第 62 卷　第 5 期
2026 年 9 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 62, No. 5
Sept., 2026

DOI:10.13232/j.cnki.jnju.2026.05.005

基于语义与样本双重增强机制的X光图像定向目标检测研究

刘小亮[1],申富饶[2,3*],王国强[1],徐百乐[2,3],林　毅[1],黄　原[1]

(1.温州商学院信息工程学院,温州,325035;2.计算机软件新技术全国重点实验室(南京大学),南京,210023;
3.南京大学人工智能学院,南京,210023)

摘　要:针对安检X光图像中小目标占比高、目标重叠严重、背景干扰复杂以及目标姿态随机分布等问题,提出一种基于语义与样本双重增强机制的定向目标检测方法.在语义增强方面,以面向旋转框检测的You Only Look Once-Oriented Bounding Box(YOLO-OBB)为主干,引入冻结的Segment Anything Model(SAM)图像编码器作为语义辅助分支,并通过通道注意力融合模块实现异构特征自适应融合,增强模型对目标结构信息和高层语义信息的表征能力.在样本增强方面,设计上下文约束复制粘贴增强策略(Context-Constrained Copy-Paste Augmentation,CCPA),在空间占用约束、碰撞检测和泊松融合机制下对小目标实例进行受控增强,扩展训练样本在尺度、位置和姿态空间中的分布,缓解小目标样本不足与分布长尾问题.同时,引入高斯平滑分布焦点损失,提高旋转框回归稳定性.实验结果表明,提出的方法在安检场景定向目标检测任务中优于基线模型,消融实验验证了各模块的有效性与互补性.

关键词:安检X光图像,定向目标检测,小目标检测,语义增强,样本增强

中图分类号:TP391.41　　**文献标志码:**A

Oriented object detection in X-ray images with semantic and data dual-augmentation

Liu Xiaoliang[1], Shen Furao[2,3*], Wang Guoqiang[1], Xu Baile[2,3], Lin Yi[1], Huan Yuan[1]

(1. School of Information Engineering, Wenzhou Business College, Wenzhou, 325035, China;
2. National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China;
3. School of Artificial Intelligence, Nanjing University, Nanjing, 210023, China)

Abstract: Addressing challenges such as the high proportion of small targets, severe target overlap, complex background interference, and random distribution of target orientations in security-check X-ray images, we propose a directional object detection method based on a dual enhancement mechanism-combining semantic and sample augmentation. In terms of semantic enhancement, we adopt the You Only Look Once-Oriented Bounding Box (YOLO-OBB) framework, which is designed for rotated bounding box detection, as the backbone model. We further introduce a frozen Image Encoder from the Segment Anything Model (SAM) as a semantic auxiliary branch and employ a channel-attention fusion module to achieve adaptive fusion of heterogeneous features, thereby enhancing the model's ability to capture both target structural information and high-level semantic cues. For sample augmentation, we design a Context-Constrained Copy-Paste Augmentation (CCPA) strategy that performs controlled augmentation of small-target instances under constraints on spatial occupancy, collision detection, and Poisson blending. This strategy expands the distribution of training samples across scale, location, and

基金项目:国家自然科学基金(62276127)

收稿日期:2026－07－14

*通信联系人,E-mail:frshen@nju.edu.cn

orientation spaces, alleviating the issues of insufficient small-target samples and long-tail distribution. Additionally, we introduce a Gaussian-smoothed focal loss to improve the stability of rotated bounding box regression. Experimental results demonstrate that our proposed method outperforms the baseline model in directional object detection tasks within security-check scenarios, and ablation studies confirm the effectiveness and complementarity of each module.

