E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2024)08-2350-14

论文引用格式:Xu K, Liu X P, Wang H Y, Wan J W and Guo Y L. 2024. Infrared-visible image object detection algorithm using feature dynamic selection. Journal of Image and Graphics, 29(08):2350-2363(许可, 刘心溥, 汪汉云, 万建伟, 郭裕兰. 2024. 红外与可见光图像特征动态选择的目标检测网络. 中国图象图形学报, 29(08):2350-2363)[DOI:10.11834/jig.230495]

红外与可见光图像特征动态选择的目标检测网络

许可[1],刘心溥[1*],汪汉云[2],万建伟[1],郭裕兰[1]

1. 国防科技大学电子科学学院,长沙　410005; 2. 信息工程大学地理空间信息学院,郑州　450001

摘　要:**目的**　基于可见光和红外双模态图像融合的目标检测算法是解决复杂场景下目标检测任务的有效手段。然而现有双光检测算法中的特征融合过程存在两大问题:一是特征融合方式较为简单,逐特征元素相加或者并联操作导致特征融合效果不佳;二是算法结构中仅有特征融合过程,而缺少特征选择过程,导致有用特征无法得到高效利用。为解决上述问题,提出了一种基于动态特征选择的可见光红外图像融合目标检测算法。**方法**　本文算法包含特征的动态融合层和动态选择层两个创新模块:动态融合层嵌入在骨干网络中,利用Transformer结构,多次对多源的图像特征图进行特征融合,以丰富特征表达;动态选择层嵌入在颈部网络中,利用3种注意力机制对多尺度特征图进行特征增强,以筛选有用特征。**结果**　本文算法在FLIR、LLVIP(visible-infrared paired dataset for low-light vision)和VEDAI(vehicle detection in aerial imagery)3个公开数据集上开展实验验证,与多种特征融合方式进行平均精度均值(mean average precision,mAP)性能比较,mAP50指标相比于基线模型分别提升了1.3%、0.6%和3.9%;mAP75指标相比于基线模型分别提升了4.6%、2.6%和7.5%;mAP指标相比于基线模型分别提升了3.2%、2.1%和3.1%。同时设计了相关结构的消融实验,验证了所提算法的有效性。**结论**　提出的基于动态特征选择的可见光红外图像融合目标检测算法,可以有效地融合可见光和红外两种图像模态的特征信息,提升了目标检测的性能。

关键词:红外图像;目标检测;注意力机制;特征融合;深度神经网络

Infrared-visible image object detection algorithm using feature dynamic selection

Xu Ke[1], Liu Xinpu[1*], Wang Hanyun[2], Wan Jianwei[1], Guo Yulan[1]

1. College of Electronic Science and Technology, National University of Defense Technology, Changsha　410005, China;

2. School of Surveying and Mapping, Information Engineering University, Zhengzhou　450001, China

Abstract: **Objective** In recent years, considerable attention has been given to the object detection algorithm that utilizes the fusion of visible and infrared dual-modal images. This algorithm serves as an effective approach for addressing object detection tasks in complex scenes. The process of object detection algorithms can be roughly divided into three stages. The first stage is feature extraction, which aims to extract geometric features from the input data. Next, the extracted features are fed into the neck network for multi-scale feature fusion. Finally, the fused features are input into the detection network to output object detection results. Similarly, dual-modal detection algorithms follow the same process to achieve object

收稿日期:2023-07-19;**修回日期**:2023-10-23;**预印本日期**:2023-10-30

*通信作者:刘心溥　liuxinpu@yeah.net

基金项目:国家重点研发计划项目(2022YFB3902400)

Supported by: National Key R&D Program of China (2022YFB3902400)

localization and classification. The difference lies in the fact that traditional object detection focuses on single-modal visible images, while dual-modal detection focuses on visible and infrared image data. The dual-modal detection algorithm aims to simultaneously utilize information from infrared and visible images. It merges these images to obtain more comprehensive and accurate target information, which enhances the accuracy and robustness of the object detection process. Traditional fusion methods encompass pixel-level fusion and feature-level fusion. Pixel-level fusion employs a straightforward weighted overlay technique on the two types of images, which enhances the contrast and edge information of the targets. Meanwhile, feature-level fusion extracts features from the infrared and visible images and combines them to enhance the representation capability of the targets. However, the feature fusion process of existing dual-modal detection algorithms faces two major issues. First, the feature fusion methods employed are relatively simple, which involves the addition or parallel operation of individual feature elements. Consequently, these methods yield unsatisfactory fusion effects that limit the performance of subsequent object detection. Second, the algorithm structure solely focuses on the feature fusion process, which neglects the crucial feature selection process. This deficiency results in the inefficient utilization of valuable features. **Method** In this study, we introduce a visible and infrared image fusion object detection algorithm that employs dynamic feature selection to address the two issues mentioned above. Overall, we propose enhancements to the conventional YOLOv5 detector through modifications to its backbone, neck, and detection head components. We select CSPDarkNet53 as the backbone, which possesses an identical structure for visible and infrared image branches. The algorithm incorporates two innovative modules: dynamic fusion layer and dynamic selection layer. The proposed algorithm includes embedding the dynamic fusion layer in the backbone network, which utilizes the Transformer structure for multiple feature fusions in multi-source image feature maps to enrich feature expression. Moreover, it employs the dynamic selection layer in the neck network, which uses three attention mechanisms (i. e. , scale, space, and channel) to improve multi-scale feature maps and screen useful features. These mechanisms are implemented with SENet and deformable convolutions. In line with standard practices in target detection algorithms, we utilize the detection head of YOLOv5 to generate detection results. The loss function employed for algorithm training is the combined sum of bounding box regression loss, classification loss, and confidence loss, which are implemented with generalized intersection over union, cross entropy, and squared-error functions, respectively. **Result** In this study, we validate our proposed algorithm through experimental evaluation on three publicly available datasets: FLIR, visible-infrared paired dataset for low-light vision (LLVIP), and vehicle detection in aerial imagery (VEDAI). We use the mean average precision (mAP) for evaluation. Compared with the baseline model that adds features individually, our algorithm achieves improvements of 1. 3%, 0. 6%, and 3. 9% in mAP50 scores and 4. 6%, 2. 6%, and 7. 5% in mAP75 scores. In addition, our algorithm demonstrates enhancements of 3. 2%, 2. 1%, and 3. 1% in mAP scores on the respective datasets, which effectively reduces the probability of object omission and false alarms. Moreover, we conduct ablation experiments on two innovative modules: the dynamic fusion layer and the dynamic selection layer. The complete algorithm model, which incorporates the two layers, achieves the best performance on all three test datasets. This performance validates the effectiveness of our proposed algorithm. We also compare the network model size and computational efficiency of these state-of-the-art algorithms, and experiments show that our algorithm can significantly improve algorithm performance while slightly increasing parameter computation. Furthermore, we visualize the attention weight matrices of the three dynamic fusion layers in the backbone to better reveal the mechanism of the dynamic fusion layer. The visual analysis confirms that the dynamic fusion layer effectively integrates the feature information from visible and infrared images. **Conclusion** In this study, we propose a visible and infrared image fusion-based object detection algorithm using dynamic feature selection strategy. This algorithm incorporates two innovative modules: dynamic fusion layer and dynamic selection layer. Through extensive experiments, we demonstrate that our algorithm effectively integrates feature information from visible and infrared image modalities, which enhances the performance of object detection. However, the proposed algorithm has a little increasing computational complexity and requires pre-registration of the input visible and infrared images, which limits some application scenarios of the algorithm. The research on lightweight fusion modules and algorithms capable of processing unregistered dual light images will be the focus of future research in the field of multimodal fusion target detection.

