计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(12)-2984-15
doi: 10.3778/j.issn.1673-9418.2211102

三维坐标注意力路径聚合网络的目标检测算法

涂小妹[1,2],包晓安[2],吴　彪[3+],金瑜婷[2,4],张庆琪[5]
1. 浙江广厦建设职业技术大学 建筑工程学院,浙江 东阳 322100
2. 浙江理工大学 计算机科学与技术学院(人工智能学院),杭州 310018
3. 浙江理工大学 理学院,杭州 310018
4. 浙江广厦建设职业技术大学 信息学院,浙江 东阳 322100
5. 山口大学 东亚研究科,日本 山口 753-8514
+ 通信作者 E-mail: biaowuzg@zstu.edu.cn

摘　要:针对YOLO系列算法在实际工业应用中存在对目标预测框定位不够准确,难以适用于对定位要求较高的现实场景的问题,提出了三维坐标注意力路径聚合网络的目标检测算法YOLO-T。首先,采用短连接方式对路径聚合特征金字塔的跨层特征进行融合,保留其浅层语义信息;其次,基于坐标注意力机制提出了三维坐标注意力(TDCA)模型,利用该模型对路径聚合特征金字塔内的特征进行注意力加权(TPA-FPN),保留有用信息和去除冗余信息;然后,改进了标签分配策略中简单最优传输分配(SimOTA)的损失矩阵计算方法,在保证不损失效率的同时增强了性能;最后,利用Depthwise Separable Conv改进了主干特征提取网络中的卷积模块使模型轻量化。实验结果表明:该算法在PASCAL VOC2007+2012数据集上,检测准确率mAP@0.50比YOLOX-S提高了1.3个百分点,mAP@0.50:0.95提高了3.8个百分点;在COCO2017数据集上平均检测精度mAP@0.50:0.95提高了2.4个百分点。

关键词:目标检测;三维坐标注意力(TDCA);注意力路径聚合特征金字塔(TPA-FPN);YOLOX-S算法;改进SimOTA策略
文献标志码:A　　**中图分类号:**TP391

Object Detection Algorithm for 3D Coordinate Attention Path Aggregation Network

TU Xiaomei[1,2], BAO Xiao'an[2], WU Biao[3+], JIN Yuting[2,4], ZHANG Qingqi[5]
1. School of Civil Engineering and Architecture, Zhejiang Guangsha Vocational and Technical University of Construction, Dongyang, Zhejiang 322100, China
2. School of Computer Science and Technology (School of Artificial Intelligence), Zhejiang Sci-Tech University, Hangzhou 310018, China
3. School of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
4. School of Informatics, Zhejiang Guangsha Vocational and Technical University of Construction, Dongyang, Zhejiang 322100, China
5. The Graduate School of East Asian Studies, Yamaguchi University, Yamaguchi 753-8514, Japan

基金项目:浙江省重点研发计划项目(2020C03094);国家级大学生创新创业训练计划项目(202010338024);浙江省教育厅一般科研项目(Y202250677,Y202250706,Y202147659,Y202250679)。
This work was supported by the Key Research and Development Program of Zhejiang Province (2020C03094), the National College Students Innovation and Entrepreneurship Training Program (202010338024), and the General Scientific Research Project of the Department of Education of Zhejiang Province (Y202250677, Y202250706, Y202147659, Y202250679).
收稿日期:2022-11-25　　**修回日期:**2023-02-24

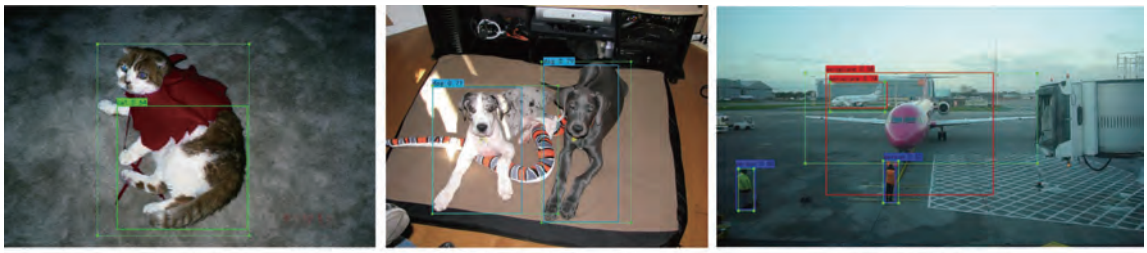
Abstract: In practical industrial applications, YOLO series algorithms are not accurate enough to locate the object prediction boxes, and it is difficult to apply to realistic scenarios with high positioning requirements. The object detection algorithm YOLO-T of the three-dimensional coordinate attention path aggregation network is proposed. Firstly, the shortcut connection method is used to fuse the cross-layer features of the path aggregation feature pyramid to retain its shallow semantic information. Secondly, based on the coordinate attention mechanism, a three-dimensional coordinate attention (TDCA) model is proposed, which is used to pay attention weight to the features in the path aggregation feature pyramid (TPA-FPN (TDCA path aggregation feature pyramid networks)) to retain useful information and remove redundant information. Thirdly, the loss matrix calculation method of SimOTA (simplify optimal transport assignment) in the label allocation strategy is improved, which enhances the performance while ensuring no loss of efficiency. Finally, Depthwise Separable Conv is used to improve the convolution module in the backbone feature extraction network to make the model lightweight. Experimental results show that the detection accuracy mAP@0.50 of the algorithm is 1.3 percentage points higher than that of YOLOX-S on the PASCAL VOC2007+2012 dataset, and the mAP@0.50:0.95 is improved by 3.8 percentage points. The average detection accuracy mAP@0.50:0.95 is improved by 2.4 percentage points on the COCO2017 dataset.

Key words: object detection; three-dimensional coordinate attention (TDCA); TDCA path aggregation feature pyramid networks (TPA-FPN); YOLOX-S algorithm; improved SimOTA strategy

目标检测一直以来是计算机视觉领域的研究热点之一，其任务是返回给定图像中的单个或多个特定目标的类别与矩形包围框坐标[1-3]。目前两大主流目标检测算法：(1)基于候选区域的双阶段目标检测算法，以RCNN(region-CNN)为代表，双阶段检测算法准确率高，但是训练和推理阶段速度慢，不能满足实时要求[4-5]；(2)基于直接回归的单阶段目标检测算法，以SSD(single shot multi-box detector)和YOLO(you only look once)为代表[6-10]，单阶段检测算法在准确率和运行速度上能达到一个均衡，是目前目标检测中使用较多的一种检测框架[11-12]。本文主要以单阶段YOLO系列算法为研究基础，针对现实场景中对目标预测框定位要求较高的场景，提出了一种检测精度较高、定位较准确的检测模型(YOLO-T)。

YOLO系列检测框架主要分为主干网络、特征融合网络、特征解码网络。主干网络提取特征，特征融合网络融合多层语义特征信息，特征解码网络根据具体任务解码网络的输出。为了充分利用主干网络提取的特征，2017年，Lin等人[13]提出了特征金字塔网络(feature pyramid networks，FPN)，用于构建多尺度特征获取高级语义信息。FPN以及基于FPN的改进版[14-18]在单阶段检测算法上表现出不错的效果。2018年，Liu等人[14]考虑到网络浅层特征信息对于目标分割的重要性，提出了PANet(path aggregation network)网络，该网络从一个多尺度特征金字塔中捕获远程浅层特征，提高了模型检测精度。2020年，Tan等人[15]在PANet的基础上提出了BiFPN(bi-directional feature pyramid network)，该网络在每个层级添加残差连接进行反复堆叠来融合特征。2022年，Luo等人[17]提出了通道增强特征金字塔网络(channel enhancement feature pyramid network，CE-FPN)，该网络既实现了通道增强又实现了上采样的亚像素跳跃融合方法。以上这些网络模型检测准确率较高，但模型推理速度较慢，参数量较大，用于YOLO网络会使其失去实时性。于是许多基于YOLO系列的轻量级模型应运而生。2021年，Hu等人[19]将YOLOv3-Tiny网络中的卷积层替换为深度分布偏移卷积和移动反向瓶颈卷积，并设计渐进式通道级剪枝算法在保持检测性能的同时减少了参数量和计算成本。2022年，邱天衡等人[20]基于YOLOv5网络提升检测精度的同时，使用Ghost模块对网络进行轻量化，减少模型复杂度和参数量。2022年，杨小冈等人[21]在基于改进YOLOv5的基础上，使用深度可分离卷积以及对网络进行迭代通道剪枝，以降低模型的参数量和计算量。

YOLO系列网络在实际工业应用中备受青睐，2021年，Ge等人[10]提出了YOLOX，在网络宽度和深度不断递增的过程中，按照主干特征提取网络大小，YOLOX可以分为S、M、L、X。YOLOX-S使用的主干特征提取网络网络最小，模型更轻量化，但在实际工业应用场景中发现YOLOX-S对目标边界框的回归不够准确，如图1所示。YOLOX-M/L/X模型随着网络宽度和深度的加深，模型具有更好的检测和识



