resnet-34分别提取可见光以及红外图像特征.但由于红外图像缺少场景匹配的可见光图像，利用训练好的CycleGAN生成伪可见光图像.之后，将通过双通道残差网络Resnet-34得到的伪可见光和红外图像特征进行融合，输入到PMFD单阶段检测网络部分.本文算法通过改造Baseline，能够充分快速地提取红外图像以及利用CycleGAN生成其对应伪可见光图像的特征，并能够将两者特征融合以完成更好的红外

目标检测，在效率和准确率方面占绝对优势．表3给出了两个数据集红外图像经过五种不同算法处理后检测准确率、召回率、F1-score以及mAP，用于定量化不同算法的性能．实验结果表明，本文方法在测试数据集上的平均准确率达到0.813.

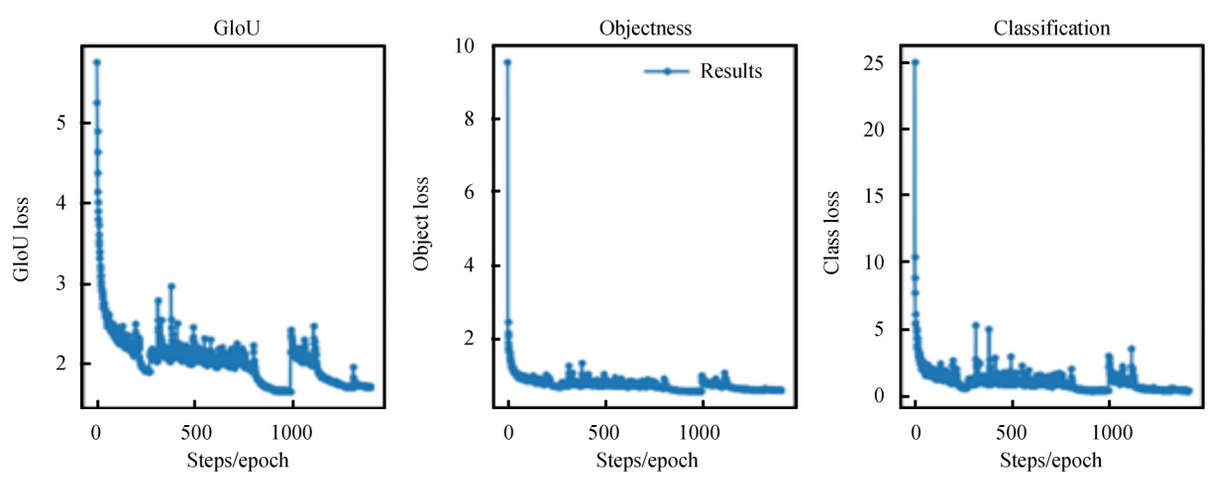
表3　在测试集上的实验结果

Table 3　Experimental results on test set

Networks		Precision	Recall	F1-score	mAP
SSD	All	0.564	0.836	0.674	0.571
	Car	0.603	0.881	0.716	0.628
	Person	0.525	0.790	0.631	0.514
Faster-RCNN	All	0.581	0.841	0.687	0.612
	Car	0.625	0.880	0.731	0.676
	Person	0.536	0.801	0.642	0.547
Baseline	All	0.608	0.892	0.723	0.786
	Car	0.630	0.909	0.744	0.82
	Person	0.586	0.874	0.701	0.752
MMTOD	All	0.629	0.893	0.738	0.800
	Car	0.640	0.902	0.749	0.835
	Person	0.618	0.884	0.727	0.765
PMFD	All	0.625	0.909	**0.741**	**0.813**
	Car	0.638	0.923	0.754	0.839
	Person	0.611	0.894	0.726	0.786

PMFD训练模型的学习率为0.001，批量（Batch Size）大小为16．如图8所示，整个训练过程包含1 400次迭代．在前1 000次训练过程中Baseline仅训练红外图像过程，后400次迭代是在前1 000次的基础上训练本文所提出的模型，发现中间有段mAP特别不稳定的状态，原因在于训练方式不同．训练得到的最终模型对单个图像进行检测平均需要48.76 ms，Baseline耗时23.5 ms，SSD耗时24.56 ms，Faster-RCNN耗时80.2 ms，MMTOD耗时110 ms.