Keywords: security X-ray image, oriented object detection, small object detection, semantic enhancement, sample enhancement

在民航、轨道交通及现代物流等安全检查场景中，利用X射线成像技术自动识别违禁品是保障公共安全的核心环节．随着深度学习的发展，目标检测技术在安检图像识别中得到了广泛应用[1−3]．然而，与常规的自然光学图像相比，X光安检图像具有其特有的物理成像难点，给检测算法带来了巨大挑战：首先，X射线的穿透性成像机制导致不同材质、不同深度的物体在二维图像上发生严重的物理重叠与视觉混叠，破坏了目标的结构完整性；其次，安检包裹内部件杂乱，复杂的背景纹理极易对目标边界造成严重干扰；最后，部分违禁品（如刀片、打火机零件等）尺寸微小，在整幅图像中像素占比较低，且在包裹内的摆放呈现出任意的方向性．因此，传统的水平边界框已无法满足需求，模型必须具备高精度的定向边界框（Oriented Bounding Box，OBB）回归能力，以实现对任意方向微小目标的精准定位．

近年来，以YOLOv11/v12-OBB[4−5]为代表的主流单阶段旋转目标检测器在遥感等自然场景的定向检测任务中取得了显著进展，然而，X光安检图像具有目标尺度小、类别分布不均、物体重叠严重、纹理边界弱和成像对比度复杂等特点，使得现有方法在数据分布适应和特征表达能力方面仍面临挑战．一方面，在数据输入分布上，真实复杂的安检数据集中违禁品样本通常呈现明显的长尾效应，特定极端角度和重度遮挡状态下的小目标样本较为匮乏．常规数据增强方法，如随机裁剪或基础Copy-Paste[6]，难以充分模拟X光成像下的复杂遮挡关系和姿态变化，容易引入不自然的局部结构，限制模型对罕见姿态和复杂重叠场景的泛化能力．另一方面，在网络特征提取上，小目标通常仅占据有限像素区域，其形状、边界和纹理信息较弱．在多物体重叠的X光背景下，局部卷积特征容易受到周围高响应背景区域和相邻物体结构的干扰，导致前景目标与背景之间的判别边界不够清晰．由于缺乏全局语义先验的辅助，模型在复杂背景中对小目标的稳定表征仍然不足，增加漏检和目标-背景混淆的风险．

针对复杂X光场景下定向小目标样本匮乏与目标-背景混淆问题，本文提出一种数据端与特征端协同增强的定向目标检测框架．在数据输入端，通过面向小目标的复制增强策略扩展长尾样本的空间分布和姿态多样性；在网络特征端，引入冻结的SAM（Segment Anything Model）语义辅助分支，为轻量检测网络提供高层语义先验，以增强复杂背景下的前景目标判别能力．

本文的主要贡献如下．

（1）提出一种上下文约束复制粘贴增强策略（Context-Constrained Copy-Paste Augmentation，CCPA），通过对定向小目标实例进行自动化的多维仿射变换（缩放、旋转、翻转等）与空间掩膜寻址，并结合泊松融合（Poisson Blending）技术消除边缘伪影，动态生成高复杂度的训练样本，从数据端有效解决了极端姿态匮乏与分布长尾问题．

（2）提出一种SAM增强型YOLO-OBB定向目标检测框架，以YOLO-OBB为检测主干，引入冻结的SAM图像编码器作为语义辅助分支，并通过通道注意力融合模块在中间特征层实现异构特征的自适应融合．该设计能够充分利用视觉基础模型提供的全局上下文与结构先验，增强检测网络在复杂X光场景下对小目标和重叠目标的特征建模能力．

（3）在复杂安检基准数据集上进行了充分的实验验证．大量消融实验表明，本文提出的双重增强机制能够产生显著的协同增益，有效提升了X光图像定向目标检测的精度．

1 相关工作

1.1 定向目标检测 传统目标检测算法大多基于水平边界框(Horizontal Bounding Box,HBB)进行目标定位,然而,在遥感、工业缺陷检测以及X光安检等场景中,目标通常呈现任意方向密集排布,HBB包含过多的背景噪声,且难以区分相邻重叠物体.为此,定向目标检测(Oriented Object Detection)引入了旋转边界框(Oriented Bounding Box,OBB),通过增加角度参数(如θ)来实现对物体的紧凑包围.