(a) IOU值为0.55
(a) IOU value is 0.55

(b) IOU值为0.69
(b) IOU value is 0.69

(c) IOU值为0.53
(c) IOU value is 0.53

图1 检测框与真实框的IOU

Fig.1 IOU of prediction box and ground truth box

别性能,但会受到硬件条件的制约,难以满足对检测实时性和目标框回归准确率要求都很高的应用场景。针对这个问题,本文基于YOLOX-S算法提出了一种检测精度较高、定位较准确的目标检测算法YOLO-T。

(1)在路径聚合特征金字塔网络(path aggregation feature pyramid networks,PAFPN)中,本文提出了采用shortcut连接方式进行跨层特征之间融合,利用TDCA(three-dimensional coordinate attention)对PAFPN内特征进行注意力加权的方法。该方法不仅能将浅层特征传递到特征解码网络中,保留浅层语义信息,又解决了融合浅层特征信息的特征金字塔网络存在信息冗余的问题。

(2)考虑到坐标注意力机制(coordinate attention,CA)[22]只在 X 和 Y 方向进行特征聚合,而忽略了通道 Z 方向的特征加权,本文在CA的基础上,提出了三维坐标注意力(TDCA),其在特征的 X、Y 和 Z 三个方向上进行注意力增强,有效地将空间坐标信息和通道特征信息整合到生成的注意权重中。

(3)在正负样本标签分配策略中,本文沿用了更精准的SimOTA采样策略,但在cost代价函数中,使用了软标签高质量焦点损失(soft label quality focal loss,soft-QFL)和GIOULoss联合作为cost代价损失以及网络的分类和回归损失,通过在目标区域采集高质量的样本来有效地加速模型收敛。

1 相关工作

YOLO系列算法直接对预测的目标物体进行回归,在精度要求不高的情况下,速度能达到实时检测。经过不断研究发展,2021年,Ge等人[10]提出了YOLOX,YOLOX由Backbone、Neck和Head等部分组成。Backbone采用CSPDarknet提取图片的特征信息;Neck部分采用PAFPN的特征金字塔结构,实现不同尺寸特征信息的传递,解决多尺度问题;Head部分采用解耦头,分别计算定位、分类和置信度任务,再通过非极大值抑制(non-maximum suppression,NMS)对最终检测结果进行后处理。2022年,汪斌斌等人[23]基于YOLOX检测模型以及迁移学习方法实现了玉米雄穗的高精度识别。2022年,杨蜀秦等人[24]提出了基于改进YOLOX的单位面积麦穗检测方法,利用采样框直接实现了单位面积麦穗计数。YOLOX在实时检测任务中有一个良好的表现,但YOLOX也还有优化的空间,如YOLOX算法的目标预测框定位不够准确,如图1所示,检测框与目标框的IOU较低。图1(a)绿色真实框与带分类置信度绿色预测框的IOU值为0.55,图1(b)左边目标dog的绿色真实框与带分类置信度蓝色预测框的IOU值为0.69,图1(c)红色飞机的绿色真实框与带分类置信度红色预测框的IOU值为0.53。

故本文将YOLOX-S作为研究的基础网络,主要改进了三方面:(1)在网络的Neck部分,采用shortcut连接方式进行跨层特征之间融合,保留浅层语义信息(如边缘轮廓特征);(2)提出了TDCA注意力算法,并利用TDCA注意力对Neck部分的内部特征进行加权融合,通过给特征赋予权重来保留有用信息和去除冗余信息;(3)改进标签分配策略与损失函数,在计算SimOTA的损失矩阵时,采用联合soft-QFL和GIOULoss的计算方法,在保证不损失效率的同时增强了性能。

YOLOX的Neck部分是PAFPN,基于PANet[14]创建了自下而上的FPN增强,加速了底层信息的流动,能快速融合各层语义信息。FPN是利用图像金字塔的方式进行多尺度变化增强,与图像金字塔不同的是,FPN是将主干网络提取的特征图垒成金字搭,使用自上而下的方式进行特征融合,目的是融合高低层语义信息提高特征的表达能力,为网络的输出提

供更多有效信息。基于FPN的改进还有BiFPN、DRFPN（dual refinement feature pyramid networks）、CE-FPN、Aug-FPN（augmented FPN）等[15-18]。其中BiFPN、CE-FPN以及Aug-FPN的结构如图2所示，BiFPN[15]在每个层级添加残差连接进行反复堆叠融合特征。CE-FPN[17]实现了通道增强和上采样的亚像素跳跃融合方法，减少了由于通道缩减而造成的信息丢失。Aug-FPN[18]通过一致监督缩小特征融合前不同尺度特征之间的语义差距，减少了金字塔最高层特征图的信息损失。上述通过反复自上而下和自下而上的特征融合结构来提高检测精度，但这样的结构增加了计算复杂度，损失了检测速度。

本文在改进FPN的同时，还使用深度可分离卷积（depthwise separable convolution）模块代替主干特征提取网络中的基础卷积结构，深度可分离卷积由逐通道卷积（depthwise convolution，DW）和逐点卷积（pointwise convolution，PW）两部分组成。DW是一个卷积核对应特征图的一个通道，一个通道只被一个卷积核卷积，生成的特征图通道数和输入通道数一样。PW与常规卷积运算类似，卷积核尺寸为1×1×$C\times N$，C为上一层的通道数，N为新特征图通道数。因此，在计算量相同的情况下，Depthwise Separable Convolution可以将神经网络层数做得更深，在实际工业应用中能轻量化网络模型，降低深度学习模型对硬件的要求。

注意力机制是认知科学领域的学者发现人类处理信息时采用的机制，后来把这种机制引入到卷积神经网络[25]。SENet（squeeze-and-excitation networks）[26]通过Squeeze和Excitation两个模块得到特征通道的注意力权值，完成通道特征重标定。CBAM（convolutional block attention module）[27]在SENet的基础上使用GAP（global average pooling）和GMP（global max pooling）两个池化进行Squeeze操作，更充分地提取通道特征。之后，有学者提出Dual Attention Network[28]、Self-calibrated Convolutions[29]和Strip pooling[30]等。2021年，周勇等人[31]提出了弱语义注意力的遥感图像可解释目标检测，利用弱语义分割网络产生强化目标特征的注意力权重值，抑制背景噪声。2022年，李飞等人[32]提出了混合域注意力YOLOv4的输送带纵向撕裂多维度检测，改进了轻量级网络MobileNetv3的特征提取性能。2022年，王玲敏等人[33]提出了一种改进YOLOv5的安全帽佩戴检测方法，该算法在YOLOv5的主干网络中添加CA注意力机制，将位置信息嵌入到通道注意力当中，使网络可以在更大区域上进行注意。本文算法是在CA[22]注意力基础上提出了TDCA注意力算法，在CA的基础上捕获了通道感知注意力特征，提高了注意力的信息融合。

2 本文算法

2.1 网络整体结构

本文以YOLOX-S为基础网络提出了改进网络YOLO-T，使用TPA-FPN（TDCA path aggregation feature pyramid networks）和TDCA改进了网络的Neck部分来提高网络的回归精度和收敛速度，使用Depthwise Separable Conv模块改进Backbone中的卷积结构来降低网络模型的复杂度，网络模型结构如图3所示。YOLO-T在特征融合部分利用Backbone中不同位置的3个特征层，分别位于中间层、中下层、底层。中间层保留了较多的浅层特征信息（如轮廓、纹理和颜色等）。中下层保留了一些属性特征（如某一时刻目标的状态）。底层则保留了高级语义信息，高层语义性越强，模型的分辨能力也越强，但高层语义信息容易

