激光与光电子学进展 先进成像

注意力机制与多层特征融合策略的安检图像目标检测方法

张弘，张思聪*

西安邮电大学自动化学院，陕西 西安 710100

摘要 YOLOv5（You only look once，v5）具有检测速度快、精度高的优点，被广泛应用于实时目标检测中。针对X光安检图像背景复杂、物体多尺度、相互重叠导致的错检、漏检问题，在YOLOv5s网络结构的基础上，通过改进注意力机制开发了新的特征融合策略，并提出了一种具有自适应特征融合策略与注意力机制的目标检测YOLOv5s-AFA网络。该网络在浅层引入扩大感受野模块与改进的空间注意力机制，在深层引入改进的通道注意力机制。新的特征融合策略可每次输出三个不同深度的特征图，通过自适应学习权重融合浅层空间信息与深层语义信息，使网络的学习更具针对性。在X光安检图像数据集上的目标检测结果表明，相比其他对比网络，YOLOv5s-AFA网络的错检率和漏检率有明显降低。

关键词 图像处理；深度学习；目标检测；X光安检图像；注意力模块；特征融合

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP202259.1610013

Security Inspection Image Object Detection Method with Attention Mechanism and Multilayer Feature Fusion Strategy

Zhang Hong, Zhang Sicong*

School of Automation, Xi'an University of Posts & Telecommunications, Xi'an 710100, *Shaanxi, China*

Abstract YOLOv5 (You only look once, v5) is widely used in real-time target recognition because of its fast detection speed and high accuracy. On X-ray security image detection errors or omissions problems with complex backgrounds, multiple scales, and overlapping. By improving the attention mechanism, a new feature fusion strategy is developed based on the YOLOv5s network structure. This study proposes a YOLOv5s-AFA object detection network with an adaptive feature fusion technique and an attention mechanism. In the shallow layer of the network, an extended receptive field module and an improved spatial attention mechanism are introduced, whereas the improved channel attention mechanism is introduced in the deep layer. The new feature fusion technique can output three feature maps of varying depths at a time and fusing shallow spatial and deep semantic information using adaptive learning weights to improve the network learning. The target results on the X-ray security image dataset show that the false and missed detection rates of the YOLOv5s-AFA network decrease considerably compared with other compared networks.

Key words image processing; deep learning; object detection; X-ray security image; attention module; feature fusion

1 引 言

X光安检仪是公共交通、快递包裹安全检查中的重要设备[1]，随着客流量和物流量的增加，用计算机和人工智能技术辅助进行安检以实现快速且准确的危险物品检测已成为重要的应用和研究方向。针对X光安检图像中的物品检测问题，Mery[2]制作了伽马X射线（GDX-ray）安检图像数据集，该数据集包含铸件、行李、焊缝等五大类共19407张安检图像，专门用于安检人员训练和计算机视觉算法的研究。Jaccard等[3]用一个19层的卷积神经网络（CNN）框架对X射线货物图像进行分类，但网络结构过于简单，检测速度和精度都不能满足实际应用的要求。

为了提高深度学习网络模型的性能，Girshick等[4]

收稿日期：2021-07-22；**修回日期**：2021-08-23；**录用日期**：2021-09-24

基金项目：陕西省重点研发计划（2021SF-478）

通信作者：*1622065516@qq.com

在CNN的基础上提出了区域卷积神经网络(RCNN)目标检测模型。首先,用该模型提取输入图像的候选区域,对所有区域进行归一化后逐个将其输入CNN中提取特征。然后,用支持向量机(SVM)进行分类、回归,借鉴空间金字塔池化网络(SPP-Net)[5]提出了一种感兴趣区域池化(ROI pooling)方法并将其应用于Fast RCNN[6]中,统一了候选区域特征的尺寸,构建了多任务损失的思想。最后,将分类损失和边框回归损失进行统一训练,使分类任务和定位任务共享卷积特征、相互促进,提升检测效率。为了避免过于依赖选择性搜索算法产生候选区域,Ren等[7]提出了Faster RCNN算法,用区域选择网络(RPN)产生候选区域框,实现了端到端的训练,大幅提升了网络的检测速度和精度。Liu等[8]借鉴Fast RCNN中的Anchor思想提出了单步多框检测(SSD)算法。近年来,Redmon等[9]提出的YOLO(You only look once)网络融合了目标的分类、定位和检测,只需一次计算就能得到输入图像中的目标边界框和类别概率,极大缩短了检测时间。Redmon等[10]用Anchor机制对YOLOv1进行改进,提出了YOLOv2网络,将原始网络中每个网格随机生成的边界框(Bounding box)模板改为K均值(K-means)聚类方法获得的Anchor,通过引入多尺度训练提出了YOLOv3网络[11],进一步借鉴残差网络(ResNet)构建了DarkNet-53结构,只采用尺寸为1×1、3×3的卷积层,最终输出三种不同尺寸的特征图,且每个特征图都是由高层特征和浅层特征融合得到,可用于处理多尺度目标。Bochkovskiy等[12]将DarkNet-53改进为跨阶段局部网络(CSP DarkNet)[13]并提出了YOLOv4网络,该网络使用了Mish激活函数,新增了Mosaic数据处理、跨Batch的批归一化(CmBN)等技巧,在COCO数据集上的每秒检测帧数(FPS)为65,平均精度(AP)达到了43.5%。为了解决航空发动机部件表面缺陷特征相近的问题,李彬等[14]在YOLOv4网络前加入了多次小卷积处理,提高了网络对缺陷的特征提取能力,将网络的类平均精度(mAP)提升了3.71个百分点。YOLOv5网络新增了Focus结构,通过改进两种CSP结构和Anchor生成策略并加入了PAN结构,使网络的灵活性与检测速度远超YOLOv4网络。在口罩佩戴情况的检测中,YOLOv5网络的精确率达到了92.4%[15]。

为了使神经网络将关注点聚焦在需要注意的部分,使信息处理更高效,基于深度学习的注意力机制得到了迅速发展。刘建男等[16]在YOLOv3主干网络末端引入空间注意力机制,用CSP结构替换原有的残差结构,并在特征融合时加入一条浅层到深层的直连路径,解决了浅层信息经多层传递导致的位置信息丢失问题,在COCO2017数据集上的检测精度达到了65%。徐诚极等[17]提出的Attention-YOLO网络在残差结构中加入串行的通道与空间注意力机制,在残差连接时加入一个二阶项,通过增加残差结构的非线性程度提升网络的泛化性能,在VOC2007数据集上的mAP达到了81.9%。李浪怡等[18]将YOLOv5中的CSP模块替换为GhostBottleneck,并在第5、第7层加入压缩与激励(SE)通道注意力模块,在Ⅱ型钢轨公开数据集上的检测精度达到了91.8%。针对X光安检图像中的目标检测问题,张友康等[19]提出了一种非对称多视野网络(ACMNet),采用空洞卷积模块(DCM)扩大视野,建立全局与局部的联系,以提高遮挡情况下违禁物品的检出率;用小卷积非对称模块(ATM)学习不同尺度违禁物品的局部特征,降低了小目标的漏检率,在X光安检领域公开数据集上的检测精度为84.3%。张震等[20]将YOLOv3中的残差结构(Res unit)替换为密集块(Dense block),通过密集连接将特征融合复用,提高了网络对安检异常图像中小目标的检测能力,在自制异常图像数据集上的检测精度达到了91.68%,FPS达到59。郭守向等[21]借鉴复合骨干网络(CBNet)[22]的思想提出了YOLO-C目标检测算法及双层DarkNet-53骨干网络,该网络在SIXray-OD数据集上mAP达到了73.68%,但FPS仅为40。

