引用格式：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

一种基于跨域融合网络的红外目标检测方法

赵明[1,2],张浩然[1]
(1 上海海事大学 信息工程学院,上海 201306)
(2 中国科学院上海技术物理研究所 中国科学院智能红外感知重点实验室,上海 200083)

摘　要:由于红外图像本身缺乏纹理信息,多数目标检测网络针对红外图像难以达到理想的检测效果,该方法提出了一种跨域融合网络结构,结合多个模态进行红外目标检测。首先,采用无需成对的图像转换网络,对已有的红外数据集进行模态转换,生成伪可见光数据集;然后,提出了红外域和伪可见光域双通道的多尺度特征融合结构,采用特征金字塔网络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;最后,为了弥补融合过程中的纹理缺失,提出软权重分配模块,通过拼接参数化后的源域、目标域和融合域特征,自适应分配和优化网络权重,从而提高特征提取与目标检测的精度。与常规方法相比该方法方法具有更好的红外目标检测性能。

关键词:红外图像;红外目标检测;模态转换网络;跨域融合;软权重分配
中图分类号:TP391.4　　**文献标识码**:A　　**doi**:10.3788/gzxb20215011.1110001

An Infrared Object Detection Method Based on Cross-domain Fusion Network

ZHAO Ming[1,2], ZHANG Haoran[1]
(1 *College of Information Engineering, Shanghai Maritime University, Shanghai* 201306, *China*)
(2 *Key Laboratory of Intelligent Infrared Perception, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai* 200083, *China*)

Abstract: Because the infrared image lacks certain texture information, most target detection networks cannot achieve great detection results for infrared images. This paper proposes a cross-domain fusion network structure that combines multiple modal for infrared target detection. Using image conversion network without pairing, modal conversion of existing infrared dataset to generate a pseudo-visible light dataset. Then, this paper proposes a dual-channel multi-scale feature fusion structure in the infrared domain and the pseudo-visible light domain, uses feature pyramid network to obtain the feature map of each mode, and performs dual-modal feature fusion for multi-scale features. Finally, in order to make up for the lack of texture in the fusion process, this paper proposes a soft weight distribution module. By splicing the parameterized source domain, target domain and fusion domain features, the network weight is assigned and optimized through learning, thereby improving the accuracy of feature extraction and target detection. The experimental results show that the method in this paper has better infrared target detection performance compared with the conventional method.

Key words: Infrared image; Infrared target detection; Modal transformation network; Cross-domain fusion; Soft weight distribution
OCIS Codes: 100.3008; 200.4260; 100.4996; 110.3080

基金项目:上海市自然科学基金(No.20ZR1423500),中国科学院智能红外感知重点实验室开放课题(No.CAS-IIRP-04)
第一作者:赵明(1984—),女,副教授,博士,主要研究方向为红外弱小目标的图像处理。Email: mingzhao@shmtu.edu.cn
通讯作者:张浩然(1997—),男,硕士研究生,主要研究方向为红外目标检测。Email:hhhxl_zhr@163.com
收稿日期:2021-04-30;**录用日期**:2021-07-01

0 引言

与可见光成像系统相比，红外成像系统受光照影响较小、探测距离较远、成像相对稳定。红外目标检测是红外成像中的一项关键技术，其在军事预警、安全监测和红外制导等领域有着重要的应用价值[1-4]。

自从深度卷积神经网络(Convolution Neural Network，CNN)[5]出现以来，目标检测在可见光领域有了飞速的发展。常见的双阶段网络，例如，R-CNN(Regions with CNN features)[6]采用选择性搜索算法划分区域(Selective Search，SS)，Fast-RCNN[7]引入感兴趣区域池化(Region Of Interest Pooling，ROI Pooling)以池化不同大小的特征图加快处理速度，Faster-RCNN[8]采用区域生成网络(Region Proposal Network，RPN)识别感兴趣区域，进一步提升了网络的精度和速度。与此同时一些单阶段网络也应运而生，例如You Only Look Once(YOLO)系列和Single Shot MultiBox Detector(SSD)[9]将生成备选框与检测一体化，即不单独提出备选框，从而保证更快的检测速度。YOLO-V3[10]利用多尺度特征进行目标检测，虽然在精度上不及双阶段网络，但是其检测速度有优势。YOLO-V4[11]采用CSPDarknet-53骨干网以提高检测精度和检测速度。然而，基于深度学习的目标检测网络多是针对自然光图像提出的，采用对应的公开数据集(如ImageNet、PASCAL-VOC和MS-COCO)进行框架和模型的训练。直接将经典的深度学习网络应用于红外目标检测效果不好。由此，研究者们在经典检测网络的基础上开展了对红外图像目标检测的研究。BAEK J等[12]使用YOLO-V3目标检测网络对红外图像进行检测，结果表明检测精度与自然光图像相比明显降低。LEE E J等[13]设计了一种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，然后结合一个随机森林分类器用于检测红外图像中的行人目标，其精度高于某些与CNN相关的算法，并且处理时间更短。RODGER I等[14]利用长波红外传感器(Long Wave Infrared Sensor，LWIR)开发了一种中短程红外高分辨率图像的卷积神经网络训练器，网络能够完成中短程红外目标的检测，然而却难以检测到远程目标。BERG A[15-16]利用车载红外相机提出一种基于异常障碍物检测的方法。LEYKIN A等[17]设计了一种用于多光谱行人检测的融合跟踪器和行人分类器的系统，该系统通过使用动态适应的背景模型将前景区域从每帧中分割出来，采用周期性步态分析进行评估，对光照噪声具有鲁棒性，在室外环境中表现较好。