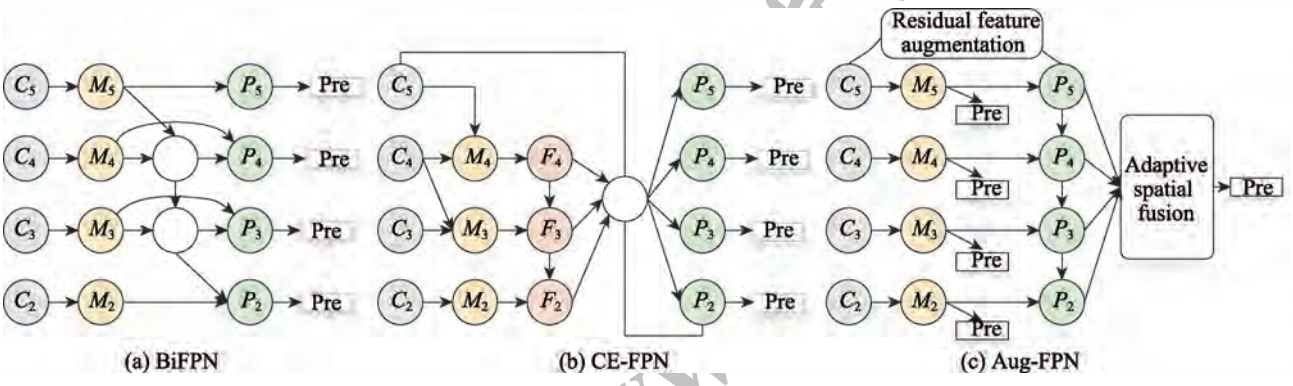


图2 BiFPN、CE-FPN以及Aug-FPN的网络结构图

Fig.2 Network structure diagram of BiFPN, CE-FPN and Aug-FPN

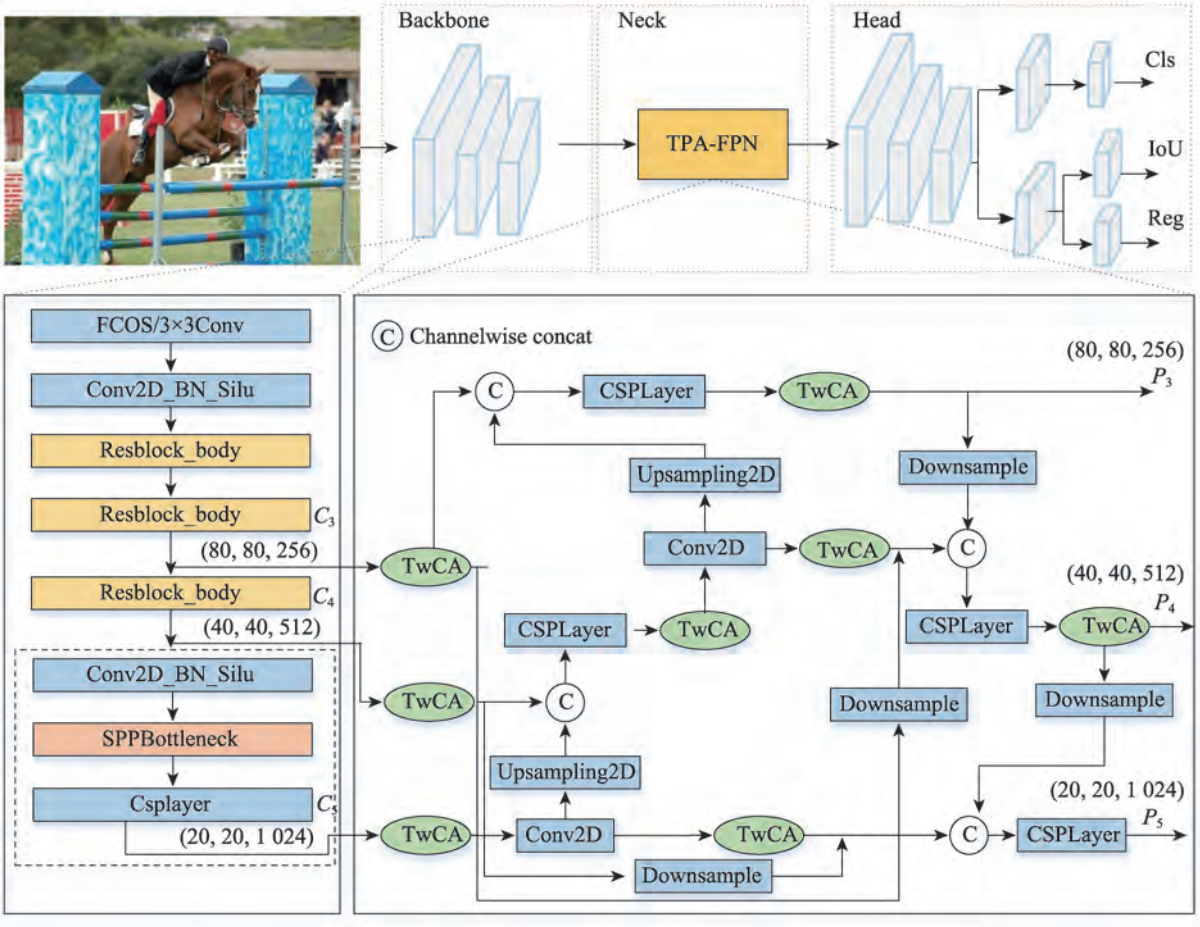


图3 YOLO-T网络结构

Fig.3 YOLO-T network structure

丢失小目标特征。当网络输入大小为640×640的3通道RGB图像时，经实验得出用于特征融合的中间层feat1=(80, 80, 256)、中下层feat2=(40, 40, 512)和底层feat3=(20, 20, 1 024)在保持计算量的情况下效果最好。在feat1、feat2和feat3特征层输入Neck部分之前使用TDCA注意力机制完成特征重标定。Neck部分采用TPA-FPN，Head部分沿用YOLOX-S的解耦头结构。

2.2 三维坐标注意力(TDCA)

注意力机制通过权重重标定，给特征图中的信息赋予不同的权重，达到加强有用信息、抑制无用信息的目的。Hou等人[22]通过在X和Y两个方向上聚合特征的信息，提出了一种为轻量级网络设计的协调注意力机制(coordinate attention，CA)，CA模型的加入可以有效提高模型的收敛速度和监测精度。事实上，在深度卷积神经网络中，特征图不仅有X和Y两个方向的空间信息，还存在通道Z方向的通道信息，而CA注意力机制忽略了Z方向的通道信息。对于特征图数据，分别利用不同的卷积模块学习不同方向的信息权重，然后通过可学习的加权融合方式获取输入特征图的权重，也以此进行X、Y和Z方向的信息交流，充分利用X、Y方向的空间信息和Z方向的通道信息，称为三维坐标注意力(TDCA)，结构如图4所示。

在结构中对于输入特征图$\boldsymbol{F}\in\mathbb{R}^{C\times H\times W}$，$X$、$Y$和$Z$方向的注意力模块计算公式如下所示：

$$\begin{cases}\boldsymbol{M}_C^X(\boldsymbol{F})=AvgPool^{1\times 1\times W}(\boldsymbol{F})\\ \boldsymbol{M}_C^Y(\boldsymbol{F})=AvgPool^{1\times H\times 1}(\boldsymbol{F})\\ \boldsymbol{M}_C^Z(\boldsymbol{F})=\sigma(BaseConv1(GAP^{1\times H\times W}(\boldsymbol{F})))\end{cases}\tag{1}$$

式中，$\boldsymbol{M}_C^X(\boldsymbol{F})\in\mathbb{R}^{C\times H\times 1}$为注意力模块$X$方向的输出，$\boldsymbol{M}_C^Y(\boldsymbol{F})\in\mathbb{R}^{C\times 1\times W}$为$Y$方向的输出，$\boldsymbol{M}_C^Z(\boldsymbol{F})\in\mathbb{R}^{C\times 1\times 1}$为$Z$方向的输出，$AvgPool^{1\times 1\times W}$为池化大小为$1\times 1\times W$的平均池化，$AvgPool^{1\times H\times 1}$同理，$GAP^{1\times H\times W}$为池化大小为$1\times H\times W$的全局平均池化，$\sigma$表示sigmoid激活函数，$BaseConv1$表示一个基础卷积单元，包含卷积层、BN层和卷积层。

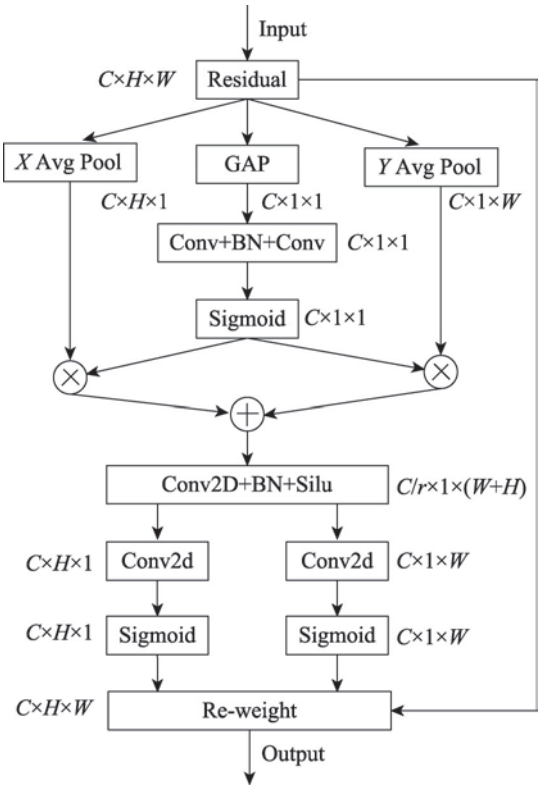


图4 三维坐标注意力TDCA

Fig.4 3D coordinate attention TDCA

经过 X、Y 和 Z 方向的权重特征提取后，将 Z 方向的特征分别与 X 和 Y 方向的特征融合，公式如下所示：

$$\boldsymbol{Z}_X = \boldsymbol{M}_C^X(\boldsymbol{F}) \times \boldsymbol{M}_C^Z(\boldsymbol{F}) \tag{2}$$

$$\boldsymbol{Z}_Y = \boldsymbol{M}_C^Y(\boldsymbol{F}) \times \boldsymbol{M}_C^Z(\boldsymbol{F}) \tag{3}$$

$$\boldsymbol{Z} = BaseConv2(Concat(\boldsymbol{Z}_X^{\mathrm{T}}, \boldsymbol{Z}_Y)) \tag{4}$$