Key words: infrared image; object detection; attention mechanism; feature fusion; deep neural network

0 引 言

在实际的二维目标检测任务中，检测场景往往较为复杂，需要模型和算法来应对多种挑战，如雨、雾、遮挡、光照不良和低分辨率等。仅使用可见光相机传感器数据在这些条件下往往不能实现高精度的目标检测。随着红外图像传感器的加入，多光谱成像技术更为重要。融合不同模态的互补特征可以提高检测算法的准确性、可靠性和稳健性。

可见光与红外相机成像质量对比如图1所示，两种传感器有其各自的优势和劣势：可见光优势在于其具有更多的纹理特征和细节表现能力，利于目标的分类和定位，如图1(a)，远处的车辆在可见光成像中比红外成像中更易区分，但可见光的劣势在于其在低光照条件下的表现较差，如图1(b)，处于黑暗中的行人无法探测到，或者易受图1(c)中的强光干扰；红外图像的优势在于其在低光照条件下的表现优异，如图1(b)，红外图像下处于黑暗中的行人清晰可见，但红外图像的劣势在于其易受热源干扰出现热噪声，如图1(d)，出现了红外“鬼影”，易造成虚警现象。综上可知，融合可见光和红外图像两种模态的特征信息，进行双光检测是必要的。

目标检测算法流程大致分为3个阶段。首先是特征提取阶段，其目的是提取输入数据的几何特征。接着，将提取的特征送入颈部网络，进行多尺度特征融合。最后，将融合后的特征输入检测网络，以输出目标检测结果。同理，双光检测算法也是基于上述的流程，实现目标的定位与分类。不同点在于，传统的目标检测聚焦于单模态的可见光图像，而双光检测聚焦于可见光和红外两种图像数据，因此双光检测任务的关键在于如何进行有效而高效的多源图像特征融合。在以往的研究中，He等人(2016)、Ren等人(2017)、Li等人(2018)、Zhang等人(2019)、Zhang等人(2021)和邱德粉等人(2023)提出的大部分多源特征融合机制都基于深度卷积神经网络(convolutional neural network，CNN)，CNN在单模态推理中具有强大的表示学习能力，特别是在视觉模态方面。然而，将其扩展到跨模态特征融合方式，以充分利用多源特征的互补性并不容易。因为CNN的卷积操作具有非全局感受野，特征信息的交互在局部区域内完成，其不利于多源特征信息进行充分的融合。与此相比，Transformer(Vaswani等，2017)的自注意力操作具有全局感受野，网络通过学习特征的长距离依赖关系，可以很好地建模特征之间的相关性强弱，已经在图像分类(Dosovitskiy等，2021)、目标检测(Carion等，2020；Wang等，2022；孙旭辉 等，2023；别倩 等，2023)和三维点云处理(Guo等，2021；刘心溥 等，2022)等多个视觉领域取得了不错的效果，因此基于Transformer结构的融合算法相比于CNN

(a) 可见光图像纹理细节　(b) 红外图像夜间成像　(c) 可见光图像强光模糊　(d) 红外图像热噪声

图1　可见光与红外相机成像质量对比图

Fig. 1　Comparison of imaging quality between visible and infrared cameras

((a) visible image texture details; (b) infrared image night imaging; (c) visible image light blur; (d) infrared image thermal noise)

结构更适合于双光检测任务。Fu 等人(2023)进一步将 Swin-Transformer 结构引入双光检测任务中，在高效计算的同时，较好地对双光特征进行了融合。

双光检测任务的重点是特征融合部分的算法设计。然而现有的特征融合模块存在两方面的问题：一是特征融合方式较为简单，逐特征相加或者并联操作，并且仅在颈部网络模块中融合特征，导致特征融合不够充分；二是特征融合模块仅有特征融合过程，缺少特征选择过程，导致特征融合效果不佳。为此，针对上述两个问题，本文设计了两个算法模块，分别是动态融合层(dynamic fusion，DyFu)和动态选择层(dynamic selection，DySe)，动态意为不断更新双光图像的特征矩阵。动态融合层 DyFu 主要集中在骨干网络中，利用 Transformer 机制，多次对多源的特征图进行特征融合，以丰富特征表达；动态选择层 DySe 嵌入到颈部网络中，利用 3 种注意力机制对多尺度特征图进行特征增强，以筛选有用特征。本文算法在 FLIR、LLVIP(visible-infrared paired dataset for low-light vision)和 VEDAI(vehicle detection in aerial imagery) 3 个公开数据集上进行对比实验，并设计了相关的消融实验，验证了所提算法的有效性。

本文贡献如下：1)提出了一种基于动态特征选择的可见光红外图像融合目标检测算法，将多源的特征图按照先“融合”后“选择”的范式进行特征增强，有效地利用了双光信息，提升了目标检测的性能。2)设计了动态融合层和动态选择层两个算法模块，动态融合层利用 Transformer 机制，对多源的特征图进行特征融合；动态选择层利用 3 种注意力机制对特征图进行特征增强。3)本文算法在 FLIR、LLVIP 和 VEDAI 3 个公开数据集上开展实验验证，mAP75 指标相比于基线模型分别提升了 4.6%、2.6% 和 7.5%，性能达到领域领先水平。

1 整体算法流程

基于动态特征选择的可见光红外图像融合目标检测算法的整体结构如图 2 所示。

整体来看，本文算法在传统的 YOLOv5(you only look once v5)检测器的基础上进行改进，同样由骨干网络(backbone)、颈部网络(neck)和检测头(detection head) 3 部分组成。骨干网络选用 CSP-DarkNet53，可见光图像和红外图像两条支路的结构

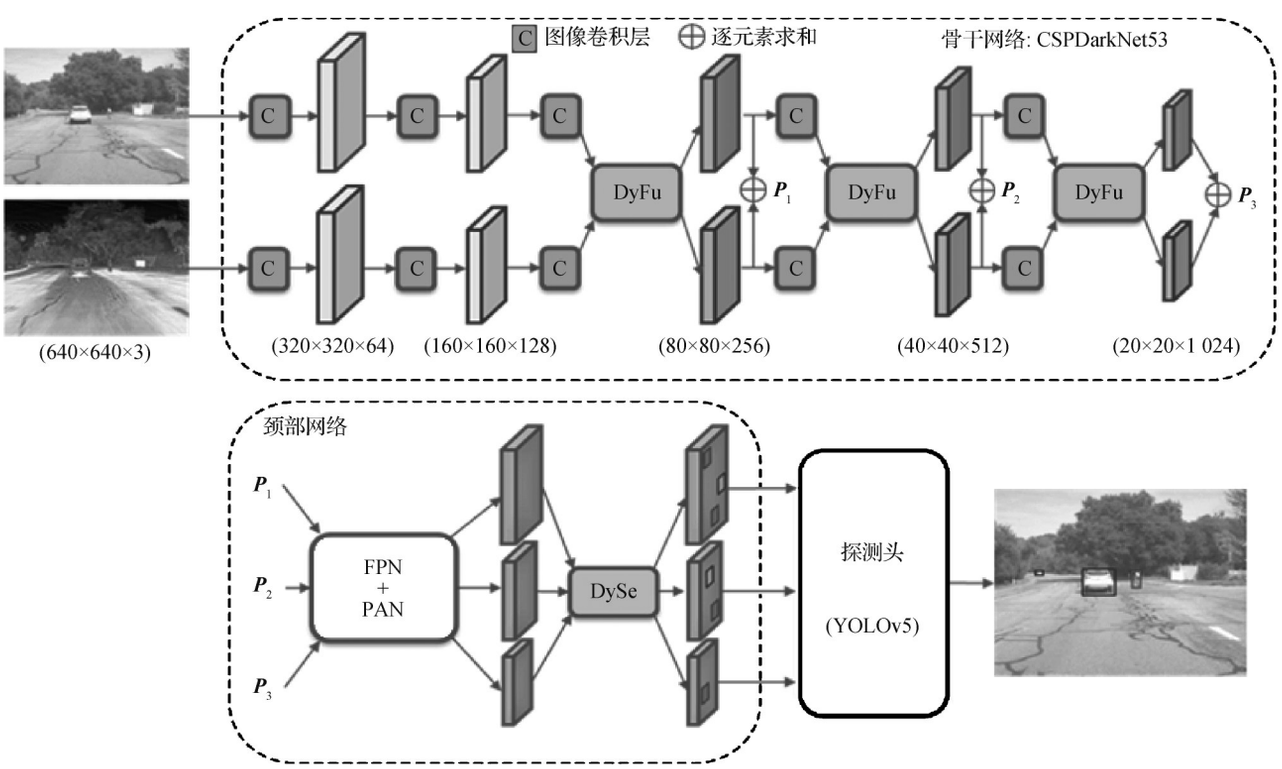


图 2　基于动态特征选择的可见光红外图像融合目标检测算法的整体结构图

Fig. 2　Overall structure of visible-infrared image fusion object detection based on dynamic feature selection