然而，由于红外图像的获取成本高，目前已经公开的大规模红外数据集相对较少。同时，与可见光图像相比，红外图像分辨率较低，同时缺乏纹理信息。单一地依靠现有红外图像样本进行目标检测具有一定的局限性。为了提高红外目标检测的效果，出现将可见光与红外图像相结合的思路进行多模态目标检测。WAGNER J等[18]应用聚合信道特征(Aggregate Channel Feature，ACF)和增强决策树(Boost Decision Tree，BDT)来生成兴趣区域，并且采用融合可见光和红外信息的卷积神经网络对这些兴趣区域进行分类。CHOI H等[19]分别对可见光图像和红外图像使用两个单独的RPN生成候选区域用以检测。GHOSE D等[20]对红外图像行人检测提供一种注意机制，该机制通过融合多光谱图像以训练YOLO-V3进行检测。然而，上述的方法几乎都要用到和红外场景相对应的可见光场景数据集，该类数据集缺乏。DEVAGUPTAPI C[21]提出一种伪多模态目标检测(Pseudo Multi-modal Object Detection in Thermal Imagery，MMTOD)。该方法通过循环生成对抗网络得到伪可见光图像，并由两个骨干网分别对两种图像进行特征提取，然后对最终的两种特征进行融合，以达到利用可见光信息完成红外图像检测的效果。但是该算法难以完成多尺度预测，同时时间复杂度较高。

本文采用跨域融合的红外目标检测网络(Cross-domain Fusion Network，CFN)，对多个模态进行操作。网络旨在生成、整合、利用多个域的特征，去弥补红外图像语义信息缺失的不足，达到增强网络目标检测精度的效果。首先，采用已有的红外数据集在模态转换网络上生成伪可见光数据集，两个数据集共享标注，作为两个并行特征提取通道的输入。为了有效利于高层次语义特征的同时，也不忽略高分辨率特征图丰富的内容信息，基于红外域和伪可见光域双通道的多尺度特征融合结构，采用特征金字塔(Feature Pyramid Network，FPN)[22]获取每个模态的特征图，针对多尺度特征进行双模态特征融合，生成融合特征金字塔。针对跨域融合过程中的特征损失，提出软特征分配模块，利用红外图像和伪可见光图像的特征金字塔，对融合特征金字塔进行优化补偿。融合后的特征送入RPN，利用RPN完成粗分类并将生成的候选框映射到经过软权重分配模块得到的特征图上。最后利用全连接层完成目标分类和定位。

1 方法

如图1所示，提出的跨域融合网络的红外目标检测框架由模态转换网络、跨域融合网络、软权重分配模块三部分组成。首先，采用无需成对的图像转换网络（Contrastive Learning for Unpaired Image-to-Image Translation，CUT）[23]实现红外域（源域）到可见光域（目标域）的模态转换，获取色域信息丰富的可见光图像。其次，两种模态的图像分别经过残差网络获取特征图，利用FPN完成高层次特征图语义信息向低层次特征图的过渡，形成两个语义信息丰富的特征金字塔。融合的方式选择特征层对应1×1卷积并通过通道乘法完成信息交互，如图1（a）所示。然后，构建软权重分配模块，利用搭建的特征融合网络，以动态的形式融合源域、目标域、合成域的特征图，从而实现多域信息的互补。最后，融合的特征输入RPN，RPN生成锚框，Bounding box对锚框的边界框进行偏移量回归。同时，将RPN生成的兴趣区域映射到软权重模块得到的特征图上，通过全连接层进行目标分类和定位，如图1（b）所示。