式中，$\boldsymbol{Z}_X \in \mathbb{R}^{C\times H\times 1}$ 和 $\boldsymbol{Z}_Y \in \mathbb{R}^{C\times 1\times W}$ 是 X 与 Z 方向，Y 与 Z 方向的通道权重重标定，将 $\boldsymbol{Z}_X$ 转置与 $\boldsymbol{Z}_Y$ 沿 Y 方向进行 Concat，$\boldsymbol{Z} \in \mathbb{R}^{C/r\times 1\times (W+H)}$ 为 $\boldsymbol{Z}_X$ 和 $\boldsymbol{Z}_Y$ 结合后的结果特征图，其中 r 为通道的系数，取0.5，降低通道数以减少计算量。$BaseConv2$ 表示卷积的基本单元，包含卷积层、BN层和Silu激活函数。将 $\boldsymbol{Z}$ 分割成 $\boldsymbol{Z}_X'$ 和 $\boldsymbol{Z}_Y'$，分别对它们进行卷积和激活操作，可表示为：

$$\boldsymbol{Z}_X'' = \sigma(f_{\mathrm{conv2d}}(\boldsymbol{Z}_X')) \tag{5}$$

$$\boldsymbol{Z}_Y'' = \sigma(f_{\mathrm{conv2d}}(\boldsymbol{Z}_Y')) \tag{6}$$

f_{conv2d} 为1×1卷积为恢复缩放的通道数操作，最后将得到的两个空间解码权重图与输入特征进行点乘，完成特征权重重标定，TDCA输出公式如下所示：

$$\boldsymbol{F}_{\mathrm{out}} = \boldsymbol{F} \times \boldsymbol{Z}_X'' \times \boldsymbol{Z}_Y'' \tag{7}$$

$\boldsymbol{F}_{\mathrm{out}} \in \mathbb{R}^{C\times H\times W}$ 为TDCA网络结构的输出。通过 X 和 Y 方向捕获方向感知和位置感知信息，利用 Z 方向捕获跨通道信息，使模型更加精准地定位和识别感兴趣的目标，能够更加有针对性地提取图像特征，提升图像识别效果。表1是从对比实验角度证明TDCA网络结构的有效性。

表1 各种注意力机制与TDCA在YOLOX-S下对比

Table 1 Comparison of various attention mechanisms and TDCA in YOLOX-S

算法	mAP@0.50:0.95/%	mAP@0.50/%	FPS	模型大小/MB
YOLOX-S(base)	59.0	83.9	53.1	34.4
+SE	60.2	84.0	55.2	36.3
+CBAM	60.7	84.7	48.3	34.5
+CA	62.3	84.8	52.6	34.5
+TDCA	62.7	85.4	49.4	35.3

表1为YOLOX-S与加入SE、CBAM、CA以及TDCA注意力模块后在PASCAL VOC2007+2012数据集上训练测试的精度对比。从表1中可以看出，加入注意力机制后普遍能够提升网络的精度，而在加入TDCA后，在mAP@0.50:0.95指标上比最优的CA机制提高了0.4个百分点，比YOLOX-S算法提高了3.7个百分点。在mAP@0.50指标上，TDCA相较于最优的CA机制提高了0.6个百分点，相比原始算法提高了1.5个百分点。对算法成本进行分析，从表1中可以看出，TDCA与基础网络YOLOX-S相比检测速度和模型大小有所增加，但是与其他注意力机制相比都相差不大。因此，从3个方向提取特征注意力信息是对CA模块有效的改进方向。

相比此前的轻量级网络上的注意力方法，TDCA存在以下优势：首先，它不仅能捕获跨通道的信息，还能捕获方向感知和位置感知的信息，这能使模型更加精准地定位和识别感兴趣的目标；其次，TDCA灵活且轻量，可以很容易地插入经典模块，如Mobile-NeXt[34]提出的sandglass block；最后，作为一个预训练模型，TDCA可以在轻量级网络的基础上给下游任务带来增益，特别是那些存在密集预测的任务。

2.3 TPA-FPN网络结构设计

FPN自顶而下的融合方式极大地利用了高低层特征语义信息，从而提高了特征的表达能力。因此，基于FPN的改进算法常在融合方式上进行创新，例如PANet网络[14]从自顶向下再从自底向上融合方式提高特征融合能力，BiFPN[15]网络利用反复堆叠的方式进行特征融合。上述两种融合方式能有效地保证

特征之间的信息交流，但是特征内的信息重要程度却被忽略了。此外，特征融合过程中的自顶而下和自底向上的融合方式会使特征内的语义信息被稀释，从而会损失特征图内部的一些较重要的信息。本文提出了TPA-FPN网络结构，如图5所示，在PAFPN特征融合网络中采用shortcut连接方式进行跨层特征融合，保留其浅层语义信息。但是，融合了跨层特征的特征图信息存在冗余，于是在PAFPN网络结构中加入TDCA网络结构，对特征内的重要信息进行 X、Y 和 Z 方向的加权，TDCA网络结构通过给特征赋予权重来保留有效信息和去除冗余信息。

主干网络不同阶段的特征图对应的感受野不同，它们表达的信息抽象程度也不一样，在Backbone网络中抽取特征丰富的 C_3、C_4、C_5 三层做特征融合，C_3、C_4、C_5 除了在自顶向下过程中与邻层特征融合之外，还通过短连接（shortcut）进行跨层级特征融合，将保留的浅层信息传递到间隔层。但以这种方式做信息融合，容易产生信息冗余。解决方法是在特征融合之前以及在自顶向下和自底向上特征融合过程中，加入TDCA三维坐标注意力网络，通过注意力调节特征信息的重要程度，提高特征的信息表达能力，从而保留有用信息和去除冗余信息。最后特征融合网络TPA-FPN输出尺度大小为80×80×256、40×40×512、20×20×1 024的特征层 P_3、P_4、P_5，作为目标检测网络中分类和回归特征的依据。

在TPA-FPN特征融合网络中，不同特征层之间自上向下融合需要对尺寸较小的特征进行上采样，自下向上的特征融合过程中需要对尺寸大的特征进行下采样，本文使用的上采样方法为双线性插值法，下采样方法为普通卷积操作，将下采样或上采样后的特征图与底层特征或高层特征进行concat连接。

表2为在PASCAL VOC2007+2012数据集上训练测试的精度对比结果。以YOLOX-S网络结构为基础，在Neck部分分别使用FPN、PAFPN和TPA-FPN作为算法1、算法2和算法3。在算法3的基础上，将主干网络结构中的卷积结构替换成Depthwise Separable Conv模块降低网络模型的复杂度，作为对比算法4。

表2 FPN、PAFPN和TPA-FPN在YOLOX-S下对比

Table 2 Comparison of FPN, PAFPN and TPA-FPN in YOLOX-S

实验方法	mAP@0.50:0.95/%	mAP@0.50/%	FPS	模型大小/MB
算法1	58.3	83.7	67.2	30.1
算法2	59.0	83.9	53.1	34.4
算法3	62.7	85.4	49.4	35.3
算法4	62.3	84.8	65.7	12.6

由表2可知，算法3中的TPA-FPN在推理速度上略低于FPN和PAFPN，但在高交并比要求下TPA-FPN比PAFPN的mAP@0.50:0.95指标提升了3.7个百分点，mAP@0.50指标提升了1.5个百分点。表明跨层级特征融合和利用TDCA网络结构保留有效信