早期的OBB检测器多由基于锚框(Anchor-Based)的HBB算法演变而来,但其庞大的超参数搜索空间限制了模型的灵活性.随着无锚框(Anchor-Free)架构的兴起,基于YOLO系列[4−5]的旋转检测器因其卓越的推理效率成为主流.在边界框回归方面,由于角度的周期性与边界的不连续性,传统的平滑$L1$损失难以实现高精度优化.为此,研究者们提出基于交并比(Intersection over Union,IoU)的旋转损失以及分布焦域损失(Distribution Focal Loss,DFL)[7].DFL将连续的边界框坐标转化为离散的概率分布,有效提升网络对模糊边界的感知.本文针对X光图像中小目标边界模糊的特点,引入带高斯平滑的DFL机制,进一步增强了任意方向小目标回归的稳定性.

1.2 视觉基础模型与知识注入 近年来,以SAM[8−9]为代表的视觉基础模型(Vision Foundation Models,VFMs)在计算机视觉领域展现了惊人的零样本(Zero-Shot)泛化能力.SAM依托海量的图像与掩膜对进行预训练,其提取的高维特征蕴含了极其丰富的全局拓扑、轮廓以及目标独立性等高级语义先验.

现有研究多致力于基础模型的下游适配与表征迁移,主要范式包括知识蒸馏(Knowledge Distillation)与离线伪标签生成(Pseudo-Labeling).然而,这些间接的知识传递机制在处理复杂X光图像时面临显著的理论局限.常规的知识蒸馏在特征压缩的过程中,易导致基础模型中细粒度的拓扑结构发生不可逆的信息衰减,遭遇特征表达的泛化上限[10];离线伪标签生成不仅流程繁琐,且高度依赖过滤阈值的设定,在重叠杂乱的安检背景下极易引入并在训练中放大噪声误差[11−12].

针对上述间接迁移范式的局限性,本文提出一种更直接且高效的特征级先验注入策略.区别于破坏性微调或有损蒸馏,通过冻结SAM图像编码器以完整保留其原始的泛化流形,并设计通道注意力融合(Channel Attention Fusion,CAF)模块,在特征空间对基础模型的高阶语义先验与轻量级网络的多尺度特征进行动态调制.该机制在不显著增加网络推理复杂度的前提下,实现了复杂背景下微小目标特征的有效解耦.

1.3 数据增强 数据增强是提升深度学习模型泛化能力、缓解过拟合的核心技术.为了丰富训练样本的多样性与特征空间的宽泛度,现有目标检测研究主要依赖几类基础的增强机制.在空间与几何变换方面,随机翻转与面内旋转(Flip+Rotation)[13]通过物理视角的改变模拟了目标姿态与观测角度的随机变化.在图像级混合(Image-Level Mixing)层面,MixUp[14]通过对两幅图像的像素及标签进行全局线性插值,促使模型在不同类别间学习平滑的过渡特征;CutMix[15]通过局部图块的替换与标签按比例混合,引导网络关注区域级注意力;Mosaic[16]进一步将四张训练图像拼接为一幅,显著丰富了单次迭代内的上下文语义与多尺度特征.

在通道与像素级扰动方面,RandAugment[17]能在极简的超参数空间内自动采样并应用多种图像变换序列(如对比度、锐度调整),而AugMix[18]通过多条数据增强链的凸组合(Convex Combinations)来保持语义的一致性,有效提升了模型的抗干扰鲁棒性与不确定性校准能力.此外,在实例级增强(Instance-Level Augmentation)中,基础的Copy-Paste[6]通过裁剪源图像中的目标实例并将其随机粘贴至其他场景中,极大扩充了实例的组合方式并改善了长尾类别的平衡问题.

然而,在复杂重叠的X光安检场景中,传统增强机制常因忽略空间与视觉连贯性而产生边缘伪影,使网络陷入过拟合高频噪声的“捷径学习(Shortcut Learning)”.针对此局限,本文提出CCPA,通过多维仿射变换极大丰富了微小目标的定向姿态,并结合空间掩膜寻址与泊松融合(Poisson Blending)彻底消除视觉拼接伪影.该策略在

物理逻辑与视觉连贯性的双重约束下，有效化解了极端姿态样本的长尾分布难题.