上述模型为X光安检图像的目标检测研究提供了重要的方法,但当图像中的目标较小或受背景干扰、相互重叠等复杂情况影响时,仍存在错检、漏检以及检测精度较低等问题。本文针对这类目标的检测问题进行研究,提出了一种结合自适应特征融合(AF)策略与注意力机制的YOLOv5s-AFA网络。在X光安检图像数据集上的检测实验结果表明,该网络具有损失收敛迅速、检测速度快、精度高的优点,能实现目标物体多尺度、背景复杂、相互重叠情况下的检测识别,可达到比Faster RCNN、RetinaNet、PP-YOLOv2(在PaddlePaddle框架下改进的YOLOv3模型)、YOLOv4、YOLOv5s网络更高的精确率、召回率及mAP。

2 理论基础

2.1 YOLOv5s网络

YOLOv5网络具有模型尺寸小、训练和检测速度快以及检测精度高等优点,可适用于安检场景下的目标检测。根据模型深度和参数量可将YOLOv5网络分为YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x四类,不同网络中Backbone的卷积核数量如表1所示。

表1 不同YOLOv5网络中Backbone的卷积核数量

Table 1 Number of convolution kernels of Backbone in different YOLOv5 networks

Network	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
Focus	32	48	64	80
CBL-1	64	96	128	160
CBL-2	128	192	256	320
CBL-3	256	384	512	640
CBL-4	512	768	1024	1208
Model size /MB	27	84	192	367

其中，CBL模块为卷积(Conv)操作、批归一化(BN)操作以及带泄漏的修正线性单元(Leaky ReLU)激活函数。可以发现，YOLOv5s网络是所有结构中深度最浅、特征图宽度最小的模型，其模型尺寸最小、训练速度最快，有利于模型的快速部署。因此，本算法用YOLOv5s作为基础网络进行改进。YOLOv5s网络包括Input、Backbone、Neck、Head四个部分，具体结构如图1所示。其中，SPP为空间金字塔池化，Maxpool为最大值池化，Add为特征图相加操作，Concat为通道数合并操作，Slice为裁剪操作。

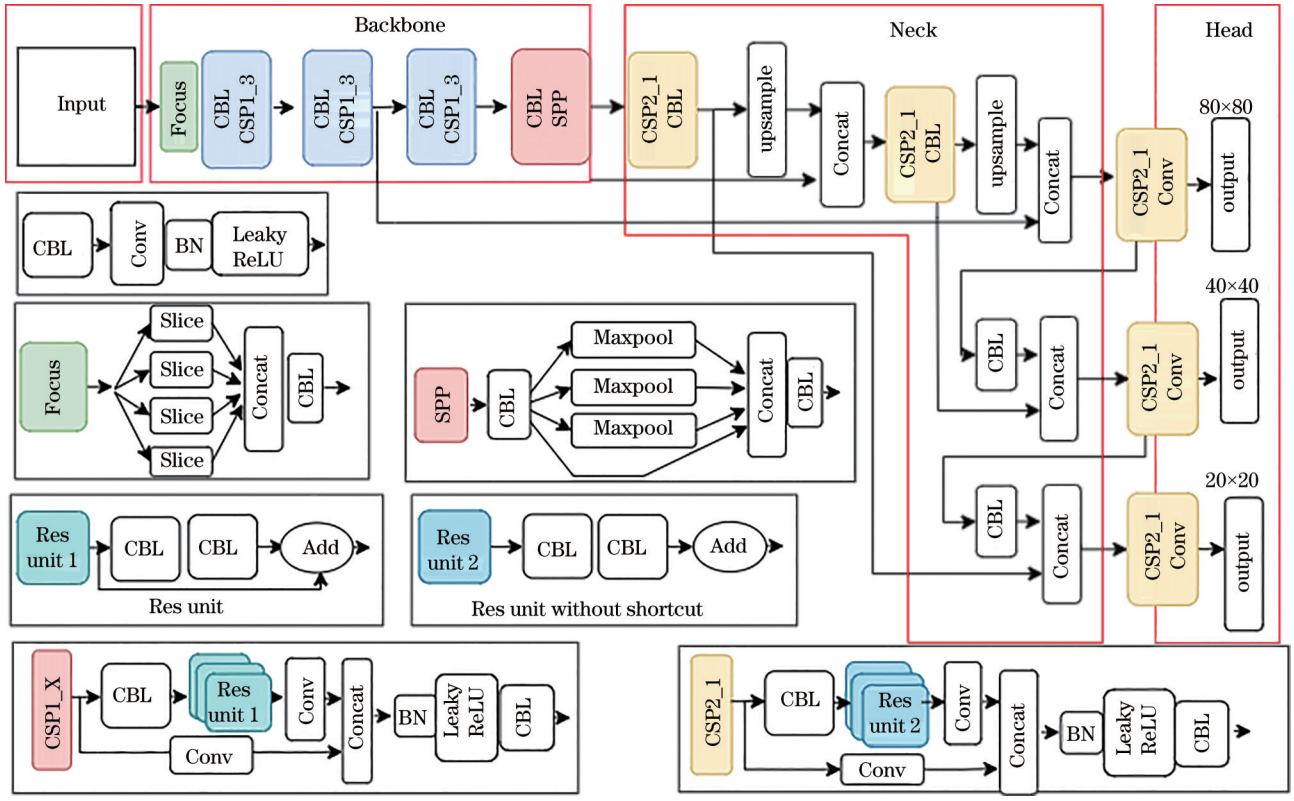

图1　YOLOv5s网络的结构

Fig. 1　Structure of the YOLOv5s network

1) Input端使用了Mosaic数据增强：随机使用四张图像进行拼接，丰富检测物体的背景；自适应锚框(Auto learning bounding box Anchor)在预设的三组锚框基础上根据训练数据自适应计算最佳锚框尺寸；自适应图像缩放可在训练时自动填充输入图像的黑边，不需要单独对训练集图像进行缩放、填充。

2) Backbone端将图像特征组合生成特征金字塔。其中：Focus结构利用切片操作将尺寸为640×640×3的输入图像裁剪成尺寸为320×320×12的特征图，再经过卷积操作输入后续网络，相比传统的卷积下采样，减少了模型计算量且不会导致信息丢失；CSP1_X和CSP2_1两种CSP结构中残差模块的数量由参数控制，分别被添加在网络的Backbone和Neck中，从而将梯度变化信息反映在特征图中，减少了模型的参数量和每秒浮点运算量(FLOPS)，在保证准确率的同时缩小了模型尺寸。

3) Neck的作用是对图像特征进行检测，利用锚框生成带有类概率、置信度、边界框坐标的最终输出。FPN+PAN[23]结构在获得丰富语义特征的同时能得到较强的定位特征，提升了特征融合的能力。

4) Head的输出端用广义交并比(GIoU)[24]作为损失函数，解决了普通交并比(IoU)面对真实框与预测框没有重叠时梯度为0无法优化的问题。

YOLOv5s网络不仅基于目标框中心点所在网格产生Anchor的负责网格，还需根据中心点所在网格的位置选取临近的两个网格，三个网格共产生3×3个Anchor。计算出Anchor的宽(W)、高(H)比并将其与预测框的形状进行匹配，相差较大的预测框被当作背景舍弃。