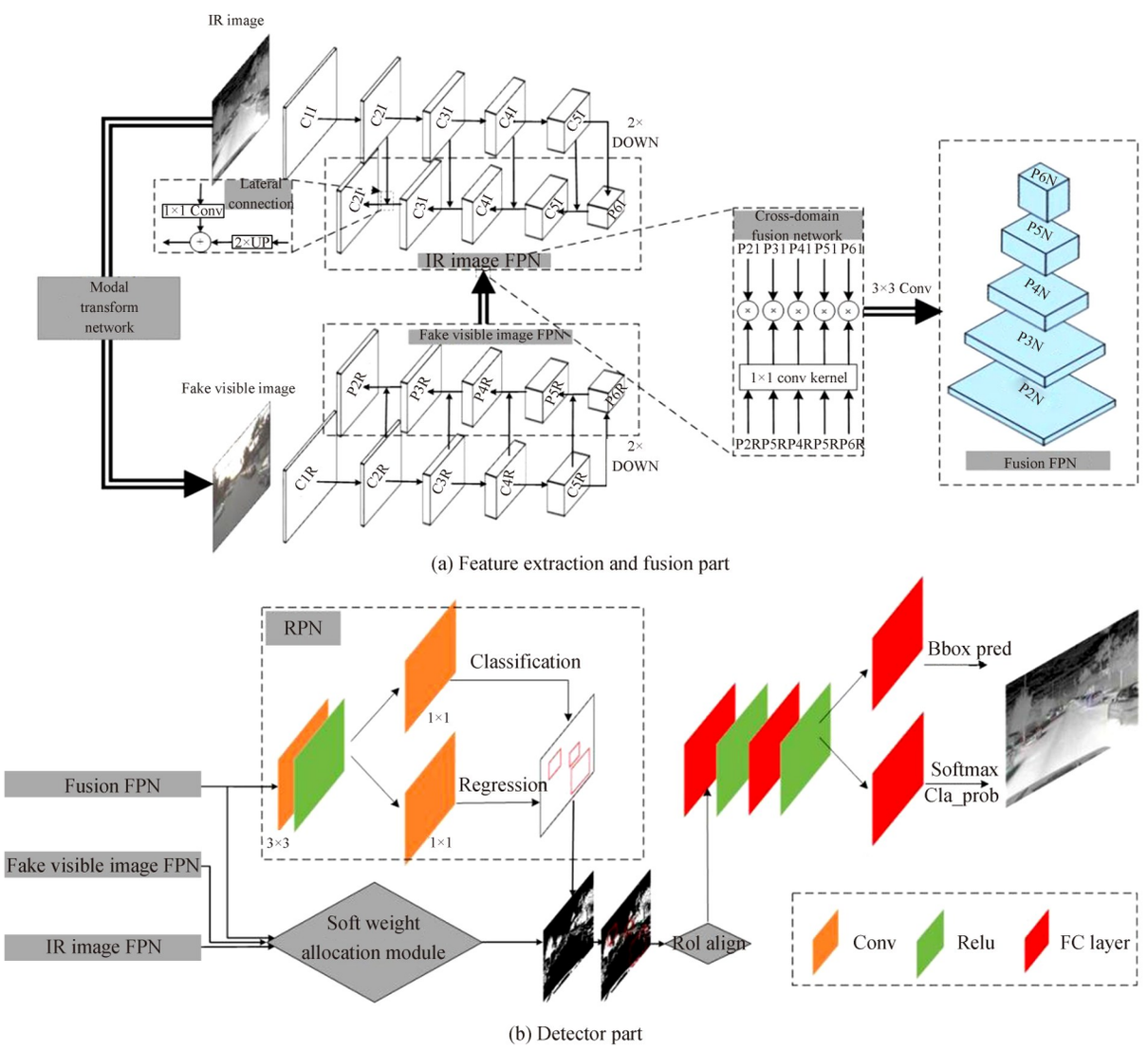


图1 网络整体框架

Fig. 1 Overall network architecture

1.1 模态转换网络

CycleGAN（Cycle Global Area Network）[24]是一种在缺少成对数据的情况下，学习从源数据域X到目标数据域Y的方法。而CUT作为最新且具有最先进性能的风格转换网络，保留了CycleGAN不需要成对源域和目标域数据集进行训练的优势，并且体系结构简单。同时，不同于CycleGAN，CUT允许只学习一个方向

的映射，只使用一对生成器和鉴别器，属于轻量级网络。如图2所示，CUT利用对比学习的框架来最大化源域和目标域中对应图像块之间的互信息。其具体做法是，在输入图像和输出图像的同一位置上设置图像块，并将输出图像所生成的图像块嵌入到原图中。与此同时，原图也会选取一些和输出图像相似但并不在同一位置的图像块，以此来计算输入和输出图像间的对比损失。具体对比损失公式如式1所示，其中z表示输出图像所生成的图像块，z^+表示输入图像中对应位置的图像块，z^-表示输入图像中非对应位置的图像块，N表示非对应图像块的个数，n表示当前非对应图像块的索引，τ取常数0.07。

$$\ell\left(z,z^+,z^-\right)=-\log\left[\frac{\exp\left(z\cdot z_n^+/\tau\right)}{\exp\left(z\cdot z_n^+/\tau\right)+\sum_{n=1}^{N}\exp\left(z\cdot z_n^-/\tau\right)}\right] \tag{1}$$

本方法中，CUT主要负责将源域$X\subset R^{H\times W\times 1}$的红外图像转换为目标域$Y\subset R^{H\times W\times 3}$的伪可见光图像。输出的目标域图像表示为

$$\hat{y}=G(z)=G_{\mathrm{dec}}\left(G_{\mathrm{enc}}(x)\right) \tag{2}$$

式中，网络框架包括编码器G_{enc}和解码器G_{dec}，$\hat{y}$表示生成的伪可见光目标域图像，x表示红外源域图像。对比损失为来自两个域的图像编码器编码所产生信息的差值。