2　本文方法

针对X光安检图像中背景复杂、目标遮挡重叠严重以及目标姿态变化随机等问题，提出一种融合视觉基础模型语义先验的增强型YOLO-OBB定向目标检测框架. 该方法以YOLO主干网络作为检测主体，并引入冻结的SAM图像编码器提取高层语义特征，通过通道注意力融合模块将语义先验注入检测分支，从而提升复杂场景下尤其是小目标与重叠目标的定向检测能力. 整体框架由任务主干分支、语义辅助分支以及通道注意力融合模块三部分组成，如图1所示.

2.1　整体架构设计　采用双分支异构架构，输入为安检场景下的X光扫描图像I. 推理过程中，输入图像同时送入任务主干分支与语义辅助分支进行特征提取，随后在主干网络的中间高分辨率特征层处完成异构特征融合，最终输出定向边界框检测结果. 具体结构如下.

(1)任务主干分支(Backbone Branch)基于YOLOv11/v12[4−5]架构进行改造，并结合OBB检测头实现定向目标回归. 该分支主要负责提取与检测任务直接相关的局部纹理、边缘结构和几何信息，通过多尺度特征建模，主干网络能够有效表征不同尺度目标，尤其适用于安检图像中尺度差异明显、目标分布密集的检测场景.

(2)语义辅助分支(Auxiliary Branch)采用SAM的图像编码器作为高层语义特征提取器. 首先，将输入图像进行尺寸调整与标准化预处理；随后送入冻结参数的SAM图像编码器，提取包含全局上下文信息与目标结构先验的高层语义特征，记为：

$$f_{\mathrm{SAM}} \in R^{B\times C_2\times H_2\times W_2} \tag{1}$$

由于该分支在训练阶段保持参数冻结，因此能够在不破坏基础模型泛化能力的前提下，为检测网络提供稳定的语义补充信息.

(3)特征融合流(Feature Fusion Flow). 输入图像I经主干分支后得到中间检测特征f_1，经语义辅助分支后得到高层语义特征f_{SAM}. 考虑到两路特征在空间分辨率上存在差异，首先对f_{SAM}进行插值对齐，使其与主干中间层特征具有一致的空间尺寸，得到对齐后的语义特征$f_2 \in R^{B\times C_2\times H\times W}$.

随后，将f_2通过本文设计的CAF模块注入主干分支，在通道维上完成语义增强与特征重标定，生成融合特征F'. 融合后的特征继续送入YOLO后续网络和OBB检测头，以完成最终的定向目标预测. 该过程本质上属于中间层特征增强，即利用SAM语义先验提升YOLO检测特征的判别性，而非在检测结果层面进行后融合.

2.2　通道注意力融合(CAF)模块　为充分利用SAM提供的高层语义先验，并抑制X光成像中复杂背景与穿透性噪声对检测特征的干扰，本文设计了通道注意力融合模块. 该模块以主干检测特征和SAM语义特征为输入，通过通道级自适应重加权机制实现异构信息筛选与融合，从而突出对目标检测更有贡献的响应通道.

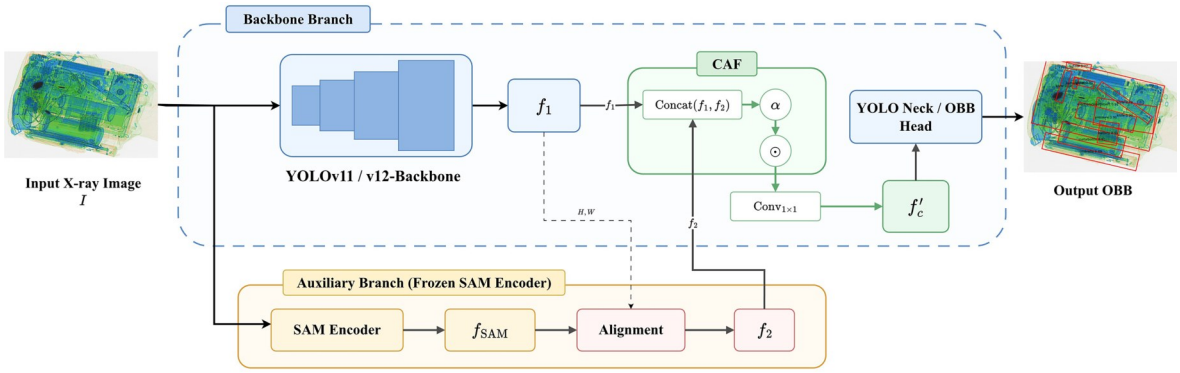


图1　SAM增强型YOLO-OBB框架的整体结构

Fig. 1　The overall structure of the SAM-enhanced YOLO-OBB framework

给定主干分支特征$\left(f_1\in R^{B\times C_1\times H\times W}\right)$和对齐后的SAM特征$\left(f_2\in R^{B\times C_2\times H\times W}\right)$,CAF模块通过特征拼接、通道权重学习和特征重标定来完成融合.