2.2　通道注意力模块

在ImageNet2017比赛中，Hu等[25]提出SE通道注意力模块，其核心思想在于通过学习的方式自动获取每个通道权重，再依照该权重增强有用的特征通道、抑制无用的特征通道，使网络学习到不同特征通道的重要程度，进而获得通道上的注意力(Channel attention)。SE模块主要由Squeeze和Excitation两部分构成，如图2所示。

Squeeze操作通过全局均值池化(GAP)将C个通道的二维特征($H\times W$)压缩为一个表征特征通道上响应全局分布的实数，将原始尺寸为$H\times W\times C$的特征

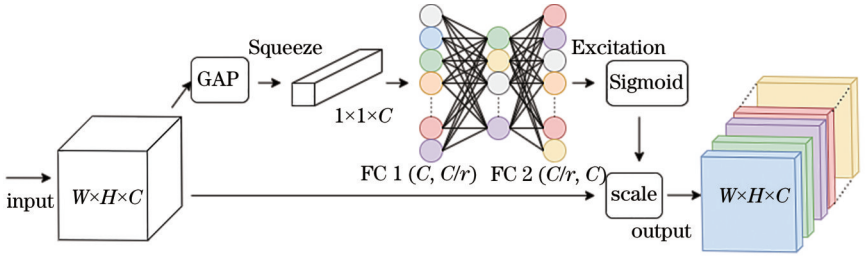

图 2　SE模块的结构

Fig. 2　Structure of the SE module

图转换为尺寸为1×1×C的输出。Excitation用包含两个全连接(FC)层的瓶颈(Bottleneck)结构：先使用缩放参数r对通道特征进行降维，以减少模型复杂度；然后将其映射回来，以构建通道之间的相关性；最后将每个特征通道的权重经过Sigmoid激活后，作用到原始特征图对应的特征通道上。

Excitation过程中，全连接层的降维操作破坏了通道特征与权重的直接对应关系。而高效通道注意力(ECA)[26]模块提出了一种跨信道交互的策略，用一维卷积替代全连接层，使特征图间可以共享卷积信息，在保持网络性能的同时降低了网络计算量。

2.3　空间注意力模块

空间注意力(SA)主要用来弥补只使用通道注意力的不足。通道注意力集中在给定的输入图像“what”是有意义的，空间注意力关注图像“where”的特征是有意义的[27]。SA模块的结构如图3所示。首先，分别用最大值池化和均值池化在通道维(C)上将特征图进行压缩，输出两个尺寸为$W\times H\times 1$的通道描述；然后，将两个描述拼接并通过7×7卷积还原通道维度，经Sigmoid激活后得到权重系数M_s；最后，将权重系数与输入的特征图($W\times H\times C$)按元素相乘，得到空间注意力特征图。

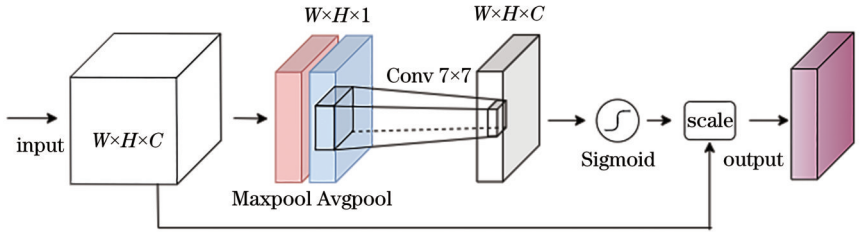

图 3　SA模块的结构

Fig. 3　Structure of the SA module

3　YOLOv5s-AFA网络

X光射线机使用两个射线管，根据物体对两种射线吸收系数的差异产生X光图像[28]。X光图像具有的特性：1)物体重叠、细节特征丢失；2)颜色信息丢失、杂乱无序。将X光射线机用于行李、包裹安检时，物体重叠、尺度各异、背景复杂等情况均会增大相应图像的目标检测难度。为了解决这些特定的问题，提出了一种针对X光安检图像的YOLOv5s-AFA网络。该网络基于YOLOv5s网络框架，通过改进空间和通道注意力模块，在浅层加入改进的空间注意力模块、在深层加入改进的通道注意力模块，应用注意力机制增强受背景干扰、相互重叠目标的特征信息。此外，YOLOv5s-AFA通过改进特征融合策略，使每个输出都得到融合网络三处不同深度特征层的特征，并将浅层和深层的特征信息进行自适应融合，使网络获得更精细的特征，提高其对较小目标的检测能力。YOLOv5s-AFA网络的结构如图4所示。其中，ERF为扩大感受野模块，SAd为加入空洞卷积的空间注意力模块，ASFF为自适应空间特征融合模块，iECA为改进后的高效通道注意力。

3.1　改进的空间注意力模块

深度学习网络在不同深度时，卷积捕获的特征信息不同，深层次的特征图具有更多的全局和语义信息，而浅层特征图包含更丰富的空间结构细节。随着网络的加深，图像分辨率逐渐减小，较小目标的信息容易丢失，目标的位置特征也更加模糊。因此，在YOLOv5s网络Backbone中第二个CBL模块后加入SA模块，获得空间注意力特征，使网络具有捕获浅层特征图中目标边界、轮廓及位置信息的能力。为了降低SA模块计算量对网络的影响，将SA模块中7×7的卷积替换成卷积核尺寸为3×3、膨胀率为2(kernel：3×3，d=2)的空洞卷积(Dilated convolution)[29]，在不改变感受野大小的同时降低SA模块的参数量，其映射关系可表示为

$$M_s(F)=\sigma\left\{f_{d=2}^{3\times 3}\left[X_{\text{Avgpool}}(F),X_{\text{Maxpool}}(F)\right]\right\},\quad(1)$$

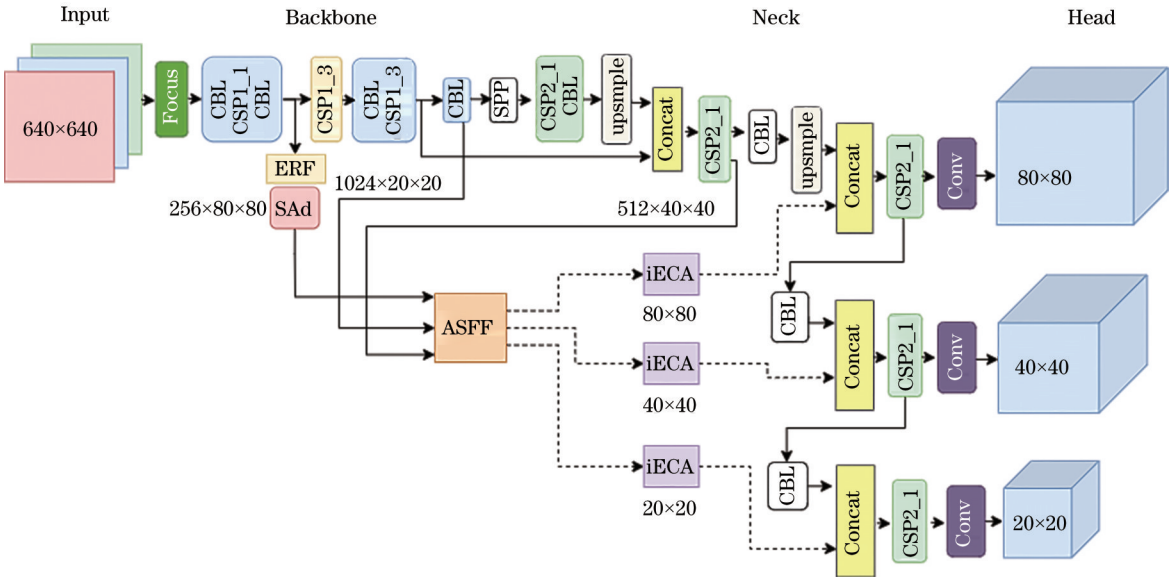

图 4　YOLOv5s-AFA 网络的结构

Fig. 4　Structure of the YOLOv5s-AFA network

$$F_s = F \times M_s(F), \tag{2}$$

式中，$f_{d=2}^{3\times3}$ 为卷积核尺寸为 3×3 且膨胀率为 2 的空洞卷积，σ 为 Sigmoid 激活函数，F 为初始输入的特征图，F_s 为最终的空间注意力特征图。