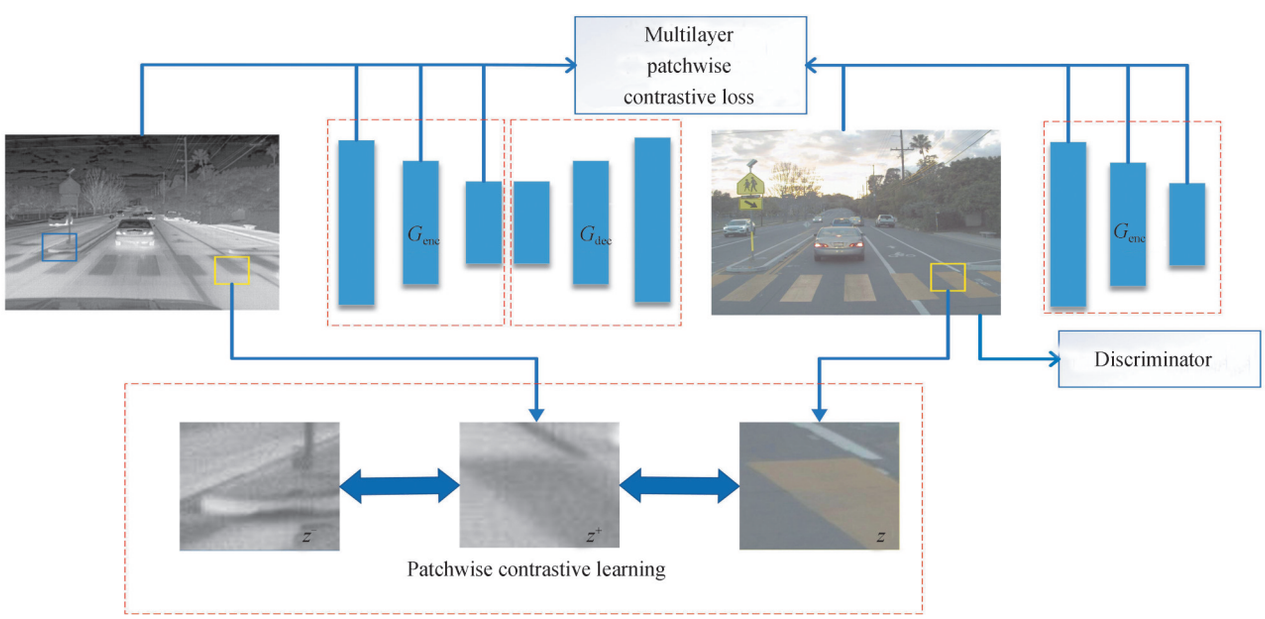


图2 红外域和可见光域对比学习框架

Fig. 2 Comparative learning framework

1.2 跨域融合网络

FPN的多尺度网络具有横向连接的自上而下的体系结构，用于在所有尺度上构建高级语义特征映射。基于该结构框架，构建多层特征图跨域融合网络。特征提取网络分为红外通道和可见光通道，均采用Resnet-101[25]作为backbone进行特征提取。

红外域图像通道经过backbone提取图像特征，得到特征图$\{C_{1\mathrm{I}},C_{2\mathrm{I}},C_{3\mathrm{I}},C_{4\mathrm{I}},C_{5\mathrm{I}}\}$。$C_{5\mathrm{I}}$作为最高层次的特征层，包含了最丰富的语义信息。随着卷积层数的增加，对应特征层的分辨率不断降低，结构中最高层$C_{5\mathrm{I}}$所包含的纹理信息最少。将这个特征层进行1×1卷积降维得到特征层$P_{5\mathrm{I}}$。以横向连接的自下而上的体系结构为基础，对该层上采样后与其对应的特征层叠加，得到$\{P_{2\mathrm{I}},P_{3\mathrm{I}},P_{4\mathrm{I}},P_{5\mathrm{I}},P_{6\mathrm{I}}\}$特征层，其中$P_{6\mathrm{I}}$为$P_{5\mathrm{I}}$下采样所得。具体结构结合图1中lateral connect部分。

跨域融合网络的另一条通道输入是由模态转换网络所得的伪可见光域图像。可见光域图像通过骨干网Resnet-101提取特征，得到各个尺度分辨率对应的特征图$\{C_{1\mathrm{R}},C_{2\mathrm{R}},C_{3\mathrm{R}},C_{4\mathrm{R}},C_{5\mathrm{R}}\}$。如图1(a)中Fake Visible FPN所示，其获取方式与红外图像特征金字塔相同，定义该部分参与特征融合的特征金字塔为

$\{P_{2R}, P_{3R}, P_{4R}, P_{5R}, P_{6R}\}$。

基于上述两个通道进行特征提取，既得到了源域和目标域图像的低级网络特征映射的精准定位信息，也获取了高级网络特征映射的丰富语义信息。FPN为网络构建了多层次的特征图，通过采用对应尺度特征图相融合的方式，交互源域和目标域特征金字塔的信息，构建出包含两种模态语义信息的多尺度特征融合金字塔$\{P_{2N}, P_{3N}, P_{4N}, P_{5N}, P_{6N}\}$。为解决不同域特征图因通道数不同所造成的融合特征缺失的问题，在该部分首先采用1×1的卷积核对三通道伪可见光图像进行卷积，这个过程定义1×1卷积层含有一个卷积核，这个卷积核在卷积过程中将原来三个通道进行跨通道线性组合变成单个通道的特征图；然后引入激活函数的非线性表达，强化网络学习能力；最后通过通道乘法将经过1×1卷积层后的伪可见光特征和红外特征融合在一起，表示为

$$f_{\mathrm{FPN}}(\theta)=\mathrm{Conv}_{1\times1}^{s=1}\left(f_{\mathrm{RGB}}(\theta)\right)\otimes f_{\mathrm{IR}}(\theta) \tag{3}$$

式中，$\otimes$表示信道乘法，s表示卷积运算的步长，f_{RGB}表示伪可见光图像分支的输出特征，f_{IR}表示红外图像分支的输出特征，θ表示特征层的索引。最后，针对每个合并的特征图分别采用一个3×3的卷积来生成特征金字塔的融合特征，从而减少上采样的混叠效果。

1.3 软权重分配模块

针对不同域图像融合的过程相当于对可见光图像特征进行降维操作。可见光特征图经过1×1卷积层，将RGB三个通道信息合并成一个通道后，再与单通道的红外信息进行交互，这个过程势必会带来一定的信息损失。因此，提出一种软权重分配策略，分别充分利用两个模态图像的特征金字塔，对上述三种不同域的特征进行融合。

软权重分配模块的目的是通过参数化不同域的特征金字塔，利用所有金字塔级别的特征去学习并生成更好的拼接特征。如图3所示，软权重分配模块以三个域的特征金字塔作为输入，首先，将这些特征在列方向上进行首尾拼接，并为每一个特征金字塔生成一个空间权重映射，权重用于各个特征金字塔之间的聚合。获得拼接特征后，采用通道乘法保持输出拼接特征尺度不变。然后，在含有少量卷积层的情况下能够给来自不同域的特征自适应分配权重，从而使这些特征具有各个域的多尺度信息，达到弥补信息缺失的目的。最后，软权重分配模块通过反向传播方式调整网络中的其他参数，而不依赖硬性分配原则。上述拼接方法定义为

$$f_{\mathrm{CON}}=\mathrm{Concat}(f_{\mathrm{FPN}}, \alpha f_{\mathrm{RGB}}, \gamma f_{\mathrm{IR}}) \tag{4}$$