相同，本文算法在两条支路中间嵌入了3个动态融合层DyFu，通过对两种图像的特征图进行特征融合，以更新特征表示，具体为

$$\boldsymbol{F}'_{\mathrm{RGB}}, \boldsymbol{F}'_{\mathrm{IR}} = DyFu\left(\boldsymbol{F}_{\mathrm{RGB}}, \boldsymbol{F}_{\mathrm{IR}}\right) \tag{1}$$

式中，$\boldsymbol{F}_{\mathrm{RGB}}$和$\boldsymbol{F}_{\mathrm{IR}}$分别为特征更新前的可见光和红外图像特征，$\boldsymbol{F}'_{\mathrm{RGB}}$和$\boldsymbol{F}'_{\mathrm{IR}}$分别为特征更新后的可见光和红外图像特征，$DyFu$表示动态融合层。然后，后3层的特征图分别相加为融合特征图$\boldsymbol{P}_1$、$\boldsymbol{P}_2$和$\boldsymbol{P}_3$，送入颈部网络中。

在颈部网络部分，3个融合特征图输入到具有特征金字塔网络（feature pyramid network，FPN）（Lin等，2017）结构的路径聚合网络（path aggregation network，PAN）（Liu等，2018）中，生成多尺度特征图$\boldsymbol{F}_{\mathrm{U1}}$、$\boldsymbol{F}_{\mathrm{U2}}$和$\boldsymbol{F}_{\mathrm{U3}}$，随后将其输入到所提出的动态选择层DySe中，利用3种注意力机制对多尺度特征图进行特征增强，生成增强特征图$\boldsymbol{F}'_{\mathrm{U1}}$、$\boldsymbol{F}'_{\mathrm{U2}}$和$\boldsymbol{F}'_{\mathrm{U3}}$，具体为

$$\boldsymbol{F}_{\mathrm{U1}}, \boldsymbol{F}_{\mathrm{U2}}, \boldsymbol{F}_{\mathrm{U3}} = PAN_{\mathrm{FPN}}\left(\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3\right) \tag{2}$$

$$\boldsymbol{F}'_{\mathrm{U1}}, \boldsymbol{F}'_{\mathrm{U2}}, \boldsymbol{F}'_{\mathrm{U3}} = DySe\left(\boldsymbol{F}_{\mathrm{U1}}, \boldsymbol{F}_{\mathrm{U2}}, \boldsymbol{F}_{\mathrm{U3}}\right) \tag{3}$$

式中，PAN_{FPN}表示具有特征金字塔结构的路径聚合网络，$DySe$表示动态选择层。最后，增强特征图$\boldsymbol{F}'_{\mathrm{U1}}$、$\boldsymbol{F}'_{\mathrm{U2}}$和$\boldsymbol{F}'_{\mathrm{U3}}$输入到YOLOv5的检测头中，输出检测结果。本文两个创新算法模块的动态融合层DyFu和动态融合层DySe将分别在第2节和第3节中介绍。

2 动态融合层DyFu

动态融合层DyFu的目的是多次对可见光和红外图像的特征图进行特征融合，以丰富特征表达，其算法结构图如图3所示。

受到多模态聚合Transformer（cross modality fusion Transformer，CFT）（Fang等，2022）工作的启发，本文在骨干网络中间得到可见光和红外图像的特征图$\boldsymbol{F}_{\mathrm{RGB}}$和$\boldsymbol{F}_{\mathrm{IR}}$之后，将其展平后并联拼接起来，得到拼接特征向量$\boldsymbol{I}_{\mathrm{FU}}$，具体为

$$\boldsymbol{I}_{\mathrm{FU}} = f_{\mathrm{Concat}}\left(f_{\mathrm{flatten}}\left(\boldsymbol{F}_{\mathrm{RGB}}\right), f_{\mathrm{flatten}}\left(\boldsymbol{F}_{\mathrm{IR}}\right)\right) \tag{4}$$

式中，f_{Concat}和f_{flatten}分别代表并联和展平操作，然后依次通过3个多层感知机生成对应的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$矩阵，后输入到多头注意力模块，利用Transformer结构，计算特征融合后的交互特征向量$\boldsymbol{I}_{\mathrm{mid}}$，具体为

$$\begin{cases} \boldsymbol{Q} = MLP_Q\left(\boldsymbol{I}_{\mathrm{FU}}\right) \\ \boldsymbol{K} = MLP_K\left(\boldsymbol{I}_{\mathrm{FU}}\right) \\ \boldsymbol{V} = MLP_V\left(\boldsymbol{I}_{\mathrm{FU}}\right) \end{cases} \tag{5}$$

$$\boldsymbol{I}_{\mathrm{mid}} = MHA\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \boldsymbol{\alpha V} = f_{\mathrm{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{6}$$

式中，MLP表示多层感知机，MHA表示多头注意力层，注意力权重矩阵$\boldsymbol{\alpha}$的示意图如图4所示，本文采用先拼接后计算注意力权重的方式，可以实现同时进行4个关于图像特征的相关运算，按照图4中的区域顺序分别划分为：可见光图像特征的自相关、可见光/红外图像特征的互相关、红外/可见光图像特征的互相关和红外图像特征的自相关。通过上述图像特征的相关运算，可以更好地量化两种图像模态特征矩阵$\boldsymbol{F}_{\mathrm{RGB}}$和$\boldsymbol{F}_{\mathrm{IR}}$之间的特征相关性。

最后，交互特征向量$\boldsymbol{I}_{\mathrm{mid}}$经归一化和多层感知机层后，与前面的变量构成残差连接，再按照不同图像模态分离特征，生成更新后的特征图$\boldsymbol{F}'_{\mathrm{RGB}}$和$\boldsymbol{F}'_{\mathrm{IR}}$，具体为

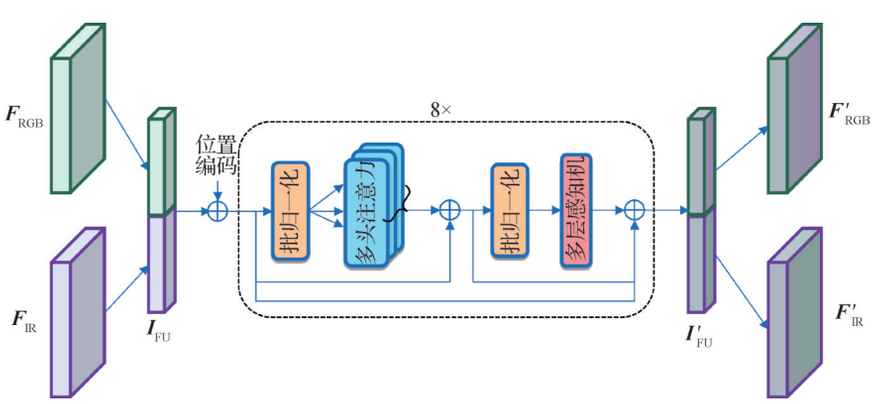


图3 动态融合层算法结构图

Fig. 3 Structure of dynamic fusion layer

$$\boldsymbol{F}_{\mathrm{RGB}},\boldsymbol{F}_{\mathrm{IR}}=f_{\mathrm{Split}}\left(\boldsymbol{I}_{\mathrm{mid}}+\boldsymbol{I}_{\mathrm{FU}}+MLP\left(BN\left(\boldsymbol{I}_{\mathrm{mid}}\right)\right)\right) \tag{7}$$