浅层特征的细节虽然丰富，但存在图像感受野小的问题，而较小的感受野会使注意力模块捕获的特征信息变少[30]，因此，在空间注意力模块前加入一个 ERF 模块，利用不同尺度的卷积核扩大感受野，其结构如图 5 所示。其中：ERF 模块的输入为第二个 CBL 模块输出的特征图（尺寸为 160×160×128），先由 1×1 卷积降维至 160×160×64，以减少计算量；然后，分别由 1×1、3×3、膨胀率为 1 的 3×3 卷积进行下采样，得到小、中、大三种尺度的感受野，并用 Maxpool 保留显著特征，再通过 1×1 卷积调整维度；最后，将四种不同感受野的特征图进行相加，得到尺寸为 80×80×256 的混合感受野特征图。由于单独加入 ERF 只会增加网络的深度与计算量，对网络性能的提升没有帮助，因此，将 ERF 与 SAd 作为一个模块作用于网络，记为 E-SAd。

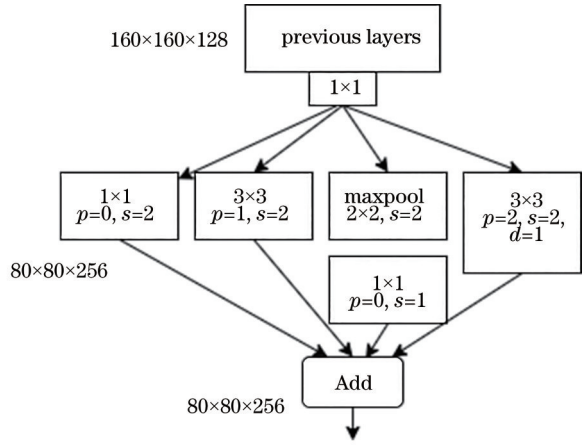

图 5　ERF 模块的结构

Fig. 5　Structure of the ERF module

3.2　改进的通道注意力模块

SE 与 ECA 模块在压缩特征图时都仅使用了均值池化，这不利于提取到最突出的特征信息。因此，改用最大值池化与均值池化并行的方式压缩特征，其结构如图 6 所示。

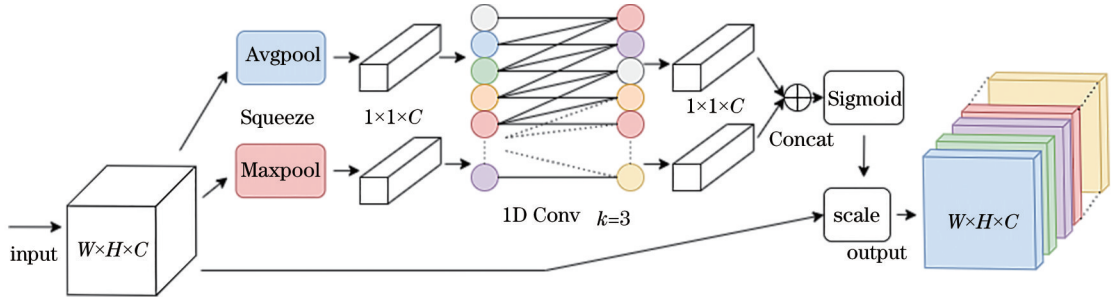

图 6　iECA 模块的结构

Fig. 6　Structure of the iECA module

输入特征图分别经过均值、最大值池化后，特征图的尺寸从 $W\times H\times C$ 压缩为 $1\times1\times C$，由卷积核大小为 k、步长为 1 的一维（1D）卷积得到具有信道交互信息的 $1\times1\times C$ 特征，进行特征拼接后，经 Sigmoid 激活得到通道注意力权重，与输入相乘获得通道注意力特征图。其映射关系可表示为

$$M_c(F)=\sigma\left\{X_k^{1D}\left[X_{Avgpool}(F)\right]+X_3^{1D}\left[X_{Maxpool}(F)\right]\right\},\quad(3)$$

$$F_c=F\times M_c(F),\quad(4)$$

式中，X_k^{1D} 表示卷积核为 k 的一维卷积，F_c 为通道注意力特征图。

3.3　多尺度特征融合策略

大多数目标检测算法使用的特征融合方式以 FPN 为主，融合过程是基于特征图尺寸与通道数的对应关系进行，一般采取级联（Concatenation）和元素拼接（Element-wise）方式将相对应的特征图进行拼接或相加。如图 1 中为了不改变特征图的大小与通道数，尺寸为 80×80 的输出仅融合了 Backbone 中第一个 CSP1_3 模块的输出。原始 YOLOv5s 网络中特征金字塔结构在检测目标时会输出三种大小的特征层，若不同特征层判断目标的结果发生冲突（某一特征层上的目标被判断为正样本，但在其余特征层上相应位置的目标却被视为背景），会干扰训练时的梯度计算，从而降低特征金字塔的有效性。为了同时融合不同尺度的特征图，Liu 等[31]提出一种 ASFF 方法，将输入的多层特征与权重参数相乘并相加得到新的融合特征，充分获取了网络深层的语义特征与浅层的位置特征。

本算法设计的多层特征融合策略可同时融合 YOLOv5s 网络三处不同深度的特征层，使每次的输出结果能同时获得网络三处不同深度的特征信息，充分学习网络浅层空间注意力模块提取到的空间结构信息与网络深层的语义信息，并用自适应方法调整三层融合特征层的权重，滤除特征层上的无用信息，从而提高网络性能。最后，加入改进的通道注意力模块对融合后的特征信息进行加权筛选。该自适应特征融合的注意力结构如图 7 所示。其中，将 SAd 的输出作为特征融合的第 3 层（Level 3，尺寸为 80×80×256），将第 4 个 CBL 模块的输出作为特征融合的第 2 层（Level 2，尺寸为 40×40×512），将第 2 个 CSP2_1 模块的输出作为特征融合的第 1 层（Level 1，尺寸为 20×20×1024）。