分别从两类数据集中选取五幅背景复杂包含目标较多的红外图像作为测试图像．图像中既有较长又有较短成像距离的目标，同时光线以及街面建筑物对目标成像对比度影响较大．图9～10分别给出了用SSD、Faster-RCNN两类经典目标检测网络对两个数据集的红外图像的检测效果，以及用Baseline、相关研究领域先进算法MMTOD和本文算法获得的效果．从对比图明显看出通过SSD、Faster-RCNN两类方法对红外图像进行检测时，对于轮廓不清晰以及尺寸较小的人和车目标，存在预测不准确以及漏检情况．此外，SSD尽



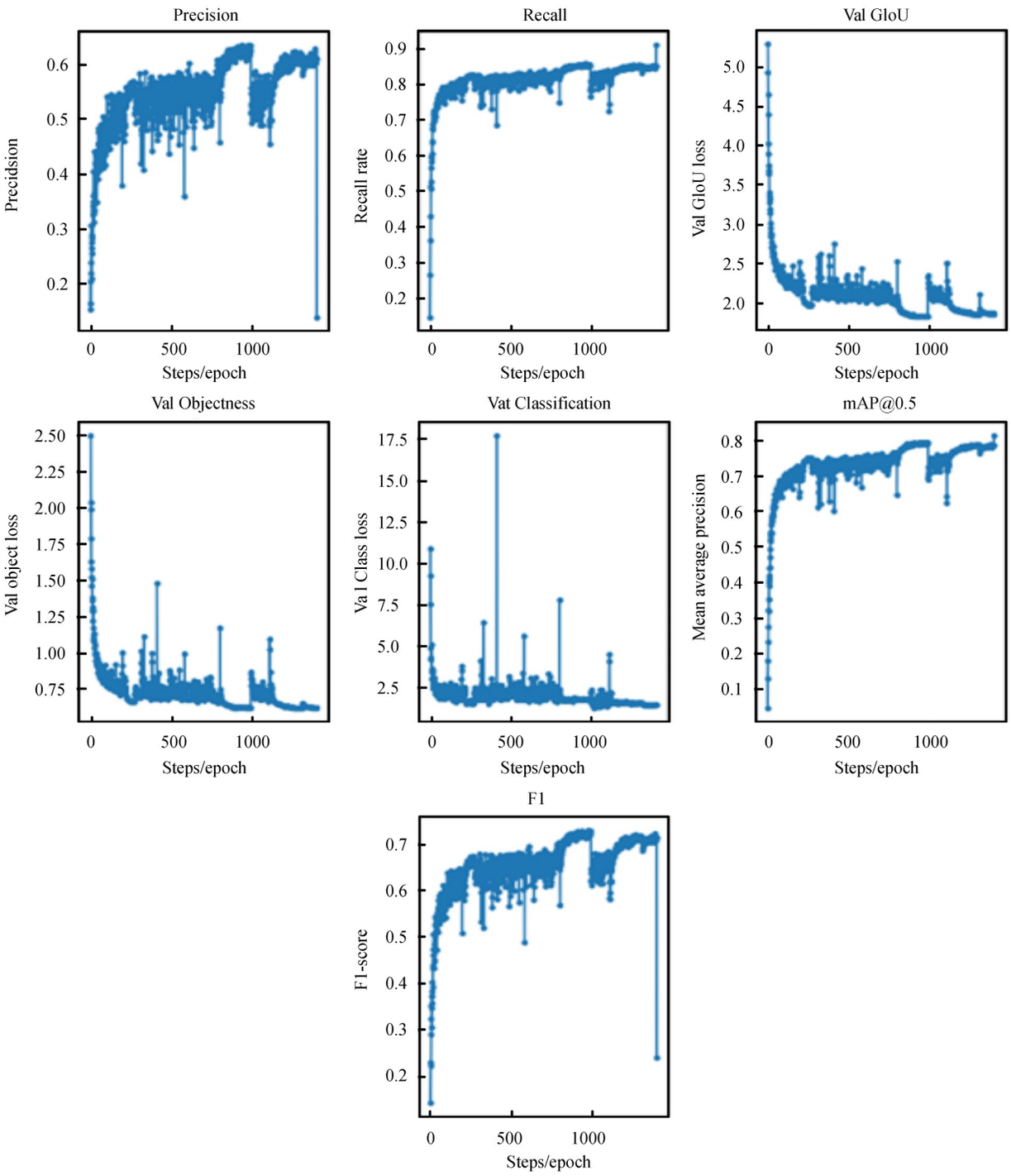
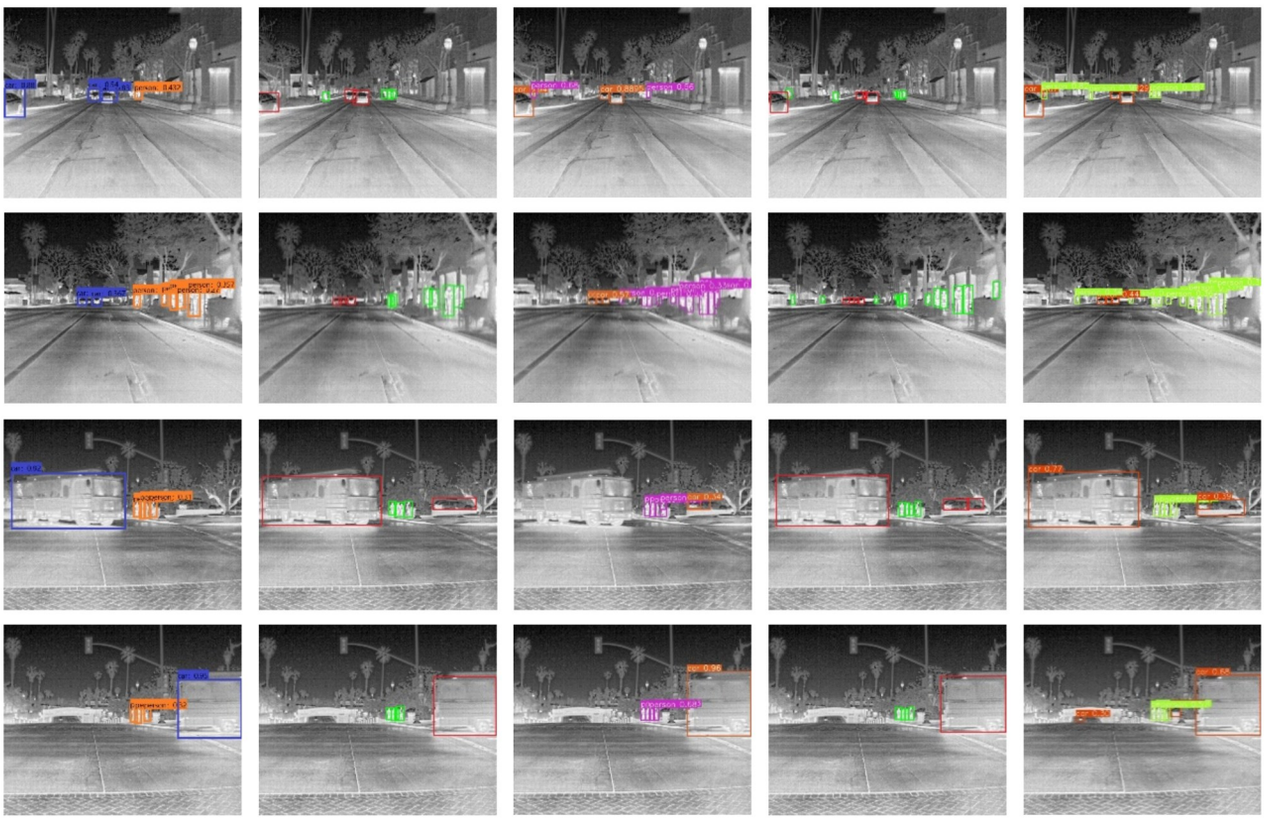