式中，f_{Split}表示分离操作，BN表示批归一化层。值得注意的是，本文算法模型中的多头注意力模块重复了8次。

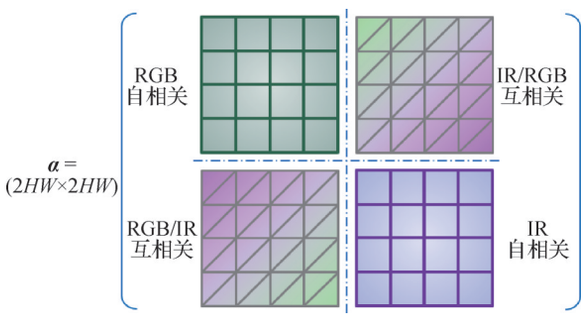


图4　注意力权重矩阵示意图

Fig. 4　Schematic diagram of attention weight matrix

3　动态选择层DySe

动态选择层DySe的目的是利用注意力机制对多尺度特征图进行特征增强，以筛选有用特征。其算法结构图如图5所示。

在得到多尺度特征图$\boldsymbol{F}_{\mathrm{U1}}$、$\boldsymbol{F}_{\mathrm{U2}}$和$\boldsymbol{F}_{\mathrm{U3}}$之后，均送入多层感知机层进行降维，使得特征维度统一为128。然后为了统一特征图的尺寸，$\boldsymbol{F}_{\mathrm{U2}}$和$\boldsymbol{F}_{\mathrm{U3}}$经过双线性插值上采样后与$\boldsymbol{F}_{\mathrm{U1}}$并联，得到多尺度聚合特征$\boldsymbol{F}_{\mathrm{U}}$，此时$\boldsymbol{F}_{\mathrm{U}}$的特征维度为$(3\times80\times80\times128)$，具体为

$$\boldsymbol{F}_{\mathrm{U}}=f_{\mathrm{Concat}}\left(MLP\left(\boldsymbol{F}_{\mathrm{U1}}\right),up\left(MLP\left(\boldsymbol{F}_{\mathrm{U2}}\right)\right),up\left(MLP\left(\boldsymbol{F}_{\mathrm{U3}}\right)\right)\right) \tag{8}$$

式中，up表示上采样操作，3代表3个特征尺度，用于检测不同尺寸的目标，80×80代表特征图尺寸，128代表特征通道数。

然后，将$\boldsymbol{F}_{\mathrm{U}}$依次通过尺度、空间和通道注意力层，分别更新3、$80\times80$和128维度特征，具体为

$$\boldsymbol{F}'_{\mathrm{U}}=Cha\left(Spa\left(Sca\left(\boldsymbol{F}_{\mathrm{U}}\right)\right)\right) \tag{9}$$

式中，Cha、Spa和Sca分别表示通道注意力、空间注意力和尺度注意力层。最后按照尺度维度进行切片解耦，后经下采样得到增强特征图$\boldsymbol{F}'_{\mathrm{U1}}$、$\boldsymbol{F}'_{\mathrm{U2}}$和$\boldsymbol{F}'_{\mathrm{U3}}$，具体为

$$\begin{cases}\boldsymbol{F}'_{\mathrm{U1}}=\boldsymbol{F}'_{\mathrm{U}}[0,:,:,:]\\ \boldsymbol{F}'_{\mathrm{U2}}=down\left(\boldsymbol{F}'_{\mathrm{U}}[1,:,:,:]\right)\\ \boldsymbol{F}'_{\mathrm{U3}}=down\left(\boldsymbol{F}'_{\mathrm{U}}[2,:,:,:]\right)\end{cases} \tag{10}$$

式中，$down$表示下采样操作。

采用改进的SENet（squeeze-and-excitation network）（Hu等，2018）结构来实现尺度注意力和通道注意力过程。具体来看，首先将尺寸为(N,C)的特征矩阵$\boldsymbol{F}$压缩到通道描述符中，这是通过使用全局平均池化生成通道统计信息来实现的，其中N可以代表多个特征维度。然后将描述符输入到具有瓶颈结构的多层感知机层，以学习每个通道的注意力权重，最后通过原特征矩阵$\boldsymbol{F}$和通道注意力权重之间的点积运算生成加权特征矩阵$\boldsymbol{F}_{\mathrm{CHA}}$，具体为

$$\boldsymbol{F}_{\mathrm{CHA}}=Cha(\boldsymbol{F})=\boldsymbol{F}\odot FC(\sum_{i=1}^{N}F(i,j)/N) \tag{11}$$

式中，FC表示全连接层，$\odot$表示矩阵的哈达玛积。

本文采用可变形卷积（deformable convnet，DCN）（Zhu等，2019）结构来实现空间注意力过程。具体来看，如图5所示，DCN可以根据图像特征学习到一组卷积偏移量，从而作用于可变形卷积核上，实现动态地提取卷积核周围的局部区域特征，起到像素空间的注意力效果。

4　损失函数

本文算法整体的损失函数由3部分组成，分别是边界框回归损失L_{box}、分类损失L_{cls}和置信度损失L_{conf}，与CFT（Fang等，2022）论文保持一致。

边界框回归损失L_{box}采用GIOU（generalized intersection over union）函数计算，分类损失L_{cls}采用交叉熵的计算形式，置信度损失L_{conf}采用均方误差来计算。

5　实验结果与分析

将所提出的基于动态特征选择的可见光红外图像融合目标检测算法，分别在FLIR（Saga等，2022）、LLVIP（Jia等，2021）和VEDAI（Razakarivony和Jurie，2016）3个公开数据集上进行性能评估并做消融实验，以验证所提算法的有效性。

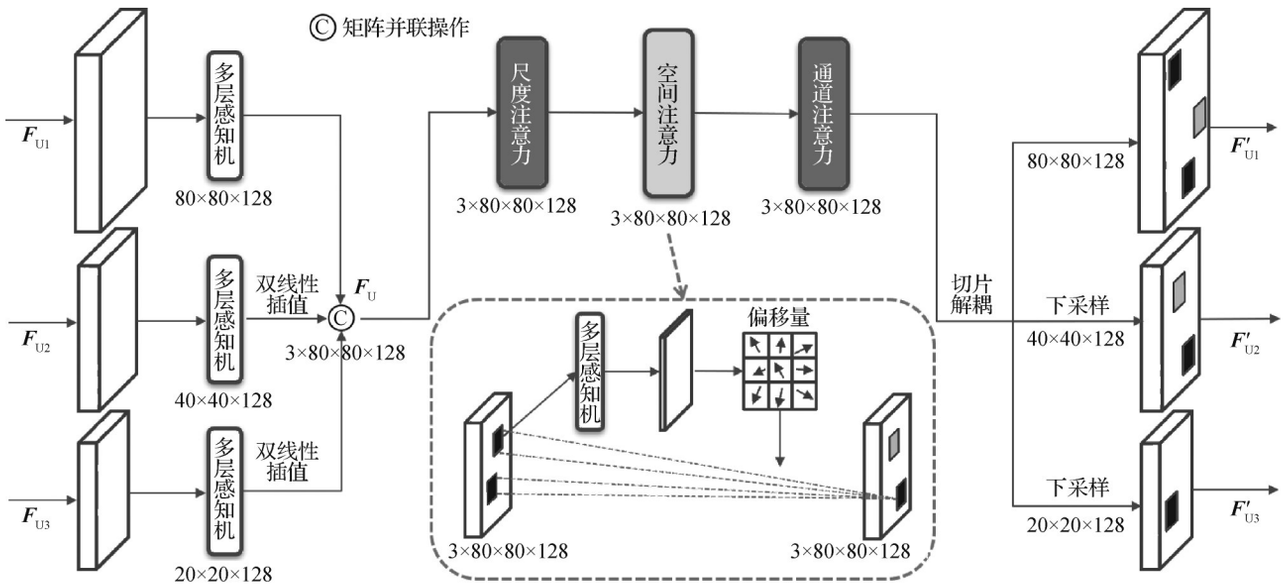


图5　动态选择层算法结构图

Fig. 5　Structure of dynamic selection layer

5.1　实验数据集和相关参数设置

实验所用3个公开数据集简介如表1所示。其中FLIR数据集是一个偏向自动驾驶的数据集，场景多样，具有行人、汽车和自行车3个类别的数据。LLVIP数据集是一个行人检测数据集，只有行人类标注，但其体量较大且夜晚场景较多。VEDAI数据集是一个无人机航拍的俯视图车辆数据集，目标较小，具有9个类别，都是不同的车辆种类。这3个数据集中的可见光与红外图像都已经预先完成了配准的过程。