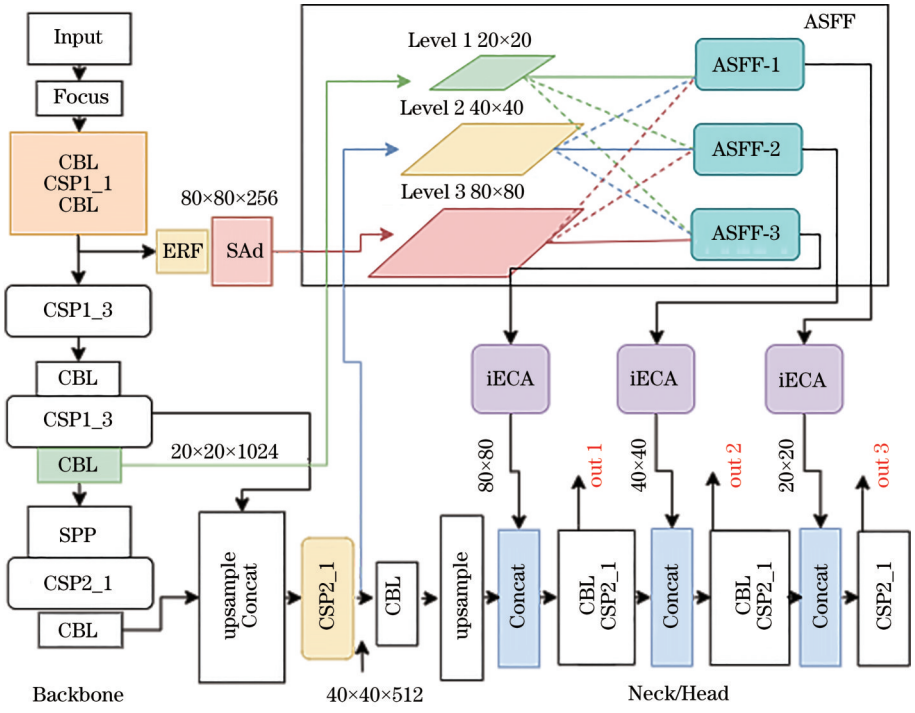

图 7　ASFF 模块的注意力结构

Fig. 7　Attention structure of the ASFF module

ASFF 模块中进行融合的三个输入特征尺寸、维度均不同，因此，需要对输入特征图进行缩放。以 ASFF-1 的输出为例，为使 Level 1、Level 2、Level 3 层的特征图尺寸统一为 20×20×1024：对 Level 2 使用 1×1 卷积层升维，用 3×3 卷积层（步长为 2）下采样减小分辨率；对 Level 3 使用 1×1 卷积层降维，并在 3×3 卷积层（步长为 2）前加入最大值池化（步长为 2）。在扩大特征图尺寸时，使用插值方式进行上采样，以提高特征图的分辨率（特征图比例为 1/2 时使用 2 倍插值，比例为 1/4 时使用 4 倍插值）。统一尺寸后的输入由三个权重参数分别相乘后相加，得到融合了浅层空间结构信息和深层细粒度语义信息的特征，映射关系可表示为

$$\boldsymbol{Y}_{ij}^{l}=\alpha_{ij}^{l}\cdot\boldsymbol{X}_{ij}^{1\to l}+\beta_{ij}^{l}\cdot\boldsymbol{X}_{ij}^{2\to l}+\gamma_{ij}^{l}\cdot\boldsymbol{X}_{ij}^{3\to l},\quad(5)$$

$$\alpha_{ij}^{l}=\frac{e^{\lambda_{\alpha_{ij}}^{l}}}{e^{\lambda_{\alpha_{ij}}^{l}}+e^{\lambda_{\beta_{ij}}^{l}}+e^{\lambda_{\gamma_{ij}}^{l}}},\tag{6}$$

$$\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}=1,\tag{7}$$

式中，$\boldsymbol{X}_{ij}^{1\to l}$为由第$n$级到第$l$级调整特征图$(i,j)$位置的尺寸，$\boldsymbol{Y}_{ij}^{l}$为输出通道特征图$\boldsymbol{Y}^{l}$第$(i,j)$个向量，$\alpha_{ij}^{l}$、$\beta_{ij}^{l}$、$\gamma_{ij}^{l}$为权重参数，$\lambda_{\alpha}^{l}$、$\lambda_{\beta}^{l}$、$\lambda_{\gamma}^{l}$为控制参数。分别对Level 1、Level 2、Level 3对应的输入特征$\boldsymbol{X}^{1\to l}$、$\boldsymbol{X}^{2\to l}$、$\boldsymbol{X}^{3\to l}$用1×1卷积得到控制参数λ_{α}^{l}、λ_{β}^{l}、λ_{γ}^{l}，再结合Softmax激活函数计算出α_{ij}^{l}、β_{ij}^{l}、γ_{ij}^{l}，最终得到的权重参数值范围为[0,1]，且权重和为1。λ_{α}^{l}、λ_{β}^{l}、λ_{γ}^{l}是由调整大小后的特征图经过1×1卷积获得，因此，可以通过标准的反向传播学习出最优的权重参数，达到自适应特征融合的目的。

4　实验结果与分析

实验研究在Google Colab平台下进行，使用的GPU为TeslaK80，CUDA版本为11.2，显存为15 G，深度学习框架为PyTorch。

4.1　数据集与模型评价指标

实验使用的数据集为在公共数据集SIXray上重新标注的6000张彩色X光安检图像，检测目标包含Knife、Gun、Wrench、Pliers四个类别的物体。将训练集和验证集以7∶3的比例随机划分。图8列出了6张具有代表性的X光安检图像。

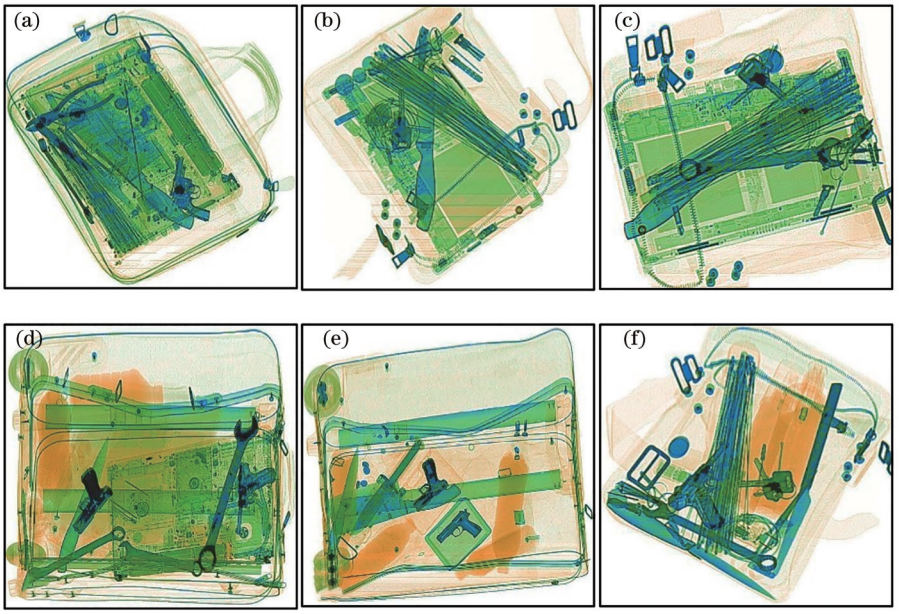

图8　X光安检图像数据集。(a)～(f)图像1～图像6

Fig. 8　X-ray security image dataset. (a)–(f) Image 1– image 6

对训练集中Label的宽、高及中心点进行统计，结果如图9所示。可以发现：图9(a)中颜色较深且密集的地方表示训练集中目标中心点，主要分布在图像的中左区域；图9(b)中目标尺寸偏中或偏小。