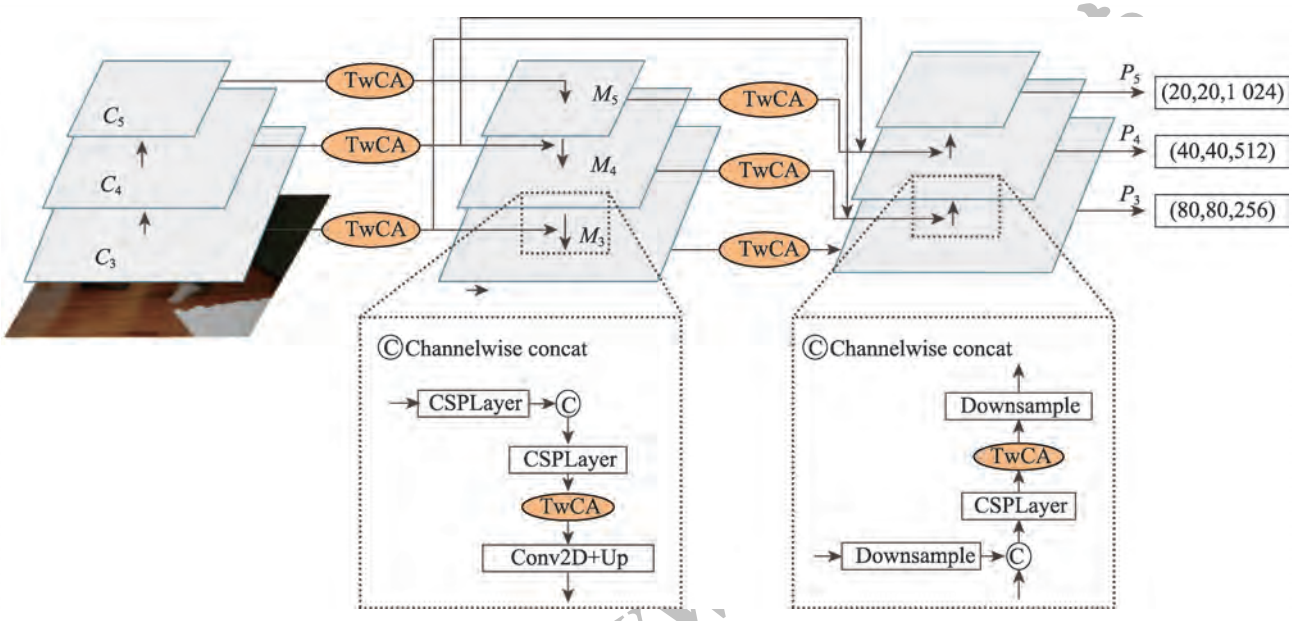


图5 TPA-FPN网络结构

Fig.5 TPA-FPN network structure

息和去除冗余信息能提高网络对边界框的回归精度。算法4使用了Depthwise Separable Conv模块，虽然在准确率上有所降低，但能有效地对网络进行轻量化，模型大小减少了64.3%，由检测速度（FPS）指标可知，模型检测速度提升了33.0%。模型大小的降低和检测速度（FPS）的提升能有效减少在实际应用场景下模型对硬件的要求。参数量的降低能有效减少在实际应用场景下模型对硬件的要求。通过TPA-FPN结构使特征金字塔网络能更好地融合各层语义信息，可以更好地回归目标边界框，契合高交并比下的工业目标检测任务。

2.4 标签分配策略与损失函数

目标检测中预测定位的过程是模型开始训练时先在图像的每个位置生成一系列锚框，网络结构按照一定的规则将锚框分成正负样本，但由于图像中目标的数量有限，这样生成的锚框大部分都是背景，导致模型训练样本不均衡。因此，在正负样本标签分配策略中，本文沿用YOLOX-S中更精准的SimOTA采样策略，但在cost代价函数中，本文使用了soft-QFL作为cost代价损失和分类损失。考虑到one-hot标签中0和1的绝对情况下，本文的soft-QFL分为两种情况：(1)half soft-QFL将正类别中的1使用IOU的值代替，其他类别的值使用0；(2)soft-QFL在正类别中的1使用IOU的值代替的情况下，其他类别的值使用$(1-IOU(gt,anchor))/C$，结构如图6所示。在原基础网络中回归损失函数使用的是IOULoss，本文针对IOU存在当锚点与真实框没有相交时，不能反映两者的距离关系，使用GIOULoss作为网络的回归损失。上述改进目的是通过在目标区域采集高质量的样本来有效地加速模型收敛，从而改善目标正负样本标签分配不均衡的问题。改进的SimOTA采样策略如式(8)所示，其中a_i的值作为样本的标签，a_i的值越大表示此锚点更接近真实框。

$$a_i=\begin{cases}IOU(gt,anchor), i=y\\(1-IOU(gt,anchor))/C, i\neq y\end{cases} \tag{8}$$

$IOU(gt,anchor)$表示目标真实框gt与生成的anchor锚点框之间的IOU值，C表示总类别。在分类损失中，将one-hot编码中的1替换为a_i的值，0替换为$(1-IOU(gt,anchor))/C$，这样更能反映出锚点与真实样本的关系，如式(9)所示：

$$L_{cls}=-\sum_{i=1}^{C}\left|a_i-p_i\right|^{\beta}((1-a_i)\ln(1-p_i)+a_i\ln p_i) \tag{9}$$

式中，p_i表示预测为第i类的概率，β为调节参数，

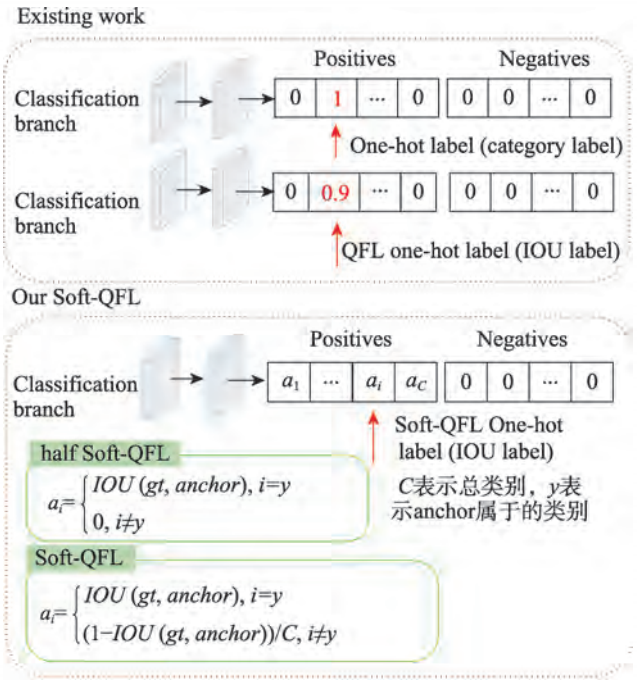


图6 改进的SimOTA采样策略

Fig.6 Improved SimOTA sampling strategy

实验取值为2。当anchor是难分的正样本时，a_i的值偏低，$1-a_i$的值偏高，而L_{cls}在逐渐降低的过程中，网络就相当于增加了难分正负样本的loss权重，使得网络在训练时不会花太多时间在易分的负样本上，加快了模型收敛。

利用IOU指导正负样本标签的分配，再与分类置信度损失函数进行联合预测，在速度无损的情况下能有效地改善正负样本不均衡问题。表3中的消融实验为在PASCAL VOC2007+2012数据集上的精度对比结果，基础网络采用YOLOX-S，“√”代表引入模块，实验环境参数以及网络超参数设置如3.1节所示。

表3 标签分配策略与损失函数的mAP对比

Table 3 Comparison of mAP of label allocation strategy and loss function

组别	half Soft-QFL	Soft-QFL	GIOULoss	mAP@0.50:0.95/%	mAP@0.50/%
1	×	×	×	59.0	83.9
2	√	×	×	57.4	83.2
3	×	√	×	60.6	84.2
4	×	×	√	60.2	84.0
5	√	×	√	60.9	84.6
6	×	√	√	61.6	84.7

在表3中对比1、2组实验发现，使用half Soft-

QFL进行改进的模型在mAP@0.50:0.95和mAP@0.50的指标上有所下降,而通过对比1、2、3组实验可以看出,Soft-QFL的表现比half Soft-QFL在mAP@0.50:0.95指标上提高了3.2个百分点,在mAP@0.50指标上提高了1.0个百分点。分析half Soft-QFL可能是使用IOU指标作为正类别的标签反而削弱了损失函数的表现,Soft-QFL则是在改变正类别时,对负类别也进行了改进,并同时作用于损失函数,从而使模型表现出良好的效果。对比1、4组实验,GIOULoss在mAP@0.50:0.95指标上比原算法提高了1.2个百分点,有一个较好的效果。通过5、6组实验可以看出,soft-QFL+GIOULoss的表现效果最好。

从结构上分析在基于锚框检测的目标检测算法中,使用Soft-QFL和GIOULoss联合能有效改善网络检测精度,使网络训练更稳定,加速网络训练收敛速度。

3 实验结果及分析

3.1 实验环境与参数设置

为了公平分析和评估本文提出的算法性能,实验测试环境配置如下:CPU为Intel® Xeon® Gold 5218R CPU@2.10 GHz,64 GB内存,Ubuntu16.04操作系统,2张GeForce RTX3090型号的显卡。运行环境配置如下:Python版本为3.7,Pytorch版本为1.9.0,CUDA版本为10.2。网络运行的超参数设置如下:网络训练分为冻结训练和解冻训练,冻结训练50个epoch后再进行解冻训练,冻结训练的batch-size设置为64,解冻训练的batch-size设置为32,动量参数为0.937,学习率初始值为0.01,最小值为0.000 1,随着网络的训练,学习率进行余弦退火衰减,解冻阶段训练300个epoch,并使用Adam优化算法更新网络权重。

超参数置信度阈值、NMS阈值的作用是剔除每一类别中的重复预测框,其取值对模型性能有一定影响。通过非极大值抑制(NMS)算法,本文设计了一组超参数置信度阈值和NMS阈值的灵敏度实验。根据NMS算法思想,置信度阈值和NMS阈值过大容易将正确的预测框剔除,过小不能达到去除重复框的效果。实验结果表示,在置信度阈值取值0.45,NMS阈值取值0.50时,本文算法展现出较好的性能。