所有实验用平均精度均值（mean average precision，mAP）进行评测，包括：mAP50、mAP75和mAP。本文采用在COCO（common objects in context）数据集（Lin等，2014）上的YOLOv5预训练模型进行模型初始化，其中VEDAI数据集采用YOLOv5s，FLIR数据集和LLVIP数据集采用YOLOv5l；所有实验采用具有动量参数设置为0.937、衰减率设置为0.000 5和初始化学习率设置为0.01的随机梯度下降（stochastic gradient descent，SGD）优化器进行训练；所有实验batchsize设为16；LLVIP数据集训练100轮，FLIR数据集和VEDAI数据集训练300轮；所有网络模型中均设置3个动态融合层和6个动态选择层；并且采用了马赛克增强和标签平滑等正则化技术。

5.2　定量实验结果比较

本文在FLIR、LLVIP和VEDAI 3个公开数据集上，对所提出的基于动态特征选择的可见光红外图像融合目标检测算法进行性能对比，结果如表2所示，为保证对比的公平性，表中结果均为本文实际复

表1　3个公开数据集的简介

Table 1　Introduction to three public datasets

数据集	训练集/幅	测试集/幅	物体类别	图像特点
FLIR	4 129	1 013	3个：{person, car, bicycle}	场景多样
LLVIP	12 025	3 463	1个：{person}	夜晚样本较多
VEDAI	1 089	121	9个：{car, truck, pickup, tractor, camping car, boat, plane, van, other}	无人机俯瞰视角

注：下载网址：FLIR（https://www.flir.com/oem/adas/adas-dataset-form/）；LLVIP（https://github.com/bupt-ai-cz/LLVIP）；VEDAI（https://downloads.greyc.fr/vedai/）。

现的结果。除本文模型外，实验额外设置了多个对照组，分别是可见光或者红外图像单模态的传统YOLOv5检测器、双光输入的前融合与后融合模型、基线模型、CFT（Fang等，2022）模型和YOLOFusion（Fang和Wang，2022）模型，其中，前融合直接将双光的初始图像通道进行拼接，后融合将双光的检测结果取并集后进行非极大值抑制（non-maximum suppression，NMS）操作，基线模型为将两种图像模态的特征图进行逐元素的相加，以完成双光特征融合的过程。可以看出，本文算法在3个数据集上均取得了性能的提升，特别是在VEDAI数据集的mAP75指标，本文算法相比于基线模型提升了7.5%的检测性能。通过表2，可以总结以下4点性能实验规律。

表2　3个数据集上的性能对比

Table 2　Comparisons of performances on three datasets

/%

模态	算法	FLIR数据集			LLVIP数据集			VEDAI数据集		
		mAP50	mAP75	mAP	mAP50	mAP75	mAP	mAP50	mAP75	mAP
RGB	YOLOv5	58.5	24.1	28.5	92.6	55.4	52.7	77.2	53.7	46.2
IR	YOLOv5	73.9	<u>35.7</u>	<u>39.5</u>	96.2	69.7	61.6	67.6	50.4	43.1
RGB + IR	YOLOv5(前融合)	69.7	33.1	37.0	94.6	57.7	56.3	69.6	51.1	44.0
RGB + IR	YOLOv5(后融合)	73.1	34.9	38.8	96.1	69.7	61.1	73.4	53.7	46.0
RGB + IR	+ Addition(baseline)	76.2	31.9	37.4	96.8	72.0	62.9	76.8	60.1	48.6
RGB + IR	YOLOFusion							78.6	61.5	49.1
RGB + IR	+ CFT	<u>77.1</u>	34.1	39.3	**97.5**	<u>73.6</u>	<u>64.5</u>	<u>80.7</u>	<u>62.7</u>	<u>51.1</u>
RGB + IR	+ 本文 (DyFu+DySe)	**77.5** (↑1.3)	**36.5** (↑4.6)	**40.6** (↑3.2)	<u>97.4</u> (↑0.6)	**74.6** (↑2.6)	**65.0** (↑2.1)	**80.7** (↑3.9)	**67.6** (↑7.5)	**51.7** (↑3.1)

注：加粗和下划线字体分别表示各列最优和次优结果，“↑”后数值表示相比基线模型的性能增量。

1）在低光照场景较多的数据集上，红外图像单模态的目标检测性能远超可见光单模态的性能，并且接近于融合检测的基线模型。例如在FLIR和LLVIP数据集上，红外图像单模态的检测性能均超过可见光单模态10%左右。原因是：一方面红外图像在黑夜情况下有更好的成像效果；另一方面仅将两种图像模态的特征图进行逐元素简单相加的基线模型，其特征融合效果一般。

2）本文算法的特征融合过程是有效的并且对于不同数据集具有鲁棒性。从表2可以看出，有些特征融合方法的性能甚至低于红外单模态的方法，例如在FLIR数据集中红外图像单模态的检测性能，超过了特征融合方法的基线模型和CFT算法。但本文算法的目标检测性能均优于单模态的性能，说明本文算法的特征融合过程是有效的。

3）相比于mAP50指标，本文算法在mAP75和mAP两个指标上的性能提升更加明显，其对应的物理含义为，在检测率基本一致的情况下，本文算法的检测定位精度更高，将其归功于动态选择层中的可变形卷积，相比于传统的卷积核，其可以对物体的边界进行更加准确的提取。

4）对于类别数越多的数据集，本文算法的性能提升越多。例如VEDAI数据集有9种物体类别，本文所提出的动态选择层对于多类别的目标检测性能提升更加明显。同时，图6和图7分别给出了CFT和本文算法在FLIR和VEDAI数据集上的P-R（precision-recall）曲线和混淆矩阵性能对比。可以看出，与CFT相比，本文算法的特征融合效果更好，检测性能更高。

5.3　可视化检测结果比较

本文在FLIR、LLVIP和VEDAI 3个公开数据集上，对所提出的基于动态特征选择的可见光红外图像融合目标检测算法进行目标检测结果的可视化分析，如图8—图10所示，从左至右依次为CFT检测结