对模型的评价指标包括精确率(P)、召回率(R)、平均精度(X_{AP})、mAP、变阈值类平均精度(mAP，0.50∶0.95)及参数增量(N_P)。精确率表示被预测为正样本中正确预测的样本数所占比例，可表示为

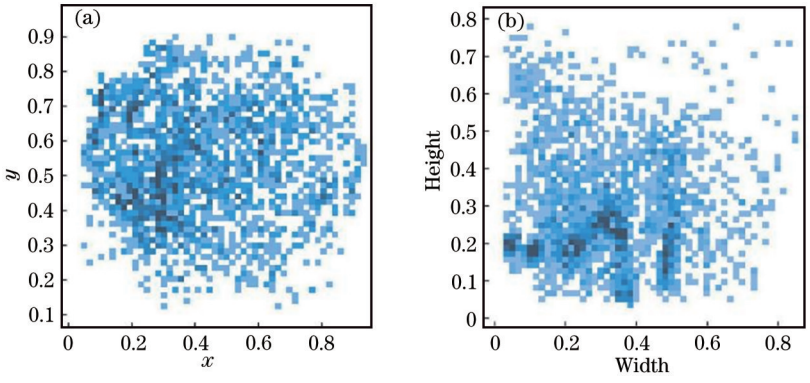

图9　训练集的Label分布。(a)中心点分布；(b)宽、高的分布

Fig. 9　Label distribution of training set. (a) Center point distribution; (b) width and height distribution

$$P=\frac{X_{TP}}{X_{TP}+X_{FP}},\tag{8}$$

式中，X_{TP}为正确预测为正样本的数目，X_{FP}为错误预测为正样本的数目。召回率表示所有正样本数目中正确预测出的正样本所占比例，可表示为

$$R=\frac{X_{TP}}{X_{TP}+X_{FN}},\tag{9}$$

式中，X_{FN}为错误预测为负样本的数目。mAP表示多个类别AP的平均值，可表示为

$$X_{mAP}=\frac{1}{C}\cdot\sum_{n=1}^{N}A\left[P(n)\right],\tag{10}$$

式中，C为类别数，A为求平均值。变阈值类平均精度(mAP，0.50∶0.95)表示阈值在0.50～0.95范围内每次增加0.05时的mAP。参数增量(N_P)为以YOLOv5s网络参数量为基准(0)的网络结构参数增量。

4.2　消融实验及分析

实验用X光安检图像数据集进行训练，设置网络的初始学习率为0.001，采用带动量的SGD和Adam优化算法，训练迭代次数为300 epochs，batchsize为64。

将SA模块、SAd模块和E-SAd模块分别加入YOLOv5s网络的相同位置进行训练，得到不同网络的性能如表2所示。可以发现：相比YOLOv5s网络，加入SA模块的YOLOv5s+SA网络mAP提升了1.5个百分点；加入SAd模块的YOLOv5s+SAd网络mAP提升了1.7个百分点；将7×7卷积替换为3×3空洞卷积的SAd模块在保持性能优势的同时，参数增量约为YOLOv5s+SA的一半。此外，相比YOLOv5s+SAd网络，加入ERF扩大感受野后再通过SA模块提取轮廓、边缘及位置信息的YOLOv5s+E-SAd网络mAP提升了1.5个百分点，但由于存在多个卷积操作，该网络的参数增量较多。

表2　SA模块的对比结果

Table 2　Comparison results of SA modules

Module	mAP /%	(mAP, 0.50∶0.95) /%	N_P
YOLOv5s	87.2	55.9	0
YOLOv5s+SA	88.7	56.3	2048
YOLOv5s+SAd	88.9	56.4	940
YOLOv5s +E-SAd	**90.4**	**57.6**	**17832**

将SE模块、ECA模块与改进后的iECA模块(k=3、5)分别加入YOLOv5s网络同一位置中进行训练，得到不同网络的性能如表3所示。可以发现：加入SE模块筛选了特征通道的YOLOv5s+SE网络比YOLOv5s网络的mAP提升了1.6个百分点；构建了特征通道间相关性的YOLOv5s+ECA网络由于舍弃了全连接层，参数量仅为YOLOv5s+SE网络的一半，mAP比YOLOv5s+SE网络高1.3个百分点；在YOLOv5s+ECA网络基础上进一步优化的YOLOv5s+iECA网络，mAP相较YOLOv5s+ECA和YOLOv5s网络分别提升了1.1个百分点和4.3个百分点。此外，增大k值并不能优化网络性能，反而会增加网络计算量，因此本算法在iECA模块中使用k=3的一维卷积。

表3　通道注意力模块对比结果

Table 3　Comparison results of channel attention module

Module	mAP /%	(mAP, 0.5∶0.95) /%	N_P
YOLOv5s	87.2	55.9	0
YOLOv5s+SE	88.1	56.3	**43008**
YOLOv5s+ECA	89.4	58.7	22868
YOLOv5s+iECA(k=3)	**91.5**	**59.5**	25668
YOLOv5s+iECA(k=5)	91.4	59.2	32786

将E-SAd、iECA、ASFF模块以6种不同的组合方式加入YOLOv5s网络中进行消融实验，得到不同网络的mAP如表4所示。由实验1、实验2可知，注意力机制对于复杂的X光图像具有更强的学习能力，使网络达到更好的检测精度，mAP相比原始网络提升了5.3个百分点。由实验1、实验3可知，同时融合三种不同深度特征的自适应特征融合策略，能使网络捕获到更丰富的空间上下文信息并提升网络性能。由实验6可知，结合注意力机制与特征融合策略的YOLOv5s+ASFF+E-SAd+iECA网络性能最好，mAP达到了94.5%，比原始网络高7.3个百分点。这表明融合浅层空间位置信息与深层语义信息后，再通过通道注意力加权筛选特征通道，能有效减少浅层信息带来的噪声，显著提升网络性能。

表4　各模块消融实验对比结果

Table 4　Comparison results of ablation experiments of each module

No.	Modules	ASFF	E-SAd	iECA	mAP /%
1	YOLOv5s	×	×	×	87.2
2	YOLOv5s+E-SAd+iECA	×	√	√	92.5
3	YOLOv5s+ASFF	√	×	×	91.3
4	YOLOv5s+ASFF+E-SAd	√	√	×	92.2
5	YOLOv5s+ASFF+iECA	√	×	√	92.7
6	YOLOv5s+ASFF+E-SAd+iECA	√	√	√	**94.5**

4.3　目标检测实验及分析

为了验证YOLOv5s-AFA网络的性能，在相同的X光安检图像数据集上对Faster RCNN、RetinaNet、YOLOv4、PP-YOLOv2、YOLOv5s、YOLOv5x网络模型以及文献[16]和文献[17]中的网络模型分别进行训练，得到不同网络的精确率、召回率、mAP、模型大小如表5所示。从检测精度上来看：YOLOv5x网络达到了最高的P、R、mAP，其次是YOLOv5s-AFA网络，评价指标仅略低于YOLOv5x；相较于PP-YOLOv2、YOLOv4、B-YOLO等网络，YOLOv5s-AFA网络在检测精度上明显更具优势。在模型大小方面：虽然YOLOv5s-AFA网络的mAP比YOLOv5x网络低1.1个百分点，但其模型大小仅为YOLOv5x网络的1/12；与检测精度较好的RetinaNet、YOLOv4、B-YOLO网络相比，YOLOv5s-AFA网络在模型大小和检测精度的平衡上更具优势。