图 8 在训练集上训练车和人每次迭代的平均精度

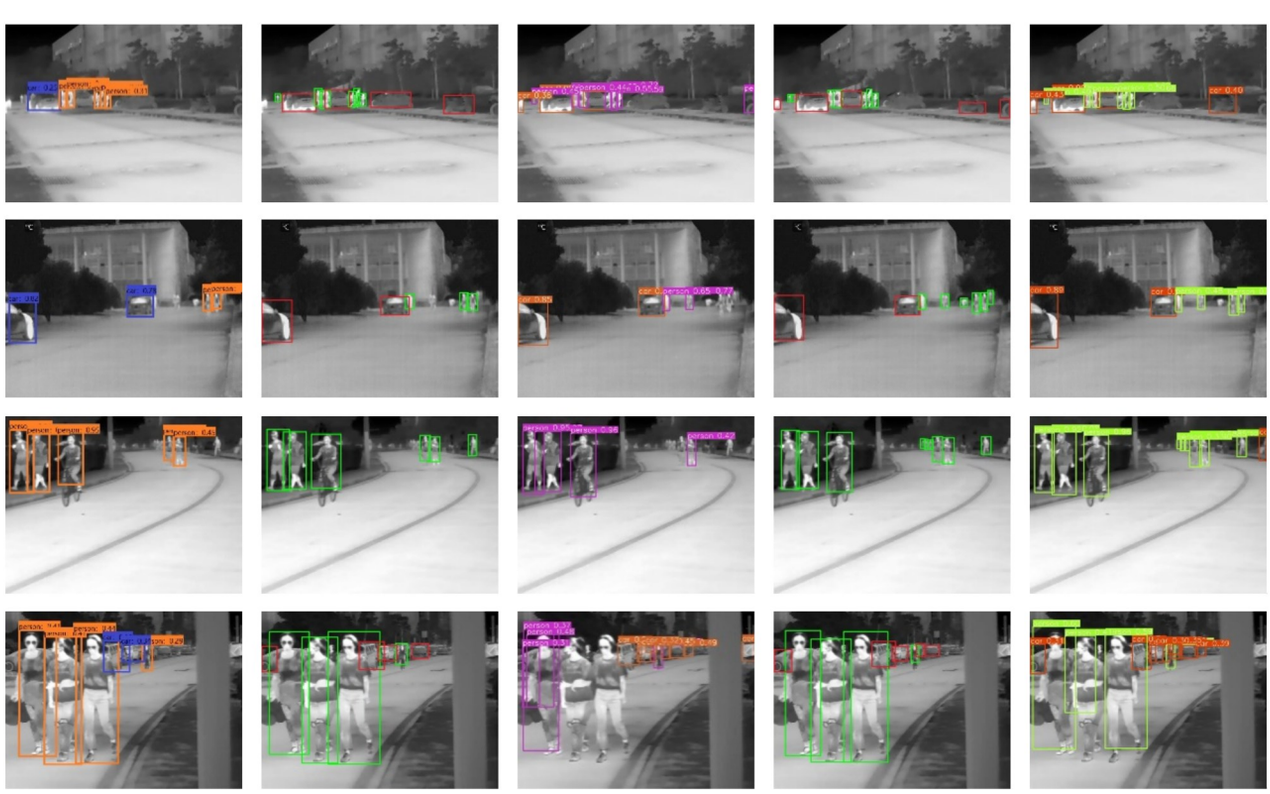
Fig. 8 Average accuracy of each iteration of training vehicle and person on training set