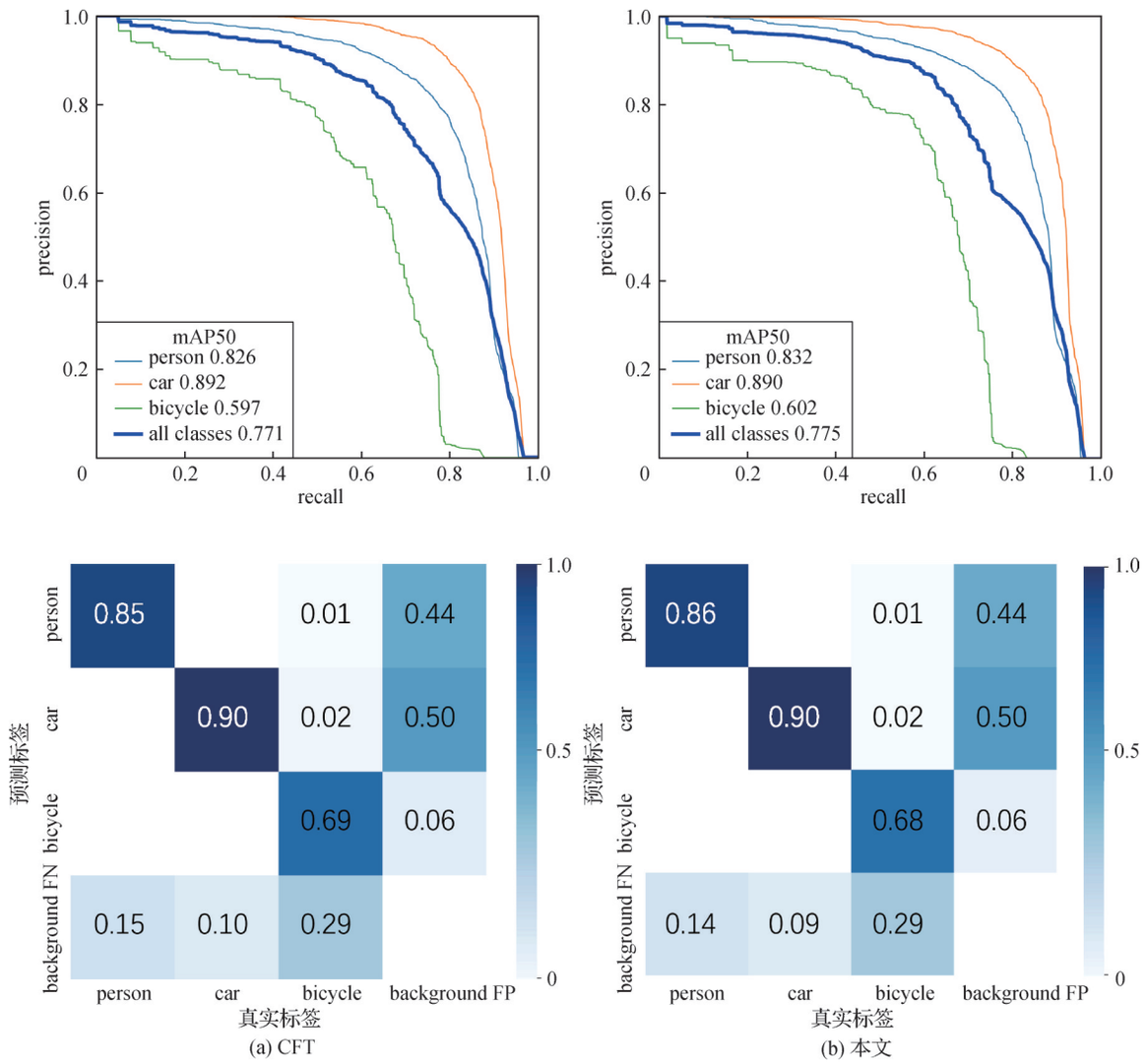


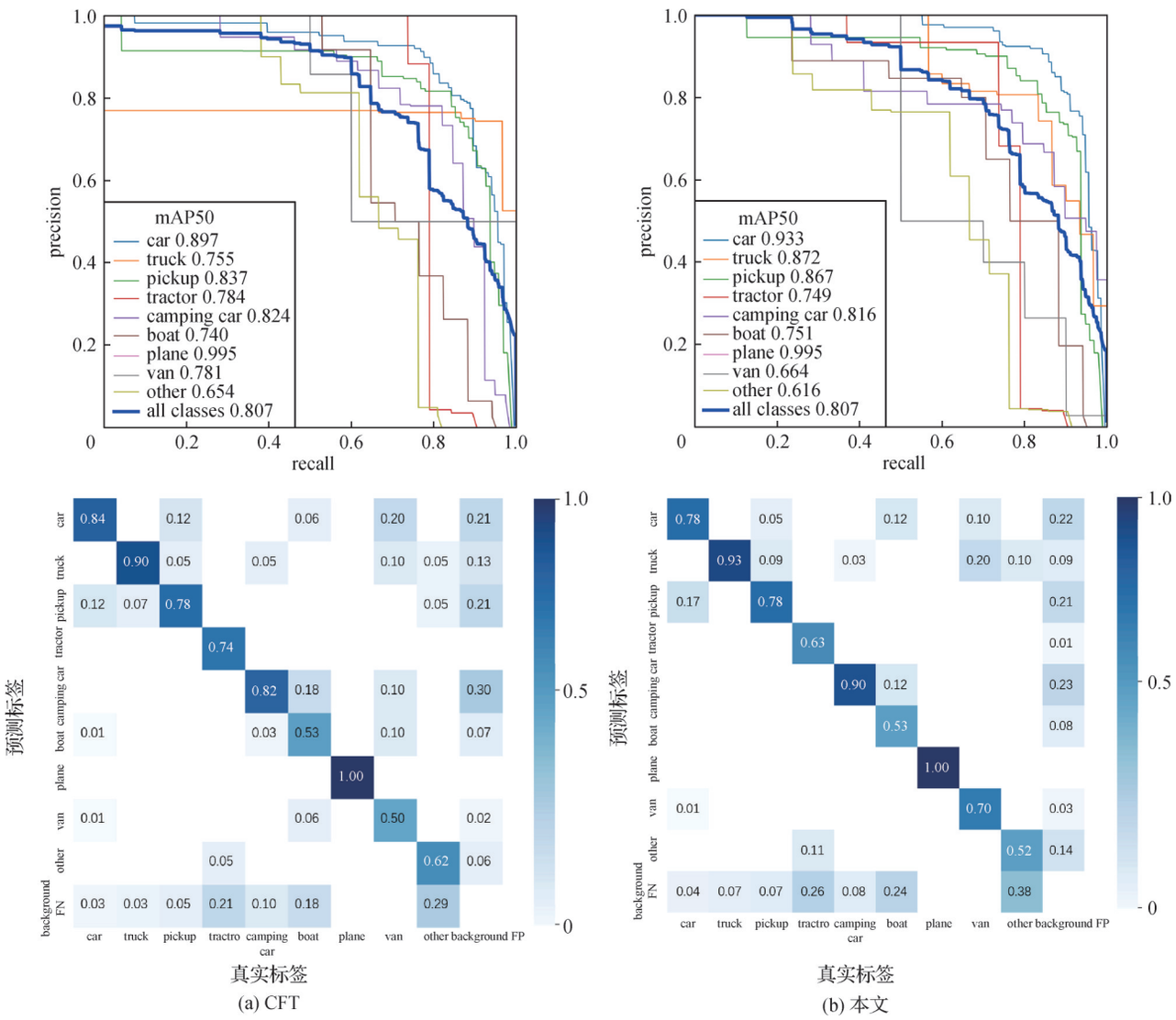
图6　FLIR数据集上的P-R曲线和混淆矩阵对比

Fig. 6　Comparisons of performances with FLIR dataset in terms of P-R curves and confusion matrix ((a) CFT; (b) ours)

果、本文算法的检测结果和目标的标注真值。

可以看出，相比于CFT算法，本文算法的目标检测结果更加精准，虚警率和漏检率较低，并且可以检测出数据集标注中漏标的物体。例如，对于FLIR数据集，在图8的场景1中，CFT模型误将远处的舰船检测为小汽车和行人，误将近岸的快艇检测为小汽车，虚警现象严重，而本文算法并未出现上述的虚警样本；在图8的场景2中，CFT模型同样误检出了自行车类别，而本文算法的检测结果与真值一致；在图8的场景3中，CFT模型漏检了图像右侧边缘的行人，而本文算法将图像中的所有行人全部检出，并且检测框的定位更加精准。对于LLVIP数据集，在图9中，在数据集标注真值漏标右上角骑电动车的人的情况下，CFT模型同样并未检测出该行人，而本文模型可以有效地克服强光反射带来的干扰，准确识别并定位该行人。对于VEDAI数据集，在图10场景1中，由于无人机航拍的俯视图中目标较小，CFT模型错将皮卡车识别为小汽车，而本文模型可以在小目标有限的像素纹理中，精确识别目标的种类；在图10的场景2中，同样遇到了小目标的情况，CFT模型漏检了道路上的黑色小汽车，而本文模型可以准确识别并定位它。

综上可知，经过定量的实验结果比较和定性的可视化检测结果比较，本文算法相比于CFT等特征融合的双光检测算法，具有更高的目标检测性能和鲁棒性。



(a) CFT　　(b) 本文

图 7　VEDAI 数据集上的 P-R 曲线和混淆矩阵对比

Fig. 7　Comparisons of performances with VEDAI dataset in terms of P-R curves and confusion matrix ((a) CFT; (b) ours)