表5　不同网络的检测结果

Table 5　Detection results of different networks

Network	Backbone	P	R	mAP /%	Module size /m
Faster RCNN	VGG	0.874	0.759	86.8	160
RetinaNet	ResNet+FPN	0.904	0.790	90.0	140
YOLOv4	CSP DarkNet53	0.920	0.812	91.7	240
PP-YOLOv2	ResNet50-vd	0.949	0.865	93.4	83
YOLOv5s	CSP DarkNet	0.890	0.782	87.2	22
YOLOv5x	CSP DarkNet	**0.968**	**0.889**	**95.6**	**320**
B-YOLO	Dark Net+CSP	0.907	0.800	90.4	39
YOLOv5+	GhostBottleneck	0.896	0.795	88.2	24
YOLOv5s-AFA	CSP DarkNet	0.967	0.871	94.5	26

检测精度较高的YOLOv4、PP-YOLOv2、YOLOv5x、YOLOv5s-AFA网络与YOLOv5s网络训练时的Loss曲线和mAP如图10所示。可以发现：五种网络的Loss值最终都达到收敛，且YOLOv5s-AFA网络在Loss曲线与mAP上均表现出优异的性能；YOLOv5s-AFA网络的Loss为0.20156，仅次于YOLOv5x网络，其余网络最终的Loss值都在0.20～0.25范围内；YOLOv5s-AFA网络最终的mAP比YOLOv4、PP-YOLOv2、YOLOv5s网络分别高2.8个百分点、1.1个百分点、7.3个百分点，略低于YOLOv5x网络。

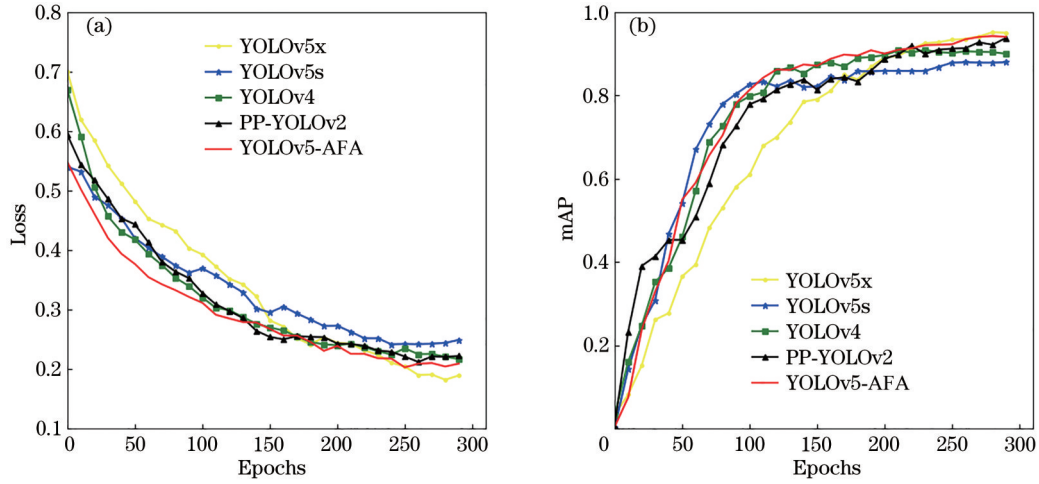

图10　五种网络的Loss与mAP。(a)Loss曲线；(b)mAP曲线

Fig. 10　Loss and mAP of five networks. (a) Loss curve; (b) mAP curve

用五种网络对图8(a)～图8(d)进行检测，结果如图11所示。可以发现，漏检的情况出现在图11(a4)、图11(b4)、图11(d1)以及图11(d4)中，原因是图11(a4)、图11(b4)、图11(d4)左下角的Knife被Wrench和Gun遮挡，而图11(d1)中的Knife受到背景的严重干扰。对比图11(a2)～图11(e2)中Knife的置信度得分可以发现，YOLOv5s-AFA网络的性能仅次于YOLOv5x网络，高于其他三种网络。

综上所述，YOLOv5s-AFA网络对受背景干扰、相互重叠遮挡的较小目标具有良好的检测效果。但当目标物体的成像角度变化较大时，如图11(e1)中最左边倾斜的Wrench目标特征明显较少，且背景中的线条对目标特征影响较大，网络依然存在漏检情况，后续会针对目标成像角度造成的漏检问题对网络做出进一步改进。