针对每种域的特征金字塔，网络将其输入软权重分配模块。其中，分别初始化α和γ用于控制两个模态特征金字塔的比例参数，在网络迭代过程中，这些参数会根据拼接特征所产生的损失函数通过反向传播寻找最优解。f_{CON}为最终拼接完成后的特征。

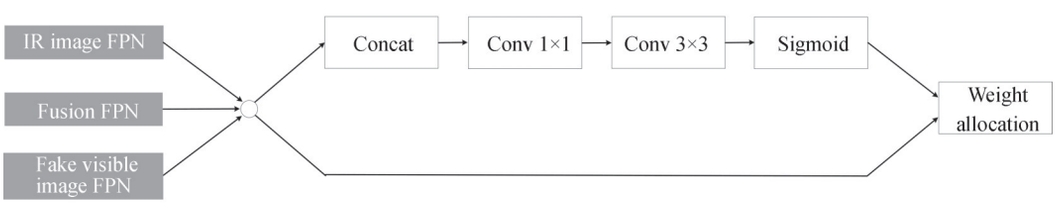


图3 软权重分配模块

Fig. 3 Soft weight allocation module

融合特征金字塔输出的特征输入RPN，由RPN生成候选锚框，然后利用softmax判断锚框属于前景还是背景。同时，另一条分支用于计算锚框的边界框回归偏移量，以获得更精确的感兴趣区域。如(5)式所示，(x, y)和(x_a, y_a)分别是真实标签和预测标签的位置信息，(w, h)和(w_a, h_a)分别是真实标签回归框和预测回归框的大小信息。(t_x, t_y)和(t_w, t_h)分别是位置信息的偏移量和回归框大小的尺度因子。RPN生成的区域映射到通过软权重分配模块得到的特征图上，将这些区域裁剪到相同尺寸。最后，利用全连接操作进行目标分类和定位。

$$\begin{cases} t_x=(x-x_c)/w_c \\ t_y=(y-y_a)/h_a \\ t_w=\log(w-w_a) \\ t_h=\log(h-h_a) \end{cases} \tag{5}$$

1.4 损失函数

基于跨域融合检测网络的损失函数包括原始损失和拼接信息损失两部分，每种损失再细分为分类损失和回归框损失。对于损失函数，分类层和回归层的输出分别由$\{p_i\}$和$\{t_i\}$表示，这两项又分别由N_{cls}和N_{reg}进行归一化，其中β为权重因子。在反向传播过程中，多分类和回归框的参数进行共享，旨在使不同的特征映射学习到标签信息外更多的语义信息。总损失函数定义为

$$\begin{aligned} L\left(\{p_i\},\{t_i\}\right)=&\frac{1}{N_{\text{cls},p}}\sum_i L_{\text{cls},p}\left(p_i,p_i^*\right)+\beta\frac{1}{N_{\text{reg},p}}\sum_i p_i^* L_{\text{reg},p}\left(t_i,t_i^*\right)+ \\ &\lambda\left[\frac{1}{N_{\text{cls},c}}\sum_i L_{\text{cls},c}\left(p_{c,i},p_i^*\right)+\beta\frac{1}{N_{\text{reg},c}}\sum_i p_i^* L_{\text{reg},c}\left(t_{c,i},t_i^*\right)\right] \end{aligned} \tag{6}$$

式中，$L_{\text{reg},p}$和$L_{\text{reg},p}$是原始损失函数，$L_{\text{cls},c}$和$L_{\text{reg},c}$是对拼接的特征所计算的损失函数。λ用于平衡拼接损失和原始损失的权重。i表示锚框的索引，p_i^*和t_i^*分别表示分类的真实标签和分类为真的边界框坐标向量的真实标签。p_i和t_i分别表示最终层对分类的预测和分类为真的边界框坐标向量的预测。$p_{c,i}$和$t_{c,i}$分别表示拼接的特征对分类的预测和分类为真的边界框的坐标向量的预测。其中p_i^*的定义为

$$P_i^*=\begin{cases} 1 & P^*>0 \\ 0 & P^*=0 \end{cases} \tag{7}$$

2 实验与分析

2.1 数据集和评价指标

实验部分采用FLIR ADAS[26]和SODA[27]红外数据集进行算法验证与比较。FLIR ADAS数据集中包含12 886幅尺寸为640×512的红外图像，SODA数据集采用其中1 500幅尺寸为640×480的红外图像。上述两个数据集总共包括31 242个person实例、61 763个car实例和4 757个bicycle实例。实验过程按照7∶2∶1的比例将数据集分为训练集、验证集和测试集。此外，为了研究伪可见光图像的有效性，利用CUT转换网络对FLIR数据集进行模态转换，得到FLIR伪可见光的数据集(Pseudo-visible FLIR)。图4为FLIR数据集和Pseudo-visivle FLIR数据集的对比图像。

图4 FLIR数据集和FLIR数据集生成的伪可见光数据集

Fig. 4 FLIR and Pseudo-visible FLIR

特征提取网络选用Resnet101-FPN，相关参数的数量和网络的结构与Resnet101-FPN相同。以$\{P_{2\text{I}},P_{3\text{I}},P_{4\text{I}},P_{5\text{I}},P_{6\text{I}}\}$举例说明，特征提取的输出层尺寸分别为{(256,160,128),(256,80,64),(256,40,32),(256,20,16),(256,10,8)}。网络经过图像融合、特征拼接后送入检测器的特征维度为(768,10,8)，网络设