算法1 非极大值抑制(NMS)算法

输入:$B=\{b_1,b_2,\cdots,b_n\}$, $S=\{s_1,s_2,\cdots,s_n\}$, T_{conf}, T_{nms}。

B表示一组预测框的集合

S表示预测框对应的分类置信度

T_{conf}表示置信度阈值

T_{nms}表示NMS阈值

输出:一组带分数的检测框集合D。

$D \leftarrow \varnothing$
for $s_j \in S$ do
  if $s_j < T_{\text{conf}}$ then
    $B \leftarrow B \backslash \{b_j\}$
    $S \leftarrow S \backslash \{s_j\}$
while $B \neq \varnothing$ do
  $m \leftarrow \text{argmax} S(s_m)$
  $B \leftarrow B \backslash \{b_m\}$
  $s \leftarrow S(s_m)$
  $D \leftarrow D \cup \{\langle b_m, s_m \rangle\}$
  for $b_j \in B$ do
    if $IoU(b_m, b_j) > T_{\text{nms}}$ then
      $B \leftarrow B \backslash \{b_j\}$
    end if
  end for
end while
return D

3.2 评价指标

本实验的评价指标使用平均检测精度(mAP@0.50、mAP@0.50:0.95)和检测速度(FPS)作为模型的衡量标准,平均检测精度能有效地评估模型的性能,包括识别准确率、定位准确率,检测速度能有效地衡量模型的推理性能,是实际工业应用中的重要指标。其中,mAP@0.50表示IOU阈值为0.50时的mAP;mAP@0.50:0.95表示步长为0.05的IOU阈值从0.50到0.95的各个mAP的平均值。mAP@0.50主要体现目标检测模型的识别能力,mAP@0.50:0.95由于IOU最高取值达到了0.95,IOU取值高主要体现目标定位效果以及边界框回归能力。mAP的值与模型的性能呈正相关,FPS表示每秒检测图像的数量,其值越大表示检测速度越快。

mAP表示平均检测精度即P-R曲线下方的面积,P-R曲线是以准确率(Precision)为纵轴,召回率(Recall)为横轴的二维曲线。具体计算公式如式(10)~式(12):

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{12}$$

式中，*TP* 表示为真正例样本数；*FP* 表示为假正例样本数；*FN* 表示为假反例样本数。Precision表示预测样本中的正样本数占所有实际正样本数的比例。Recall表示预测样本中的正样本数占所有预测样本的比例，Precision与Recall呈负相关。

3.3 PASCAL VOC2007+2012数据集对比实验

PASCAL VOC数据集是计算机视觉挑战赛公开的数据集，常被用来检验目标检测模型的性能。PASCAL VOC2007+2012是两个年份公开发布数据集的并集，此数据集更复杂，使用该数据集对模型性能进行验证可增加数据量，同时也更具说明性。该数据集包含20类检测目标，模型的训练集使用PASCAL VOC2007+2012数据集中的train+val部分，共16 551张图像，模型的测试集使用PASCAL VOC 2007数据集中的test部分，共4 952张图像。

为了验证YOLO-T模型的性能，本实验将与以下算法做对比：(1)双阶段目标检测算法Faster R-CNN[35]、Mask R-CNN[4]和Cascade R-CNN[5]；(2)高精度单阶段算法RetinaNet[36]和以SSD[6]为基础改进的ASSD(attentive single shot multibox detector)[37]算法；(3)单阶段无锚框算法FCOS(fully convolutional one-stage object detection)[38]和ATSS(adaptive training sample selection)[39]算法；(4)以YOLO系列为基础的YOLOv3[20]算法以及改进的轻量级算法YOLOv4-mobileNetv2[20]、YOLOv4-ghostNet[20]、YOLOv5-S[20]和YOLOX-S。与以上算法对比结果如表4所示。

由表4可知，本文提出的YOLO-T在检测精度上有着显著优势，在PASCAL VOC2007测试集上mAP@0.50的精度达到了85.2%，相较于基础网络YOLOX-S提高了1.3个百分点，而能体现定位效果和边界框回归能力的mAP@0.50:0.90精度达到了62.8%，相较于基础网络YOLOX-S提高了3.8个百分点，说明YOLO-T网络结构能有效提高预测定位的检测精度；与双阶段检测器相比，mAP@0.50提高了5.6~7.8个百分点；与单阶段经典算法SSD以及基于SSD改进的ASSD算法相比有4.6个百分点和2.2个百分点的提升；与高精度RetinaNet算法以及单阶段无锚框的FCOS和ATSS算法相比，YOLO-T网络结构更展现了其优势，检测精度都有大幅度提升；相比于有相同baseline的轻量化网络YOLOv4-mobileNetv2、YOLOv4-ghostNet、YOLOv5-S和YOLOX-S，虽然检测速度不如基础网络YOLOX-S，但是检测精度上有着明显的优势。总体来看，在检测精度和检测速度兼具的条件下，YOLO-T在众多模型中的表现更加出色。

由于超参数置信度阈值和NMS阈值的取值对模型性能有一定影响，本文设计了一组超参数的灵敏度实验。根据非极大值抑制(NMS)算法，选取了9组数据对模型的性能进行测试。在取值的过程中，阈值过大容易将正确的预测框剔除，过小不能达到去除重复框的效果。因此，本文的置信度阈值从0.40到0.50以0.05为步长递增，NMS阈值从0.30到0.50以0.1为步长递增，实验结果如表5所示。

根据表5的实验结果，不同的超参数置信度阈值和NMS阈值对模型性能有一定影响。当置信度阈值

表4 PASCAL VOC2007测试集上各目标检测算法对比实验

Table 4 Comparative experiment of each object detection algorithm on PASCAL VOC2007 test set

Method	Backbone	Input size	GPU	mAP@0.50/%	mAP@0.50:0.90/%	FPS
Faster R-CNN[35]	ResNet-101	~1 000×600	1080Ti	78.8	—	2.3
Mask R-CNN[4]	ResNet-50	~1 000×600	1080Ti	77.4	—	4.2
Cascade R-CNN[5]	VGGNet	~1 000×600	1080Ti	79.6	—	5.3
RetinaNet[36]	ResNet-101	~640×400	1080Ti	79.4	—	12.4
SSD[6]	ResNet-101	513×513	TitanX	80.6	—	6.8
ASSD[37]	ResNet-101	513×513	TitanX	83.0	—	16.0
FCOS[38]	ResNet-50	1 333×800	1080Ti	73.5	—	17.6
ATSS[39]	ResNet-50	1 333×800	1080Ti	75.2	—	14.9
YOLOv3[20]	DarkNet-53	640×640	RTX 3090	82.4	57.4	55.7
YOLOv4-mobileNetv2[20]	MobileNet	640×640	RTX 3090	82.0	46.8	50.9
YOLOv4-ghostNet[20]	GhostNet	640×640	RTX 3090	80.8	45.4	39.3
YOLOv5-S[20]	—	640×640	RTX 3090	78.4	51.5	74.6
YOLOX-S	CSPDarknet53	640×640	RTX 3090	83.9	59.0	53.1
YOLO-T(ours)	CSPDarknet53	640×640	RTX 3090	85.2	62.8	65.7

表5 超参数置信度阈值、NMS阈值的实验结果

Table 5 Experimental results of hyperparameter confidence threshold and NMS threshold

置信度阈值	NMS阈值	mAP@0.50:0.95/%	mAP@0.50/%
0.40		61.5	84.7
0.45	0.30	62.5	84.2
0.50		61.4	84.9
0.40		61.7	85.0
0.45	0.40	62.3	84.4
0.50		61.5	85.0
0.40		61.8	84.1
0.45	0.50	62.9	85.0
0.50		62.8	85.2

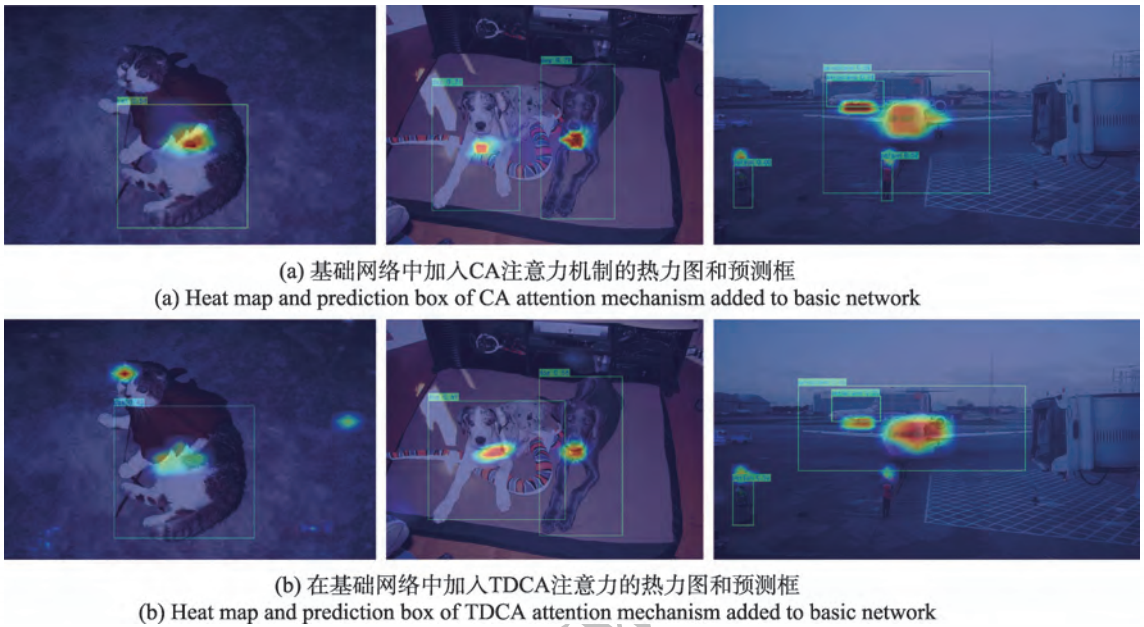
取0.45，NMS阈值取0.50时，模型的mAP@0.50:0.95指标最高；当置信度阈值取0.50，NMS阈值取0.50时，模型的mAP@0.50指标最高。综上实验结果，实验中测试的超参数置信度阈值设为0.45，NMS阈值为0.50。