5.4　消融实验

为了验证本文所提出的动态融合层和动态选择层算法的有效性，针对上述两个算法层，在 FLIR、LLVIP 和 VEDAI 3 个数据集上进行了消融实验。每个数据集各有 4 个对照组，分别对应上述两个算法层添加与否的 4 种排列组合情况，实验结果如表 3 所示。可以看出，在 3 个数据集上，添加了上述两个算法层的完整算法模型，均取得了最好的性能。并且与基线模型相比，添加了动态融合层或者动态选择层的单一模块模型，其检测性能也均有所提升。综上实验，证实了所提出的动态融合层和动态选择层算法的有效性。

为了更好地揭示动态融合层 DyFu 算法的作用机理，本文将骨干网络中 3 个动态融合层的注意力权重矩阵分别进行了可视化分析，如图 11 所示。可以看出，随着动态融合层层数的加深，骨干网络中特征提取算法的高关注度区域，逐渐由背景环境转向真实目标物体，说明动态融合层已经具备了感知物体语义特征的能力。特别地，如图 11(b)所示，随着网络的加深，动态融合层可以较好地识别出第 1 节中所述的红外“鬼影”区域，并逐渐给予较低的关注度，以防止虚警的产生。综上分析可知，本文所提出的动态融合层可以较好地融合双光图像的特征信息，验证了所提算法的有效性。

同时，为了客观地评价算法的性能，本文在表 4 中列举了 FLIR 数据集上不同算法的网络参数、计算

图8　FLIR数据集上的目标检测可视化结果对比

Fig. 8　Visual comparison of object detection on the FLIR dataset

图9　LLVIP数据集上的目标检测可视化结果对比

Fig. 9　Visual comparison of object detection on the LLVIP dataset

量与运行时间的对比结果。综合表2和表4可以看出，本文算法相比于CFT论文，虽然在参数量和计算量上均有小幅增加，但换来的是较大的性能提升。

6　结　论

1）本文提出了一种基于动态特征选择的可见光红外图像融合目标检测算法，其包含两个创新模块，分别是特征的动态融合层DyFu和动态选择层DySe。前者嵌入在骨干网络中，利用Transformer结构，多次对多源的图像特征图进行特征融合，以丰富特征表达；后者嵌入在颈部网络中，利用3种注意力机制对多尺度特征图进行特征增强，以筛选有用特征。

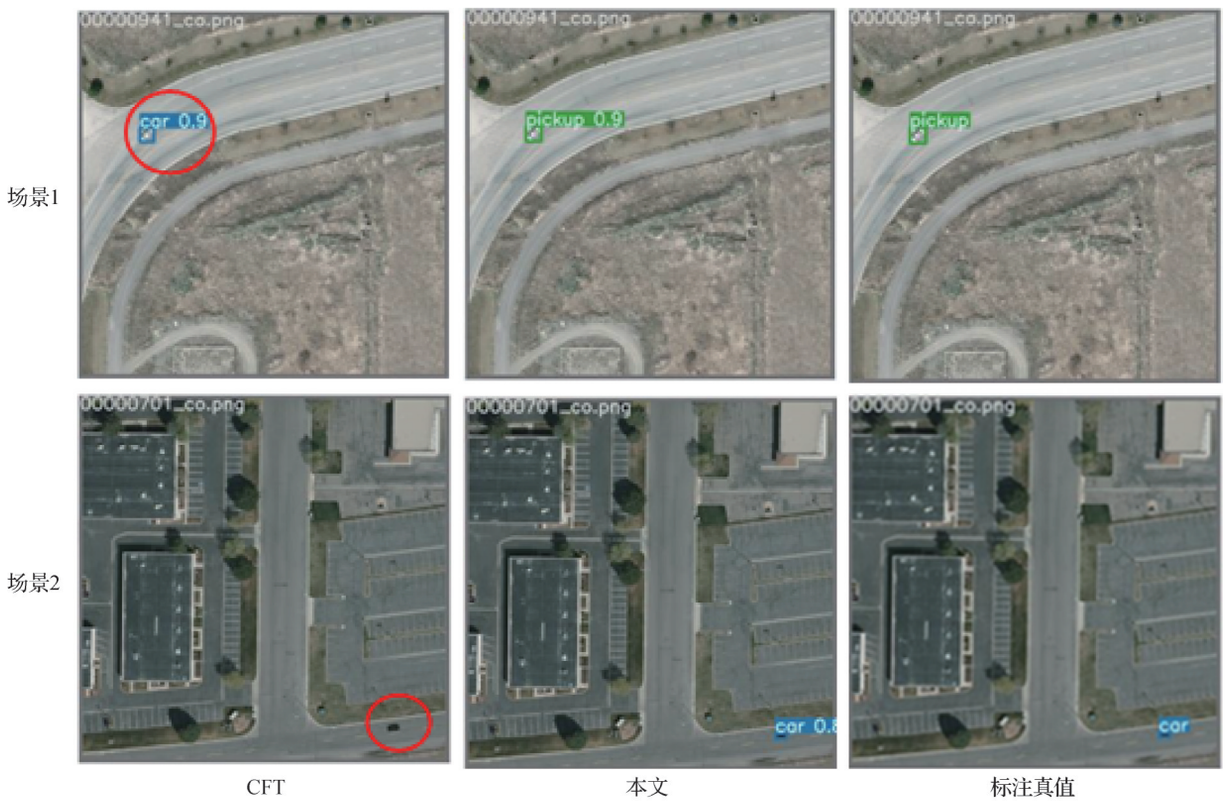


图 10　VEDAI 数据集上的目标检测可视化结果对比

Fig. 10　Visual comparison of object detection on the VEDAI dataset

表 3　3 个数据集上的算法消融实验

Table 3　Ablation studies in the three datasets

/%

DyFu	DySe	FLIR 数据集			LLVIP 数据集			VEDAI 数据集		
		mAP50	mAP75	mAP	mAP50	mAP75	mAP	mAP50	mAP75	mAP
–	–	76.2	31.9	37.4	96.8	72	62.9	76.8	60.1	48.6
√	–	77.2	34	39.4	97.3	73.8	64.5	80.9	62.5	51
–	√	77.1	34.5	39.2	97.2	74	64.8	80.1	62.5	51.5
√	√	**77.5** (↑ 1.3)	**36.5** (↑ 4.6)	**40.6** (↑ 3.2)	**97.4** (↑ 0.6)	**74.6** (↑ 2.6)	**65.0** (↑ 2.1)	**80.4** (↑ 3.6)	**67.6** (↑ 7.5)	**51.7** (↑ 3.1)

注：加粗字体表示各列最优结果，“√”表示采用，“–”表示未采用，“↑”后数值表示相比基线模型的性能增量。

表 4　不同算法在 FLIR 数据集的网络参数、计算量和运行时间对比

Table 4　Comparisons of parameters, computation and runtime among different algorithms on FLIR dataset

模态	算法	参数量/M	计算量/(GFlop/s)	帧速率/(帧/s)
RGB + IR	+ Addition (baseline)	73.72	190.14	15.9
RGB + IR	+ CFT	206.03	224.40	7.7
RGB + IR	+ 本文 (DyFu+DySe)	216.16	236.95	7.1

2）本文算法在 FLIR、LLVIP 和 VEDAI 3 个公开数据集上进行实验验证，mAP75 指标相比于基线模型分别提升了 4.6%、2.6% 和 7.5%，性能达到领域领先水平，并且通过消融实验，验证了所提算法的有效性。