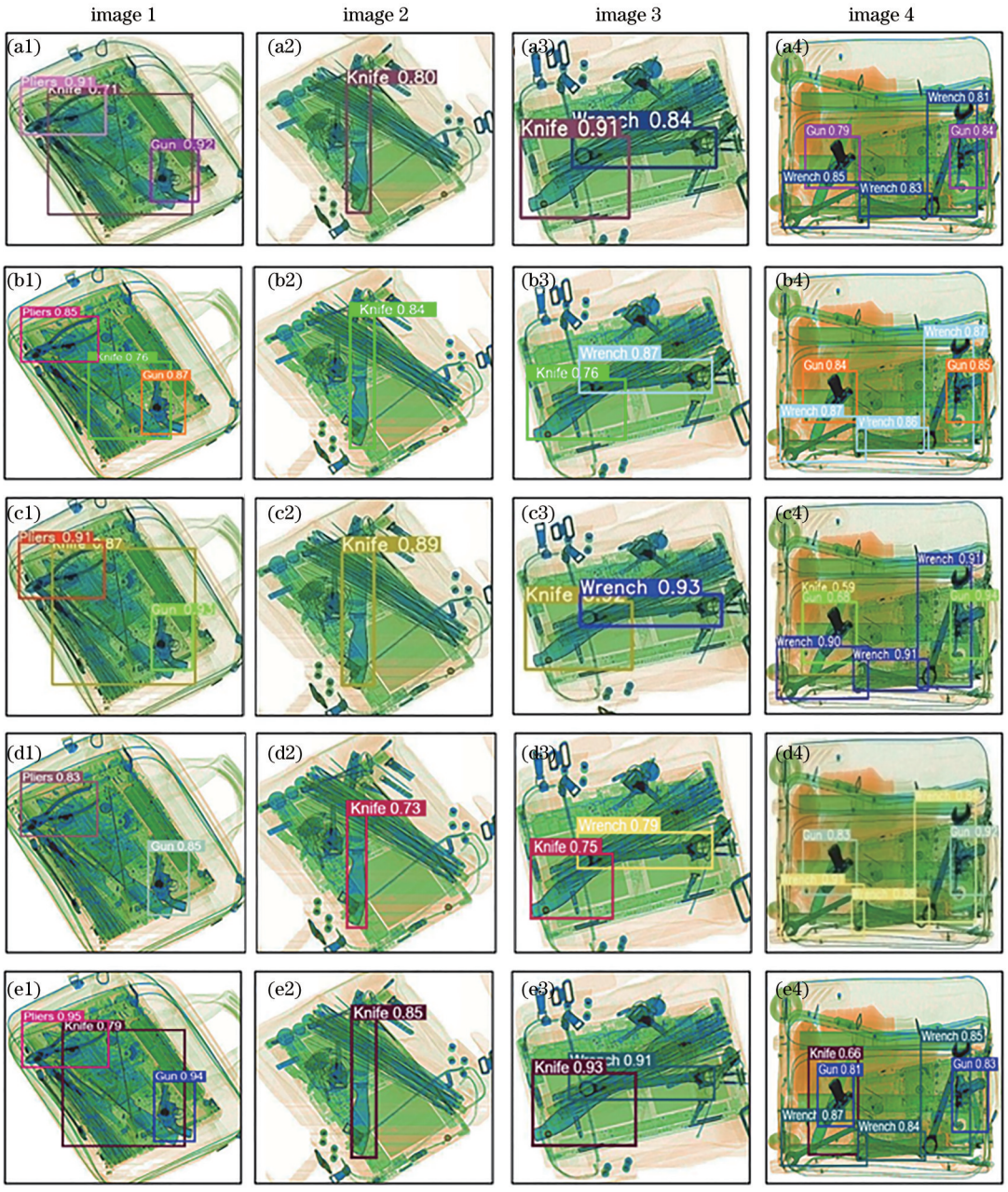

图 11　X 光图像检测结果对比。(a)YOLOv4;(b)PP-YOLOv2;(c)YOLOv5x;(d)YOLOv5s;(e)YOLOv5s-AFA

Fig. 11　Comparison of detection results for X-ray images. (a) YOLOv4; (b) PP-YOLOv2; (c) YOLOv5x; (d) YOLOv5s; (e) YOLOv5s-AFA

5　结　　论

针对 X 光安检图像中背景复杂、目标相互重叠遮挡、目标多尺度的检测难点，以 YOLOv5 作为基础网络结构，提出了适用于 X 光安检图像的 YOLOv5s-AFA 网络。该网络在浅层加入空间注意力机制提取空间位置信息，用多层特征融合的策略将浅层的空间注意力与深层的空间上下文和语义信息相结合，再由通道注意力机制对特征信息进行筛选，使网络学习到最充分的特征。此外，通过扩大浅层输入注意力模块的感受野以及优化空间注意力模块降低了计算量，改进了通道注意力模块，使其能够提取到突出的特征信息，用自适应方法融合多层特征提升了特征金字塔的有效性。通过消融实验验证了融合浅层空间注意力与深层通道注意力的 YOLOv5s-AFA 网络能获得最佳的性能指标，在实验使用的数据集上，YOLOv5s-AFA 网络能以 26.3 m 的模型大小达到 94.5% 的 mAP。

参　考　文　献

[1]　陈志强，张丽，金鑫. X 射线安全检查技术研究新进展[J]. 科学通报，2017，62(13): 1350-1365.
Chen Z Q, Zhang L, Jin X. Recent progress on X-ray security inspection technologies[J]. Chinese Science Bulletin, 2017, 62(13): 1350-1365.

[2] Mery D. Computer vision technology for X-ray testing[J]. Insight-Non-Destructive Testing and Condition Monitoring, 2014, 56(3): 147-155.

[3] Jaccard N, Rogers T W, Morton E J, et al. Using deep learning on X-ray images to detect threats[D]. London: University College London, 2016.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[5] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[6] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

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

[8] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[10] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[11] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2021-06-04]. https://arxiv.org/abs/1804.02767.

[12] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2021-03-06]. https://arxiv.org/abs/2004.10934.

[13] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 14-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1571-1580.

[14] 李彬, 汪诚, 吴静, 等. 改进YOLOv4算法的航空发动机部件表面缺陷检测[J]. 激光与光电子学进展, 2021, 58(14): 1415004.
Li B, Wang C, Wu J, et al. Surface defect detection of aeroengine components based on improved YOLOv4 algorithm[J]. Laser & Optoelectronics Progress, 2021, 58(14): 1415004.

[15] 谈世磊, 别雄波, 卢功林, 等. 基于YOLOv5网络模型的人员口罩佩戴实时检测[J]. 激光杂志, 2021, 42(2): 147-150.
Tan S L, Bie X B, Lu G L, et al. Real-time detection for mask-wearing of personnel based on YOLOv5 network model[J]. Laser Journal, 2021, 42(2): 147-150.

[16] 刘建男, 聂凯. 基于改进YOLOv3的单阶段目标检测算法[J]. 电光与控制, 2021, 28(9): 30-33, 69.
Liu J N, Nie K. One-stage object detection algorithm based on improved YOLOv3[J]. Electronics Optics & Control, 2021, 28(9): 30-33, 69.

[17] 徐诚极, 王晓峰, 杨亚东. Attention-YOLO:引入注意力机制的YOLO检测算法[J]. 计算机工程与应用, 2019, 55(6): 13-23, 125.
Xu C J, Wang X F, Yang Y D. Attention-YOLO: YOLO detection algorithm that introduces attention mechanism[J]. Computer Engineering and Applications, 2019, 55(6): 13-23, 125.

[18] 李浪怡, 刘强, 邹一鸣, 等. 基于改进YOLOv5算法的轨面缺陷检测[J]. 五邑大学学报(自然科学版), 2021, 35(3): 43-48, 54.
Li L Y, Liu Q, Zou Y M, et al. Rail surface defect detection based on improved YOLOv5 algorithm[J]. Journal of Wuyi University (Natural Science Edition), 2021, 35(3): 43-48, 54.

[19] 张友康, 苏志刚, 张海刚, 等. X光安检图像多尺度违禁品检测[J]. 信号处理, 2020, 36(7): 1096-1106.
Zhang Y K, Su Z G, Zhang H G, et al. Multi-scale prohibited item detection in X-ray security image[J]. Journal of Signal Processing, 2020, 36(7): 1096-1106.

[20] 张震, 李浩方, 李孟州. YOLO算法在安检异常图像中的研究[J]. 计算机工程与应用, 2020, 56(21): 187-193.
Zhang Z, Li H F, Li M Z. Research on YOLO algorithm in abnormal security images[J]. Computer Engineering and Applications, 2020, 56(21): 187-193.

[21] 郭守向, 张良. YOLO-C:基于单阶段网络的X光图像违禁品检测[J]. 激光与光电子学进展, 2021, 58(8): 0810003.
Guo S X, Zhang L. YOLO-C: one-stage network for prohibited items detection within X-ray images[J]. Laser & Optoelectronics Progress, 2021, 58(8): 0810003.

[22] Liu Y D, Wang Y T, Wang S W, et al. CBNet: a novel composite backbone network architecture for object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11653-11660.

[23] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8759-8768.

[24] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 658-666.

[25] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[26] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: efficient

channel attention for deep convolutional neural networks [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11531-11539.

[27] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[EB/OL]. (2018-07-17)[2021-06-05]. https://arxiv.org/abs/1807.06521.

[28] 梁添汾, 张南峰, 张艳喜, 等. 违禁品X光图像检测技术应用研究进展综述[J]. 计算机工程与应用, 2021, 57(16): 74-82.
Liang T F, Zhang N F, Zhang Y X, et al. Summary of research progress on application of prohibited item detection in X-ray images[J]. Computer Engineering and Applications, 2021, 57(16): 74-82.

[29] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2015-11-23)[2021-05-06]. https://arxiv.org/abs/1511.07122.

[30] Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection[EB/OL]. (2017-11-21)[2021-06-03]. https://arxiv.org/abs/1711.07767.

[31] Liu S T, Huang D, Wang Y H. Learning spatial fusion for single-shot object detection[EB/OL]. (2019-11-21)[2021-05-06]. https://arxiv.org/abs/1911.09516.