本文还对VOC数据集的场景图像进行定性评价分析，效果如图7所示。图7中(a)和(b)为经过CA注意力和TDCA注意力处理后的加权热力图，图7(a)中从左到右预测框与真实框的IOU值分别为0.55、0.69和0.53，图7(b)中从左到右预测框与真实框的IOU值分别为0.79、0.99和0.78。从图中可以看出，和CA注意力机制相比，加入TDCA后，网络对检测目标区域的定位和关注程度都获得了提升，证明在Neck部分加入TDCA能更好地融合关键特征信息。

第1组实验对cat进行检测，目标cat由于衣物的遮挡，将身体和头部分开了，此时图7(a)检测器出现定位不准确，只检测到了头部以下的部分，而图7(b)检测定位较图7(a)准确，IOU提高了0.24。从热力图中也可看出，TDCA比CA更能关注到目标的特征。第2组实验中两个目标较聚集，从对比图来看，图7(b)的模型对左侧的目标定位比图7(a)模型准确，IOU提高了0.3。第3组实验对较明显的大目标aircraft进行检测，图7(b)模型定位性能表现得比图7(a)好，IOU提高了0.25，但是在小目标person的检测上，YOLO-T模型检测效果略逊色。但总体来说，YOLO-T在目标定位上要优于基础网络，平均IOU提高了26%，证明网络提取到了更加丰富的语义信息，表现出更好的性能。YOLO-T更适合于定位要求较高的现实场景。

3.4 消融实验

本文算法从TDCA、TPA-FPN和Soft-QFL三方面对YOLOX-S进行改进，为探究各改进方法的有效性，在基线网络YOLOX-S的基础上设计了4组消融实验，实验数据集使用3.3节的PASCAL VOC2007+2012数据集，每组实验所采用的实验环境、网络超参数以及训练技巧均相同，实验结果如表6所示。其中，TPA-FPN代表所提Neck结构，Depthwise Separable Conv模块代表修改主干特征提取网络中的基础



(a) 基础网络中加入CA注意力机制的热力图和预测框
(a) Heat map and prediction box of CA attention mechanism added to basic network

(b) 在基础网络中加入TDCA注意力的热力图和预测框
(b) Heat map and prediction box of TDCA attention mechanism added to basic network

图7 注意力机制CA与TDCA的热力图对比

Fig.7 Heat map comparison of attention mechanism CA and TDCA

表6 各改进模块在YOLOX-S框架下的消融实验

Table 6 Ablation experiment of each improved module under framework of YOLOX-S

TPA-FPN	Depthwise Separable Conv	Soft-QFL	mAP@0.50:0.95/%	mAP@0.50/%
×	×	×	59.0	83.9
√	×	×	62.7	85.4
×	√	×	58.2	83.8
×	×	√	60.6	84.2
√	√	×	62.3	84.8
×	√	√	60.5	83.8
√	×	√	62.4	85.4
√	√	√	62.8	85.2

卷积结构，Soft-QFL代表提出的标签分配策略与损失函数，由于TDCA结构是融入到TPA-FPN结构中的，不对TDCA模块进行消融实验。

由表6可知，以YOLOX-S为基础，加入TPA-FPN模块后mAP@0.50:0.95提升了3.7个百分点，mAP@0.50提升了1.5个百分点，TPA-FPN网络结构融入了TDCA注意力机制，mAP@0.50:0.95指标的提高说明模型对目标预测框的回归能力提高了，使预测的目标框与真实目标框更接近，这对需要更准确定位的回归任务来说，加入TPA-FPN是非常有效的；其次，使用了Soft-QFL改进标签分配策略以及损失函数，mAP@0.50:0.95提升了1.6个百分点，mAP@0.50提升了0.3个百分点。Soft-QFL通过改进标签分配策略以及损失函数来提升网络模型的识别能力，Soft-QFL在几乎不消耗网络的训练和推理性能的基础上，提高了网络检测精度。此外，由于网络结构引入了TPA-FPN模块，模型的复杂度增加，网络检测速度和模型参数量有所增大。

在YOLO-T主干网络中引入Depthwise Separable Conv模块代替普通卷积模块，由表2和表6可知，引入深度可分离卷积模块mAP@0.50:0.95和mAP@0.50准确率只降低了0.8个百分点和0.1个百分点，但模型的参数量和网络复杂度减少了64.3%。最终的YOLO-T网络模型达到了速度和检测精度两方的平衡，并且模型对目标预测框的拟合能力进一步增强，在实际应用中对硬件的要求更小，能被用于需要定位更加准确的工业应用场景中。

3.5 COCO数据集对比实验

为了进一步评估YOLO-T目标检测模型的精度和定位效果，本文在类型更多、图像环境更加复杂的COCO数据集上进行实验。COCO数据集是由微软提供的大型目标检测数据集，具有数据类别多和目标尺寸跨度大等特点。实验中将COCO2017数据集中的训练集随机划分为包含105 539张图像的train和11 727张图像的val，并在包含5 000张图像的COCO2017验证集上进行测试。主要评估不同IOU阈值下的平均精度。其中，不同IOU阈值下的平均精度可以体现模型的定位效果，高IOU阈值代表预测框和真实框重合度的标准更加严格。实验数据中YOLOX-S和YOLO-T通过实验得到，实验环境和参数设置如3.1节所示。

如表7所示，在COCO数据集上，YOLO-T的mAP@0.50:0.95达到了42.0%，较原YOLOX-S提高了2.4个百分点，mAP@0.50提高了0.8个百分点，mAP@0.75提高了2.8个百分点。在不同的IOU阈值下，mAP@0.50:0.95指标涨点最多，这也说明本文算法对预测框定位以及边界框回归能力有着明显的优势。对比其他的检测算法YOLOv5-S、YOLOv3、RefineDet和FAENet，mAP@0.50:0.95也有着显著的提高，由此说明YOLO-T在复杂场景下也具有较好的预测框定位效果和检测性能。

表7 COCO数据集上的对比实验

Table 7 Comparative experiments on COCO dataset

Method	Backbone	mAP@0.50:0.95/%	mAP@0.50/%	mAP@0.75/%
YOLOv3[7]	Darknet53	33.0	57.9	34.4
RefineDet512[40]	ResNet-101	36.4	57.5	39.5
FAENet[41]	—	28.3	47.9	29.7
YOLOv5-S[10]	DarkNet53	36.7	—	—
YOLOX-S	CSPDarknet53	39.6	57.5	41.3
YOLO-T(ours)	CSPDarknet53	42.0	58.3	44.1

4 结束语

本文基于YOLOX-S网络结构提出了一种改进的目标检测算法YOLO-T，目的是改进YOLOX-S算法对目标预测框定位不准确的问题。采用TDCA、TPA-FPN和Soft-QFL结构对网络的精度和目标框边界的回归能力进行提升。使用Depthwise Separable Conv改进Backbone中的卷积模块使模型轻量化，平衡了检测速度和检测精度。在PASCAL VOC2007+2012数据集上，YOLO-T和YOLOX-S相比，模型大小减少了64.3%，检测速度提升了23.7%，mAP@0.50

提高了1.3个百分点，mAP@0.50:0.95提高了3.8个百分点。因此YOLO-T是一种检测精度较高、定位较准确的目标检测模型，适用于对定位要求较高的现实场景。但YOLO-T仍有改进的空间，如Neck部分可以再使用较低分辨率的特征图，可以更好地对小目标进行检测。

参考文献：

[1] 董文轩, 梁宏涛, 刘国柱, 等. 深度卷积应用于目标检测算法综述[J]. 计算机科学与探索, 2022, 16(5): 1025-1042.
DONG W X, LIANG H T, LIU G Z, et al. Review of deep convolution applied to target detection algorithms[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(5): 1025-1042.

[2] 陈科圻, 朱志亮, 邓小明, 等. 多尺度目标检测的深度学习研究综述[J]. 软件学报, 2021, 32(4): 1201-1227.
CHENG K Q, ZHU Z L, DENG X M, et al. Deep learning for multi-scale object detection: a survey[J]. Journal of Software, 2021, 32(4): 1201-1227.