3）由于增加了双光图像特征的动态选择过程，

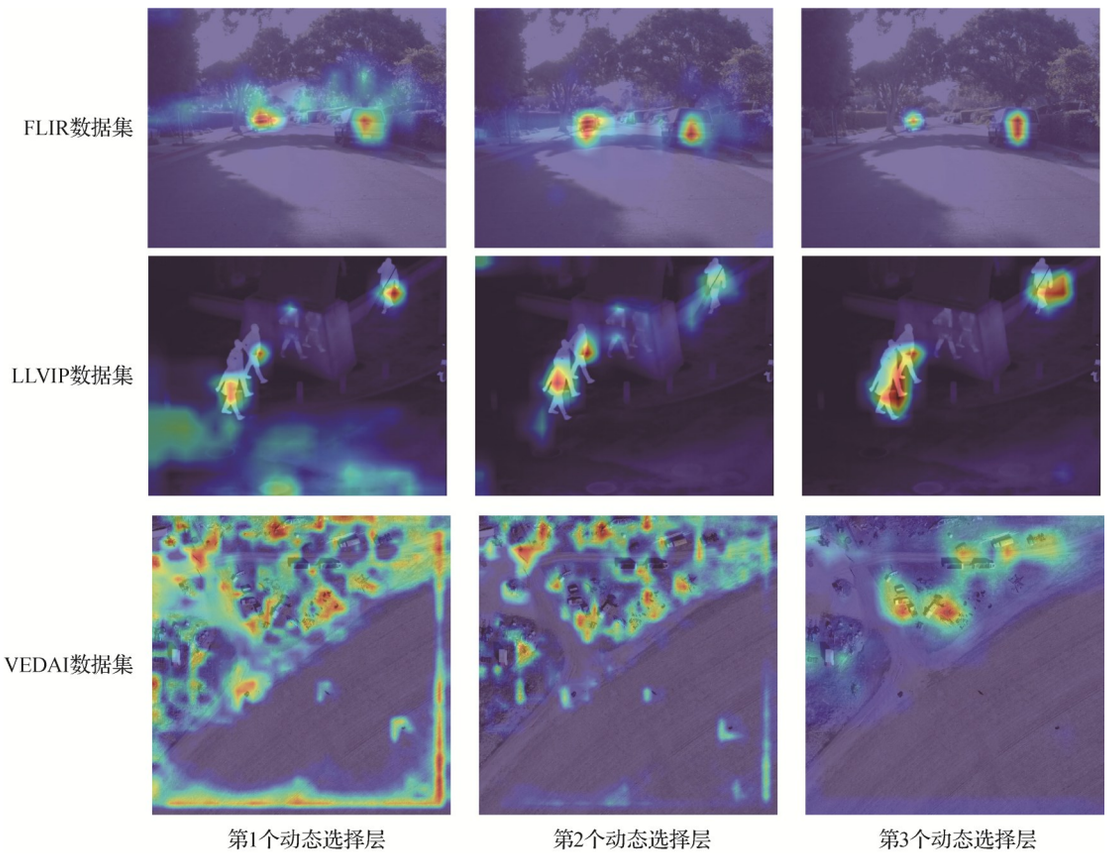


图11　不同动态融合层中注意力权重矩阵可视化结果对比

Fig. 11　Qualitative comparison of attention weight matrix in different dynamic fusion layers

故在模型参数量和计算量指标上相比于领域现有算法有所增加。同时本文要求输入的可见光和红外图像是预先配准的，这也限制了算法的部分应用场景。

4）研究具有轻量化融合模块的双光目标检测网络模型，以及能够自适应配准原始双光图像的算法处理流程，是未来多模态融合目标检测领域的研究重点。

参考文献（References）

Bie Q, Wang X, Xu X, Zhao Q J, Wang Z, Chen J and Hu R M. 2023. Visible-infrared cross-modal pedestrian detection: a summary. Journal of Image and Graphics, 28(5): 1287-1307 (别倩, 王晓, 徐新, 赵启军, 王正, 陈军, 胡瑞敏. 2023. 红外—可见光跨模态的行人检测综述. 中国图象图形学报, 28(5): 1287-1307) [DOI: 10.11834/jig.220670]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with Transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. [2023-07-19]. https://arxiv.org/pdf/2010.11929v1.pdf

Fang Q Y, Han D P and Wang Z K. 2022. Cross-modality fusion Transformer for multispectral object detection [EB/OL]. [2023-07-19]. https://arxiv.org/pdf/2111.00273.pdf

Fang Q Y and Wang Z K. 2022. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. Pattern Recognition, 130: #108786 [DOI: 10.1016/j.patcog.2022.108786]

Fu H L, Wang S X, Duan P H, Xiao C Y, Dian R W, Li S T and Li Z Y. 2023. LRAF-Net: long-range attention fusion network for visible-infrared object detection. IEEE Transactions on Neural Networks and Learning Systems: #3266452 [DOI: 10.1109/TNNLS.2023.3266452]

Guo M H, Cai J X, Liu Z N, Mu T J, Martin R R and Hu S M. 2021. PCT: point cloud Transformer. Computational Visual Media, 7(2): 187-199 [DOI: 10.1007/s41095-021-0229-5]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Jia X Y, Zhu C, Li M Z, Tang W Q and Zhou W L. 2021. LLVIP: a visible-infrared paired dataset for low-light vision//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 3489-3497 [DOI: 10.1109/ICCVW54120.2021.00389]

Li C Y, Song D, Tong R F and Tang M. 2018. Multispectral pedestrian detection via simultaneous detection and segmentation. [EB/OL]. [2023-07-19]. http://arxiv.org/pdf/1808.04818.pdf

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference Computer Vision. Zürich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu S, Qi L, Qin H F, Shi J P and Jia J Y. 2018. Path aggregation network for instance segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8759-8768 [DOI: 10.1109/CVPR.2018.00913]

Liu X P, Ma Y X, Xu K, Wan J W and Guo Y L. 2022. Multi-scale Transformer based point cloud completion network. Journal of Image and Graphics, 27(2): 538-549 (刘心溥, 马燕新, 许可, 万建伟, 郭裕兰. 2022. 嵌入Transformer结构的多尺度点云补全. 中国图象图形学报, 27(2): 538-549) [DOI: 10.11834/jig.210510]

Qiu D F, Hu X Y, Liang P W, Liu X M and Jiang J J. 2023. A deep progressive infrared and visible image fusion network. Journal of Image and Graphics, 28(1): 156-165 (邱德粉, 胡星宇, 梁鹏伟, 刘贤明, 江俊君. 2023. 红外与可见光图像渐进融合深度网络. 中国图象图形学报, 28(1): 156-165) [DOI: 10.11834/jig.220319]

Razakarivony S and Jurie F. 2016. Vehicle detection in aerial imagery: a small target detection benchmark. Journal of Visual Communication and Image Representation, 34: 187-203 [DOI: 10.1016/j.jvcir.2015.11.002]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Saga F's team. Free FLIR thermal dataset for algorithm training [DB/OL]. [2023-07-19]. https://www.flir.com/oem/adas/adas-dataset-form/

Sun X H, Guan Z and Wang X. 2023. Vision Transformer for fusing infrared and visible images in groups. Journal of Image and Graphics, 28(1): 166-178 (孙旭辉, 官铮, 王学. 2023. 红外与可见光图像分组融合的视觉Transformer. 中国图象图形学报, 28(1): 166-178) [DOI: 10.11834/jig.220515]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: MIT Press: Curran Associates Inc.: 6000-6010

Wang Y, Guizilini V C, Zhang T Y, Wang Y L, Zhao H and Solomon J. 2022. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries//Proceedings of the 5th Conference on Robot Learning. London, UK: [s.n.]: 180-191

Zhang H, Fromont E, Lefevre S and Avignon B. 2021. Guided attentive feature fusion for multispectral pedestrian detection//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 72-80 [DOI: 10.1109/WACV48630.2021.00012]

Zhang L, Liu Z Y, Zhang S F, Yang X, Qiao H, Huang K Z and Hussain A. 2019. Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion, 50: 20-29 [DOI: 10.1016/j.inffus.2018.09.015]

Zhu X Z, Hu H, Lin S and Dai J F. 2019. Deformable ConvNets V2: more deformable, better results//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9300-9308 [DOI: 10.1109/CVPR.2019.00953]

作者简介

许可，男，副教授，硕士生导师，主要研究方向为水声信号处理与计算机视觉。E-mail：xuke@nudt.edu.cn

刘心溥，通信作者，男，博士研究生，主要研究方向为三维场景感知与目标检测。E-mail：liuxinpu@yeah.net

汪汉云，男，副教授，主要研究方向为三维场景智能感知与处理。E-mail：why.scholar@gmail.com

万建伟，男，教授，博士生导师，主要研究方向为信号处理技术与系统、雷达技术和图像压缩。

E-mail：kermitwjw@139.com

郭裕兰，男，副教授，主要研究方向为三维视觉与机器学习。E-mail：yulan.guo@nudt.edu.cn