置由1 024个神经元组成的全连接层，将特征空间通过线性变换映射到样本标记空间。在预测过程中，以(1 024，1，1)的维度送入分类支路和回归支路完成检测。所有结果按照不同类别进行指标计算，采用的指标包括两个数据集中各类别准确率AP、平均类别准确率mAP以及运算速度FPS。

2.2 实验细节

实验硬件平台配置为CPU：Inteli7-9700F，GPU：NvidiaGTX2080Ti，cuda10.0和cuDNN7.6.5，内存16G，显卡内存11G。输入图像的大小调整为640×512。默认情况下，网络用GPU训练模型12 000个Epoch，并设置每个Epoch 300个Step。初始学习率设置为0.001，权重衰减率设置为0.000 1。Batch size大小设置为16。在模型迭代开始时，分别初始化式(4)中的α和γ为0.5。选取规则是希望平均化不同域各个特征的影响，便于模型收敛，在每轮训练的验证阶段，α和γ所影响的拼接特征空间将与样本标记空间进行比对，并在新的一轮迭代中以软权重分配的方式对这两个参数进行优化。经过反复的实验证明，初始化阶段两个参数未能平均分配会导致其梯度下降缓慢，甚至造成梯度消失，以至于模型未能学习到某种通道有用的特征。所以在没有特别说明的情况下式(4)中的α和γ分别初始化为0.5。经过大量实验反复论证，式(6)中的λ设置为0.25能够使网络在降低损失方面有更好的表现。

2.3 消融实验

为了分析CFN中每个模块的有效性，消融实验部分逐一地将模态转换网络、跨域融合网络和软权重分配模块应用于模型中，以验证这些模块的有效性。同时，还分析了不同模块的组合所带来的改进，消融实验的基线方法baseline采用了ResNet101-FPN为主干网的Faster R-CNN目标检测网络。消融实验结果如表1所示。

表1 消融实验

Table 1 Ablation study

Modal transfer network	Cross-domain fusion networks	Soft weight allocation module	mAP/%
×	×	×	78
√	×	√	80.3
√	√	×	81.8
√	√	√	**86.4**

设置三组对比实验分别测试不同模块组合的有效性。首先，当模态转换网络与跨域融合网络相结合，采用FPN在多尺度特征图中进行特征融合，利用特征金字塔表达特征，同时也补充了不同尺度特征图上的语义信息，检测结果中mAP比baseline提高了1.4%。然后，将模态转换网络和软权重分配模块相结合，通过本方法的特征提取阶段分别对两个域的图像进行特征提取，不经过融合直接在软权重分配模块中对两个域的特征图自适应分配和优化参数。上述组合检测结果中mAP提升了1.7%。最后，网络将模态转换网络、跨域融合网络和软权重分配模块三者结合，自适应权重分配使得拼接权重能够弥补由于降维导致缺失的空间细节。所提方法的完整检测网络的平均准确率达到86.4%，相比于baseline mAP提升了3.5%。上述结果表明，CFN模块相互之间能够协同作用，同时能够有效地提高检测性能。

2.4 伪可见光有效性验证

为了测试伪可见光图像的有效性，实施对比实验。实验选用的网络为Mask-RCNN，设置backbone为Resnet101-FPN，训练参数均不做改变，仅改变数据集。由表2可知，针对Mask-RCNN网络改变数据集训练得到的模型，在平均准确率mAP上并没有太大差别，然而，数据集由FLIR更换为Pseudo-visible FLIR，模型的准确率发生了一些变化，从64%下降到了61.7%，召回率从90.2%上升到了92.1%。分析可知，Pseudo-visible FLIR数据集针对一些特定条件下的目标，与红外数据集相比，FLIR提高了检出能力，而相应

表2 相同网络在不同数据集下的性能对比

Table 2 Performance comparison of same network under different datasets

Methods	Backbone	Datasets	Precision	Recall	F1-Score	mAP/%
Mask R-CNN	Resnet101-FPN	FLIR	0.640	0.902	0.748	84.0
		Pseudo Visible FLIR	0.617	0.921	0.738	83.3

的对于疑似该类目标的物体误检概率也会升高。

观察两类数据集的预测结果，如图5所示，当目标和背景的灰度值接近或多个目标重叠在一起，红外图像很容易造成漏检。这是由于红外图像本身的缺陷造成的，对于热度区分度低的目标和背景，红外图像难以对轮廓有准确的成像。然而，在Pseudo-visible FLIR中，这类问题得到了很好的解决，模态转换网络利用对比损失，使得两个域的图像拥有最大化的互信息。在有了RGB信息辅助的情况下，网络能够更加轻松的判断轮廓位置和形状，由此提升检测精度。

图5 FLIR和Pseudo-visible FLIR数据集的检测结果

Fig. 5 Detection results of FLIR and Pseudo-visible FLIR dataset

2.5 性能比较与分析

将本方法与目前主流的目标检测网络进行性能比较。为了实验的公平性，实验为所有的双阶段类型网络搭载相同的骨干网ResNet101-FPN，单阶段网络YOLO-V4的骨干网选择性能与之相仿的CSPDarknet-53。对比试验的基线方法采用Mask R-CNN检测网络。没有特别说明的情况下，测试遵循mmdetection的超参数设置。并用迭代过程中损失最低的模型进行推理和预测。

如表3所示，首先，针对FLIR数据集采用经典的Faster R-CNN网络进行检测，mAP达到78.9%。YOLO-V4作为最新一批的单阶段检测网络，其平均准确率为77.1%，略低于Faster R-CNN。其次，MMTOD是针对红外目标检测的网络，其网络依赖双通道的输入方式能够在精度上得到极大的提升，mAP达到了83%。CFN在所选取的Baseline的基础上提高了一定的检测精度，mAP达到了84.2%。提出的方法对于车、人和自行车实例都体现出了优越性，并且平均准确率达到了87.7%，与Faster R-CNN相比，提升了8.8%的mAP；同时与同类型的MMTOD红外目标检测网络相比，提升了4.7%的mAP。在处理速度上，YOLO-V4的单阶段模型有明显的优越性，其推理速度达到了每秒39帧，而MMTOD网络处理速率低下，难以满足实