[3] 范丽丽, 赵宏伟, 赵浩宇, 等. 基于深度卷积神经网络的目标检测研究综述[J]. 光学精密工程, 2020, 28(5): 1152-1164.
FAN L L, ZHAO H W, ZHAO H Y, et al. Survey of target detection based on deep convolutional neural networks[J]. Optics and Precision Engineering, 2020, 28(5): 1152-1164.

[4] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 2980-2988.

[5] CAI Z, VASCONCELOS N. Cascade R-CNN: high quality object detection and instance segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(5): 1483-1498.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multiBox detector[C]//LNCS 9905: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Oct 11-14, 2016. Cham: Springer, 2016: 21-37.

[7] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv:1804.02767, 2018.

[8] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv: 2004.10934, 2020.

[9] ZHU X K, LYU S C, WANG X, et al. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Oct 11-18, 2021. Piscataway: IEEE, 2021: 2778-2788.

[10] GE Z, LIU S T, WANG F, et al. YOLOX: exceeding YOLO series in 2021[J]. arXiv:2107.08430, 2021.

[11] 王鹏飞, 黄汉明, 王梦琪. 改进YOLOv5的复杂道路目标检测算法[J]. 计算机工程与应用, 2022, 58(17): 81-92.
WANG P F, HUANG H M, WANG M Q. Complex road target detection algorithm based on improved YOLOv5[J]. Computer Engineering and Applications, 2022, 58(17): 81-92.

[12] 胡皓, 郭放, 刘钊. 改进YOLOX-S模型的施工场景目标检测[J]. 计算机科学与探索, 2023, 17(5): 1089-1101.
HU H, GUO F, LIU Z. Object detection based on improved YOLOX-S model in construction sites[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(5): 1089-1101.

[13] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 936-944.

[14] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-23, 2018. Washington: IEEE Computer Society, 2018: 8759-8768.

[15] TAN M X, PANG R M, LE Q V. EfficientDet: scalable and efficient object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 10778-10787.

[16] MA J L, CHEN B. Dual refinement feature pyramid networks for object detection[J]. arXiv:2012.01733, 2020.

[17] LUO Y H, CAO X, ZHANG J T, et al. CE-FPN: enhancing channel information for object detection[J]. Multimedia Tools and Applications, 2022, 81(21): 30685-30704.

[18] GUO C X, FAN B, ZHANG Q, et al. AugFPN: improving multi-scale feature learning for object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 12592-12601.

[19] HU L N, LI Y F. Micro-YOLO: exploring efficient methods to compress CNN based object detection model[C]//Proceedings of the 13th International Conference on Agents and Artificial Intelligence, Feb 4-6, 2021: 151-158.

[20] 邱天衡, 王玲, 王鹏, 等. 基于改进YOLOv5的目标检测算法研究[J]. 计算机工程与应用, 2022, 58(13): 63-73.
QIU T H, WANG L, WANG P, et al. Research on object detection algorithm based on improved YOLOv5[J]. Computer Engineering and Applications, 2022, 58(13): 63-73.

[21] 杨小冈, 高凡, 卢瑞涛, 等. 基于改进YOLOv5的轻量化航

空目标检测方法[J]. 信息与控制, 2022, 51(3): 361-368.
YANG X G, GAO F, LU R T, et al. Lightweight aerial object detection method based on improved YOLOv5[J]. Information and Control, 2022, 51(3): 361-368.

[22] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, Jun 20-25, 2021. Piscataway: IEEE, 2021: 13713-13722.

[23] 汪斌斌, 杨贵军, 杨浩, 等. 基于YOLO_X和迁移学习的无人机影像玉米雄穗检测[J]. 农业工程学报, 2022, 38(15): 53-62.
WANG B B, YANG G J, YANG H, et al. UAV images for detecting maize tassel based on YOLO_X and transfer learning [J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(15): 53-62.

[24] 杨蜀秦, 王帅, 王鹏飞, 等. 改进YOLOX检测单位面积麦穗[J]. 农业工程学报, 2022, 38(15): 143-149.
YANG S Q, WANG S, WANG P F, et al. Detecting wheat ears per unit area using an improved YOLOX[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(15): 143-149.

[25] 王燕妮, 余丽仙. 注意力与多尺度有效融合的SSD目标检测算法[J]. 计算机科学与探索, 2022, 16(2): 438-447.
WANG Y N, YU L X. SSD object detection algorithm with effective fusion of attention and multi-scale[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(2): 438-447.

[26] HU J, SHEN L, SUN G, et al. Squeeze-and-excitation networks[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 7132-7141.

[27] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//LNCS 11211: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 3-19.

[28] FU J, LIU J, TIAN H J, et al. Dual attention network for scene segmentation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 3146-3154.

[29] LIU J J, HOU Q B, CHENG M M, et al. Improving convolutional networks with self-calibrated convolutions[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 10093-10102.

[30] HOU Q B, ZHANG L, CHENG M M, et al. Strip pooling: rethinking spatial pooling for scene parsing[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 4002-4011.

[31] 周勇, 陈思霖, 赵佳琦, 等. 基于弱语义注意力的遥感图像可解释目标检测[J]. 电子学报, 2021, 49(4): 679-689.
ZHOU Y, CHEN S L, ZHAO J Q, et al. Weakly semantic based attention network for interpretable object detection in remote sensing imagery[J]. Acta Electronica Sinica, 2021, 49(4): 679-689.

[32] 李飞, 胡坤, 张勇, 等. 基于混合域注意力YOLOv4的输送带纵向撕裂多维度检测[J]. 浙江大学学报(工学版), 2022, 56(11): 2156-2167.
LI F, HU K, ZHANG Y, et al. Multi-dimensional detection of longitudinal tearing of conveyor belt based on YOLOv4 of hybrid domain attention[J]. Journal of Zhejiang University (Engineering Science), 2022, 56(11): 2156-2167.

[33] 王玲敏, 段军, 辛立伟. 引入注意力机制的YOLOv5安全帽佩戴检测方法[J]. 计算机工程与应用, 2022, 58(9): 303-312.
WANG L M, DUAN J, XIN L W. YOLOv5 helmet wear detection method with introduction of attention mechanism [J]. Computer Engineering and Applications, 2022, 58(9): 303-312.

[34] ZHOU D Q, HOU Q B, CHEN Y P, et al. Rethinking bottleneck structure for efficient mobile network design[C]//LNCS 12348: Proceedings of the 2020 European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 680-697.

[35] 张娜, 戚旭磊, 包晓安, 等. 基于优化预测定位的单阶段目标检测算法[J]. 浙江大学学报(工学版), 2022, 56(4): 783-794.
ZHANG N, QI X L, BAO X A, et al. Single-stage object detection algorithm based on optimizing position prediction [J]. Journal of Zhejiang University (Engineering Science), 2022, 56(4): 783-794.

[36] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 2999-3007.

[37] YI J R, WU P X, METAXAS D N. ASSD: attentive single shot multibox detector[J]. Computer Vision and Image Understanding, 2019, 189: 102827.

[38] TIAN Z, SHEN C H, CHEN H, et al. FCOS: fully convolutional one-stage object detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 9627-9636.

[39] ZHANG S F, CHI C, YAO Y Q, el al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 9756-9765.

[40] ZHANG S F, WEN L Y, BIAN X, et al. Single-shot refinement neural network for object detection[C]//Proceedings of the 2017 IEEE International Conference on Image Processing, Beijing, Sep 17-20, 2017. Piscataway: IEEE, 2017: 3360-3364.

[41] LI W Q, LIU G Z. A single-shot object detector with feature aggregation and enhancement[C]//Proceedings of the 2019 IEEE International Conference on Image Processing, Taipei, China, Sep 22-25, 2019. Piscataway: IEEE, 2019: 3910-3914.

涂小妹(1995—),女,湖北黄冈人,硕士研究生,主要研究方向为视频图像处理、目标检测。

TU Xiaomei, born in 1995, M.S. candidate. Her research interests include video image processing and object detection.

包晓安(1973—),男,浙江东阳人,硕士,教授,主要研究方向为智能信息处理、人工智能。

BAO Xiao'an, born in 1973, M.S., professor. His research interests include intelligent information processing and artificial intelligence.

吴彪(1989—),男,湖北麻城人,博士,讲师,主要研究方向为计算机视觉、人体姿态估计。

WU Biao, born in 1989, Ph.D., lecturer. His research interests include computer vision and human pose estimation.

金瑜婷(1994—),女,浙江东阳人,硕士研究生,主要研究方向为图像增强、模式识别。

JIN Yuting, born in 1994, M.S. candidate. Her research interests include image enhancement and pattern recognition.

张庆琪(1996—),男,河南温县人,博士研究生,主要研究方向为目标检测、姿态估计、行为识别。

ZHANG Qingqi, born in 1996, Ph.D. candidate. His research interests include object detection, pose estimation and behavior recognition.