具体过程如下.

(1)特征聚合. 首先,在通道维度上对两路特征进行级联,构造联合特征表示:

$$f_c=\mathrm{Concat}\left(f_1,f_2\right)\in R^{B\times(C_1+C_2)\times H\times W} \tag{2}$$

该操作将来自主干网络的局部检测表征与来自SAM的全局语义先验映射到统一特征空间,为后续的通道筛选提供基础.

(2)通道注意力生成. 为了估计联合特征中各通道的重要性,首先对F施加全局平均池化,提取全局上下文描述符:

$$z=\mathrm{AvgPool}\left(f_c\right)\in R^{B\times(C_1+C_2)} \tag{3}$$

随后,将z输入带有瓶颈结构的两层MLP,生成通道注意力权重:

$$\alpha=\sigma\left(W_2\left(\delta\left(W_1(z)\right)\right)\right) \tag{4}$$

其中,$\delta(\cdot)$表示ReLU激活函数,$\sigma(\cdot)$表示Sigmoid激活函数,W_1和W_2分别表示降维与升维映射. 为了兼顾表达能力与参数开销,本文设置压缩比为4,最终得到的α表示联合特征在不同通道上的自适应响应强度.

(3)动态调制与通道适配. 利用通道权重对联合特征进行逐通道重标定:

$$\tilde{f}_c=f_c\odot\alpha \tag{5}$$

其中,$\odot$表示逐元素乘法. 该操作能够抑制无关背景通道,强化对目标轮廓、局部结构和语义区域更敏感的特征响应. 随后,采用1×1卷积对重标定后的特征进行通道压缩与适配,并结合批归一化和SiLU激活生成最终融合特征:

$$f_c'=\mathrm{SiLU}\left(BN\left(\mathrm{Conv}_{1\times1}\left(\tilde{f}_c\right)\right)\right) \tag{6}$$

(4)模块作用分析. CAF模块并非简单地将两路特征直接叠加,而是通过“级联建模-通道选择-线性适配”的方式,对主干分支与语义辅助分支的信息进行联合重构. 对于X光安检图像,该机制一方面能够引入SAM提供的全局语义结构先验,如目标闭合轮廓与实例拓扑信息;另一方面能够抑制穿透成像带来的背景冗余响应,为后续OBB检测头提供更具判别性的特征表示,提升复杂场景下的定向目标检测性能.

2.3 上下文约束复制粘贴增强策略 针对X光安检图像中小目标样本数量有限、姿态变化不足及空间分布不均衡等问题,设计了CCPA. 该策略通过对小目标实例进行受约束的复制、变换与重布局,在保持空间合理性的前提下扩展训练样本中小目标的尺度分布、位置分布与方向分布,增强检测模型对复杂安检场景中小目标的建模能力.

具体地,CCPA首先依据目标面积阈值与类别先验从原始标注中筛选小目标实例,并构建全局占用掩膜以记录图像中已有目标区域. 随后,对选中的实例施加随机缩放、旋转与翻转等仿射变换,以扩展其姿态分布. 在此基础上,算法在占用掩膜约束下搜索合法粘贴位置,并通过碰撞检测避免新生成实例与已有目标发生不合理重叠. 最后,采用泊松融合将增强实例无缝嵌入背景图像,并同步更新多边形标注,生成新的训练样本.

与传统复制粘贴增强方法相比,CCPA不仅关注实例数量扩充,还通过空间约束与无缝融合机制保证增强样本的布局合理性与视觉一致性. 该策略能够更有效地模拟安检图像中小目标密集分布、相互遮挡和姿态随机变化的实际情况,为后续定向目标检测提供更具多样性和代表性的训练数据. 增强效果示例如图2所示.