表3 不同网络在不同数据集下的性能对比

Table 3 Performance comparison of different networks under different datasets

Methods	Backbone	Datasets	AP/%			mAP/%	FPS
			Car	Person	Bicycle		
Faster R-CNN	Resnet101-FPN	FLIR	83.7	78.3	74.7	78.9	17
		SODA	79.3	78.5	73.4	77.1	
YOLO-V4	CSPDarknet-53	FLIR	80	78.8	75.7	78.1	**39**
		SODA	82.7	78.5	70.1	77.1	
MMTOD	Resnet101-FPN	FLIR	86.8	83.1	79.1	83	10
		SODA	84.2	81.4	76	80.5	
Baseline	Resnet101-FPN	FLIR	88.3	84.6	79.6	84.2	14
		SODA	84.8	83.6	76.5	81.6	
Ours	Resnet101-FPN	FLIR	91.3	88.5	83.2	**87.7**	16
		SODA	88	86.9	80.4	**85.1**	

时性，CFN在保证了检测精度的同时与MMTOD相比具有更优的推理速度，达到16帧每秒，几乎满足了实时性。

除FLIR数据集之外，在SODA数据集上做了相同的实验，实验结果显示提出的检测框架在保持相同数量级检测速度情况下，达到85.1%的mAP，与对比方法相比，展现出了优越性，从而验证了本网络的鲁棒性和泛化能力。为了更直观的体现本文方法的检测性能，对两个数据集所测得的mAP进行平均处理，并将得到的均值绘制PR曲线，结果如图6所示。

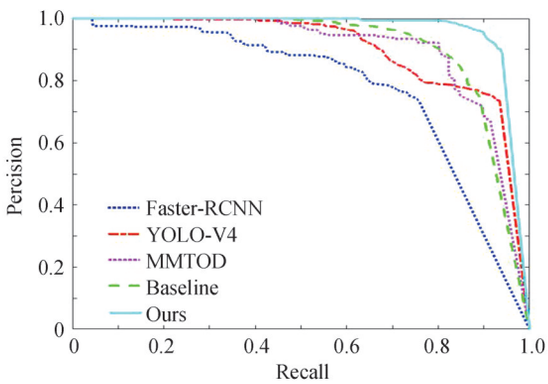
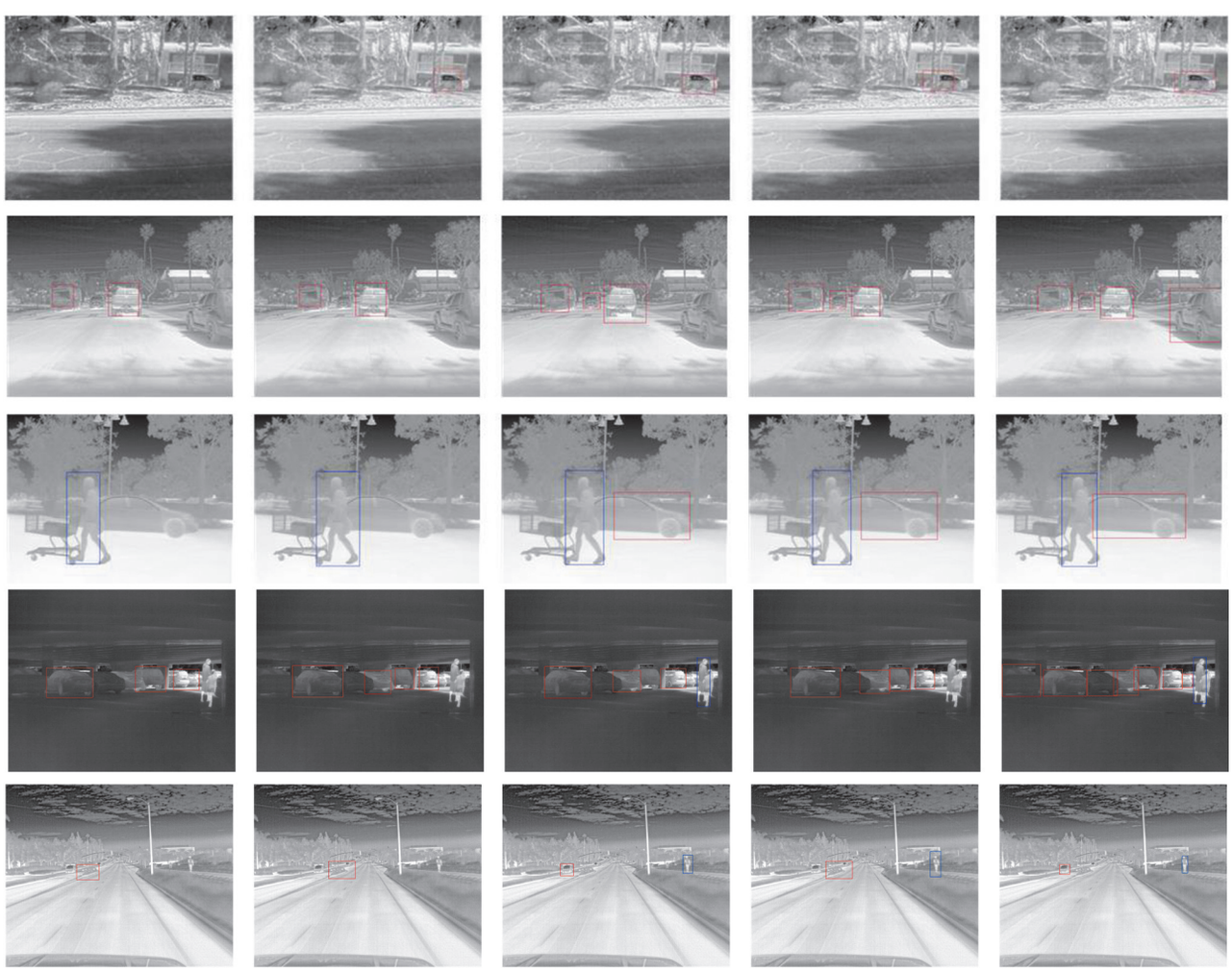