(a) SSD (b) Faster-RCNN (c) Baseline (d) MMTOD (e) PMFD

图 9 在 FLIR-ADAS 数据集上检测效果

Fig.9 Detection result on FLIR-ADAS data set



(a) SSD (b) Faster-RCNN (c) Baseline (d) MMTOD (e) PMFD

图 10 在SODA数据集上检测效果

Fig. 10 Detection result on SODA data set

管在检测目标的实时性方面具有优势，但mAP相对于其他方法低很多；Faster-RCNN相较于SSD准确率稍有提高，但耗时过长；Baseline虽采用多尺度预测，但由于红外图像纹理特征不明显，成像对比度低，噪声干扰较多，在准确率的提升上面相对于Faster-RCNN并不大，然而该算法实时性满足要求；MMTOD能够利用可见光图像纹理信息丰富红外图像目标信息以提高检测准确率，但对目标信息提取不充分，导致检测虚假目标数量较多，同时对于较小目标召回率较本文算法低很多，检测时间又过长，很难满足工程应用。本文算法PMFD相对于其他算法在实时性和准确率取得平衡，获得的检测效果也有较大提升。

3 结论

本文提出了一种基于伪模态转换的红外目标融合检测算法。为了融合红外图像生成的伪可见光图像丰富的语义信息，通过将两种图像特征融合的方式扩展和改善基于YOLOv3的目标检测器。实验部分采用FLIR ADAS和SODA两种数据集评估检测算法。实验结果证明本文框架不仅比YOLOv3有更好的性能，而且能够提供一个结构更为简单紧凑的方式来提高红外图像中目标检测的性能。但由于该框架需要使用对抗生成网络解决伪可见光图像生成问题，算法的运行时间有待进一步改善。同时，针对弱小目标、重叠遮挡等情况算法仍有较大的改进空间。

参考文献

[1] LIU S, LIU Z. Multi-channel CNN-based object detection for enhanced situation awareness[DB/OL]. [2020-04-09]. https://arxiv.org/abs/1712.00075v1.

[2] NING Qiang, QIN Peng-jie, SHI Xin, *et al*. Infrared target detection algorithm under complex ground background[J]. *Acta Photonica Sinica*, 2019, **48**(4): 0410001.
宁强, 秦鹏杰, 石欣, 等. 复杂地面背景下的红外目标检测算法[J]. 光子学报, 2019, **48**(4): 0410001.

[3] BERTOZZI M, BROGGI A, GOMEZ C H, *et al*. Pedestrian detection in far infrared images based on the use of probabilistic templates[C]. 2007 IEEE Intelligent Vehicles Symposium, 2007, **2**: 327-332.

[4] DAVIS J W, KECK M A. A two-stage template approach to person detection in thermal imagery[C]. 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05), 2005, **1**: 364-369.

[5] JUNGLING K, ARENS M, JARVERS G, *et al*. Feature based person detection beyond the visible spectrum[C]. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009, **1**: 30-37.

[6] BAY H, TUYTELAARS T, VAN G L. Surf: Speeded up robust features[C]. European Conference on Computer Vision, Springer, Berlin, Heidelberg, 2006, **3951**: 404-417.

[7] PENG M, WANG C, CHEN T, *et al*. NIRFaceNet: a convolutional neural network for near-infrared face identification [J]. *Information*, 2016, **7**(3): 61-74.

[8] LEE E J, KO B C, NAM J Y. Recognizing pedestrian's unsafe behaviors in far-infrared imagery at night[J]. *Infrared Physics & Technology*, 2016, **76**(6): 261-270.