CCPA的整体流程如以下算法所示. 为便于表述,算法中若干操作符定义如下.

SelectSmallObjects(·)表示依据目标面积阈值与类别先验筛选小目标实例.

RandomSample(·)表示从候选小目标集合中随机采样待增强实例.

ApplyAffineTransform(·)表示对实例施加随机缩放、旋转与镜像翻转等仿射变换,并返回变换后的实例区域及其对应变换矩阵.

FindValidPosition(·)表示在当前占用掩膜约束下搜索合法粘贴位置,并通过碰撞检测判断候选位置是否可用.

PoissonBlend(·)表示采用泊松融合将增强实例无缝嵌入目标图像.



图2　本文提出的CCPA策略生成的训练样本示例图（上列是原始图像，下列是增强后的图像）

Fig. 2　Training sample examples generated by the proposed CCPA strategy (upper: original images, lower: enhanced images)

Transform(•)表示对实例的像素空间标注进行同步变换.

Shift(•)表示将变换后的像素空间标注平移至目标粘贴位置.

算法　CCPA

Input: 图像 $I\in R^{H\times W\times 3}$, 多边形标注集 $\mathcal{Y}$, 复制次数区间 $[C_{\min}, C_{\max}]$, 采样数 N, 小目标面积阈值 τ_a

Output: 增强图像 I', 更新后的标注集 $\mathcal{Y}'$

$I'\leftarrow I, \mathcal{Y}'\leftarrow\mathcal{Y}, M\leftarrow 0^{H\times W}$
For each 实例 $y\in\mathcal{Y}$ do
　$P_y\leftarrow$ 将实例 y 转换到像素空间
　$M(P_y)\leftarrow 1$
End For
$\mathcal{S}\leftarrow$RandomSample(SelectSmallObjects($\mathcal{Y}, \tau_a$), N)
For each 实例 $y\in\mathcal{S}$ do
　提取实例区域 R_y 及其轮廓 P_y
　$C_y\leftarrow$ 随机整数采样于 $[C_{\min}, C_{\max}]$
　For $i=1$ to C_y do
　　$R_i, T_i\leftarrow$ApplyAffineTransform(R_y)
　　$(x_i, y_i)\leftarrow$FindValidPosition(R_i, M)
　　If 找到合法位置 (x_i, y_i) then
　　$I'\leftarrow$PoissonBlend(I', R_i, x_i, y_i)
　　$P_i'\leftarrow$Shift(Transform(P_y, T_i), x_i, y_i)
　　$\mathcal{Y}'\leftarrow\mathcal{Y}'\cup\{P_i'\}$
　　$M(P_i')\leftarrow 1$
　　End If
　End For
End For
Return $I', \mathcal{Y}'$

2.4　损失函数设计　为了同时提升类别判别能力与定向边界框回归精度，采用复合损失函数对检测网络进行训练. 总损失由旋转框回归损失 L_{IoU}、分类损失 L_{cls} 以及带高斯平滑的分布焦点损失 L_{dflsm} 三部分加权构成：

$$L_{\text{total}}=\lambda_{\text{cls}}L_{\text{cls}}+\lambda_{\text{IoU}}L_{\text{IoU}}+\lambda_{\text{dfl}}L_{\text{dflsm}} \tag{7}$$

其中，λ_{cls}，λ_{IoU} 和 λ_{dfl} 分别表示各损失项的权重系数. 记正样本数量为：

$$N_+=\sum\nolimits_i\hat{y}_i \tag{8}$$

其中，$\hat{y}_i\in\{0,1\}$ 表示第 i 个候选位置是否为正样本.

（1）旋转框回归损失 L_{IoU}. 由于安检图像中的目标具有显著的方向变化特性，传统水平框IoU难以准确刻画旋转目标之间的几何差异. 为此，采用面向定向边界框的IoU损失来监督旋转框回归，其定义为：

$$L_{\text{IoU}}=\frac{1}{N_+}\sum\nolimits_i\left(1-\text{IoU}\left(b_i,\hat{b}_i\right)\right)\hat{y}_i \tag{9}$$

其中，b_i 和 $\hat{b}_i$ 分别表示预测旋转框与真实旋转框. 该损失能够在位置、尺度与角度三个层面提供有效的几何约束，从而提高旋转框回归的稳定性与精确性[19].