图6　网络之间的PR曲线比较

Fig. 6　Comparison of PR curves of various networks

图7和图8分别展示FLIR ADAS数据集和SODA数据集有代表性的检测结果图。与其他方法相比，针对自然光图像提出的检测网络难以适应复杂环境下的红外目标检测，对边缘目标和遮挡目标没有较好的适应性。同时，针对小目标存在较多的误检和漏检情况。CFN对于各个尺度的目标检测都有较大的改善，尤其是针对复杂环境展现出了较好的鲁棒性。



(a) Faster R-CNN　(b) YOLO-V4　(c) MMTOD　(d) Baseline　(e) Ours

图7　FLIR数据集的检测对比效果

Fig. 7　Comparison of detection effect on FLIR dataset

图8 SODA数据集的检测对比效果

Fig. 8 Comparison of detection effect on SODA dataset

3 结论

提出了一种跨域融合网络结构，采用模态转换网络生成伪可见光图像，对红外图像缺失的纹理信息进行补偿，并使用软权重分配模块以动态权重的形式去优化网络检测效果，最后通过全连接层完成目标的分类和定位。实验结果表明，在数据集FLIR ADAS和SODA测试过程中，网络均展示出了更高的准确度，与Faster R-CNN、YOLO-V4、MMTOD、Baseline进行比较，验证了方法的有效性。

参考文献

[1] GAO J, GUO Y, LIN Z, et al. Infrared small target detection using multiscale gray and variance difference[C]. Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, Cham, 2018: 53-64.

[2] YI Xiang, WANG Bingjian. Fast infrared and dim target detection algorithm based on multi-feature[J]. Acta Photonica Sinica, 2017, 46(6): 0610002.
易翔, 王炳健. 基于多特征的快速红外弱小目标检测算法[J]. 光子学报, 2017, 46(6): 0610002.

[3] ZHU Guoqiang, MENG Xiangyong, QIAN Weixian. Infrared small target detection method based on curvature near the ground[J]. Acta Photonica Sinica, 2018, 47(10): 1010001.
朱国强, 孟祥勇, 钱惟贤. 基于曲率的近地面红外小目标检测算法[J]. 光子学报, 2018, 47(10): 1010001.

[4] GAO C, MENG D, YANG Y, et al. Infrared patch-image model for small target detection in a single image[J]. IEEE Transactions on Image Processing, 2013, 22(12): 4996-5009.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25(12): 1097-1105.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic

segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.
[7] GIRSHICK R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2015: 1440-1448.
[8] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[DB/OL]. https://arxiv.org/abs/1506.01497.
[9] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]. European Conference on Computer Vision, Springer, Cham, 2016: 21-37.
[10] REDMON J, FARHADI A. Yolov3: An incremental improvement[DB/OL]. https://arxiv.org/abs/1804.02767.
[11] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[DB/OL]. https://arxiv.org/abs/2004.10934.
[12] BAEK J, HONG S, KIM J, et al. Efficient pedestrian detection at nighttime using a thermal camera[J]. Sensors, 2017, 17(8): 1850.
[13] 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(10): 261-270.
[14] 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: 99870H.
[15] BERG A. Detection and tracking in thermal infrared imagery[D]. Linköping University Electronic Press, 2016.
[16] BERG A, ÖFJÄLL K, AHLBERG J, et al. Detecting rails and obstacles using a train-mounted thermal camera[C]. Scandinavian Conference on Image Analysis, Springer, Cham, 2015: 492-503.
[17] LEYKIN A, RAN Y, HAMMOUD R. Thermal-visible video fusion for moving target tracking and pedestrian classification[C]. 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007: 1-8.
[18] WAGNER J, FISCHER V, HERMAN M, et al. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks[C]. ESANN, 2016, 587: 509-514.
[19] 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), IEEE, 2016: 621-626.
[20] GHOSE D, DESAI S M, BHATTACHARYA S, et al. Pedestrian detection in thermal images using saliency maps[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019: 20-30.
[21] DEVAGUPTAPU C, AKOLEKAR N, SHARMA MM, et al. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019: 15-23.
[22] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 2117-2125.
[23] PARK T, EFROS A A, ZHANG R, et al. Contrastive learning for unpaired image-to-image translation[C]. European Conference on Computer Vision, Springer, Cham, 2020: 319-345.
[24] 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: 2223-2232.
[25] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[26] GROUPF A. FLIR thermal dataset for algorithm training[DB/OL].[2018-07-26]. https://www.flir.in/oem/adas/adas-dataset-agree.
[27] LI C, XIA W, YAN Y, et al. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020 , 47(11): 1110001.

Foundation item: National Natural Science Foundation of Shanghai(No.20ZR1423500), the Open Project Program of the Key Laboratory of Intelligent Infrared Perception, Chinese Academy of Sciences(No.CAS-IIRP-04)