[9] CHEVALIER M, THOME N, CORD M, *et al*. Low resolution convolutional neural network for automatic target recognition[C]. 7th International Symposium on Optronics in Defence and Security, 2016, **11**: 1157-1161.

[10] RODGER I, CONNOR B, ROBERTSON N M. Classifying objects in LWIR imagery via CNNs[C]. Electro-Optical and Infrared Systems: Technology and Applications XIII, International Society for Optics and Photonics, 2016, **9987**: 99870-99884.

[11] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C]. Advances in Neural Information Processing Systems, 2012, **25**: 1097-1105.

[12] REDMON J, FARHADI A. Yolov3: an incremental improvement[DB/OL]. [2020-04-09]. https://arxiv.org/abs/1804.02767.

[13] LIU Hui, HE Yong, HE Bo-xia, *et al*. Infrared target tracking algorithm based on multiple feature fusion and region of interest prediction [J]. *Acta Photonica Sinica*, 2019, **48**(7): 0710004.
刘辉, 何勇, 何博侠, 等. 基于多特征融合与ROI预测的红外目标跟踪算法[J]. 光子学报, 2019, **48**(7): 0710004.
[14] WAGNER J, FISCHER V, HERMAN M, *et al*. Multispectral pedestrian detection using deep fusion convolutional neural networks[C]. ESANN, 2016, **587**: 509-514 .
[15] CHOI H, KIM S, PARK K, *et al*. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks[C]. 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, **23**: 621-626.
[16] KONIG D, ADAM M, JARVERS C, *et al*. Fully convolutional region proposal networks for multispectral person detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, **36**: 243-250.
[17] LIU J, ZHANG S, WANG S, *et al*. Multispectral deep neural networks for pedestrian detection[DB/OL].[2020-04-09]. https://arxiv.org/abs/1611.02644.
[18] CHAITANYA D, NINAD A, MANUJ M S, *et al*. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery[DB/OL].[2020-04-09]. https://arxiv.org/abs/1905.08789.
[19] REN S, HE K, GIRSHICK R, *et al*. Faster-RCNN: Towards real-time object detection with region proposal networks [C]. Advances in Neural Information Processing Systems, 2015, **39**: 91-99.
[20] ZHU J Y, PARK T, ISOLA P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]. Proceedings of the IEEE International Conference on Computer Vision, 2017, **22**: 2242-2251.
[21] BISHOP C M. Neural networks for pattern recognition[M]. Oxford University Press, 1995, **12**: 1235-1242.
[22] REN S, HE K, GIRSHICK R, *et al*. Object detection networks on convolutional feature maps[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **39**(7): 1476-1481.
[23] VEDALDI A, FULKERSON B. VLFeat: an open and portable library of computer vision algorithms[C]. Proceedings of the 18th ACM International Conference on Multimedia, 2010, **18**: 1469-1472.
[24] LIN M, CHEN Q, YAN S. Network in network[DB/OL].[2020-04-09]. https://arxiv.org/abs/1312.4400.
[25] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]. Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, **27**: 807-814.
[26] FA Group. FLIR thermal dataset for algorithm training[DB/OL]. [2018-07-26]. https://www.flir.cn/oem/adas/adas-dataset-agree.
[27] LI C L, XIA W, LUO B, *et al*. Segmenting objects in day and night: edge-conditioned CNN for thermal image semantic segmentation[DB/OL].[2020-04-09]. https://arxiv.org/abs/1907.10303.
[28] Tzutalin. LabelImg[EB/OL]. [2018-07-06]. https://github.com/tzutalin/labelImg.
[29] LIU W, REED S, BERG A C, *et al*. SSD: Single shot multi-box detector[DB/OL].[2020-04-09]. https://arxiv.org/abs/1512.02325.

Foundation item: Open Project Program of the Key Laboratory of Intelligent Infrared Perception, Chinese Academy of Sciences (No. CAS-IIRP-04)