（2）分类损失 L_{cls}. 分类分支采用二元交叉熵损失进行监督，定义为：

$$L_{\text{cls}}=\frac{1}{N_+}\sum_i BCE\left(p_i,\hat{p}_i\right) \tag{10}$$

其中,p_i表示预测类别概率,$\hat{p}_i$表示真实类别标签. 通过正样本归一化处理,该损失能够在复杂背景与类别不均衡条件下保持训练稳定性[20].

(3)带高斯平滑的分布焦点损失L_{dflsm}. 为了增强离散距离分布建模的平滑性与鲁棒性,在分布焦点损失中引入高斯软标签机制. 该设计与分布式框回归思想一致,并可视为对DFL建模方式的平滑扩展[7]. 对于真实目标值r^*,构造对应的目标分布:

$$p_j^{\text{target}}=\frac{1}{Z}\exp\left(-\frac{\left(x_j-r^*\right)^2}{2\sigma^2}\right),\quad \sigma=\gamma R \tag{11}$$

其中,x_j表示离散采样位置,Z为归一化因子,R为回归范围,γ为控制分布宽度的比例系数. 设p^{pred}表示预测分布,则对应的DFL项为:

$$L_{\text{dflsm}}=\frac{1}{N_+}\sum_i KL\left(p_i^{\text{target}}\,\|\,p_i^{\text{pred}}\right) \tag{12}$$

其中,$KL(\cdot\|\cdot)$表示Kullback-Leibler散度[21].

综合上述各项损失,模型能够在复杂安检场景下同时优化类别识别能力与定向框定位精度,从而更好地适应目标重叠严重、尺度变化显著以及姿态随机分布的检测任务.

3 实验与分析

3.1 实验设置

3.1.1 数据集与评估指标 在自建安检X光定向目标检测数据集上开展实验. 该数据集面向安检场景中的危险品与可疑物品检测任务,具有目标尺寸小、目标间重叠严重、背景干扰复杂以及目标姿态随机分布等特点,对检测模型的特征表征能力与旋转框定位能力提出了较高要求. 数据集共包含3500张图像,其中训练集3000张,测试集500张. 按照类别划分,数据集包含10类目标,分别为knife, lighter, electronicequipment, battery, OCbottle, metalbottle, pressure, glassbottle, umbrella和gun.

由于本文的研究任务是定向目标检测,因此采用基于旋转框匹配的平均精度指标对模型性能进行评估. 设类别集合为:

$$C=\{1,2,\cdots,c\} \tag{13}$$

其中,$C=10$. 对于类别$c\in C$,在给定IoU的阈值τ下,其平均精度定义为该类别精确率-召回率曲线下面积,即:

$$AP_c^{\tau}=\int_0^1 p_c^{\tau}(r)\,\mathrm{d}r \tag{14}$$

其中,r表示召回率,$p_c^{\tau}(r)$表示在IoU的阈值τ下类别c的插值精确率函数. 对于OBB任务,预测框与真实框之间的匹配基于旋转框IoU计算.

在此基础上,采用mAP_{50}作为主要评估指标. mAP_{50}定义为在IoU阈值$\tau=0.50$时所有类别平均精度的均值:

$$mAP_{50}=\frac{1}{C}\sum_{c=1}^{C}AP_c^{0.50} \tag{15}$$

3.1.2 实施细节 选取YOLOv11-OBB和YOLOv12-OBB作为基准检测框架,以验证所提方法在定向目标检测任务中的有效性. 所有实验均在单张NVIDIA GeForce RTX 5090 GPU上完成训练与测试. 输入图像分辨率统一设置为640×640,训练轮数为120,batch size为8. 优化器采用带动量的随机梯度下降(Stochastic Gradient Descent,SGD)优化策略,初始学习率设为0.01,动量系数设为0.937,权重衰减系数设为0.0005.

在模型结构方面,语义辅助分支采用SAM-B作为基础模型,其图像编码器可提取高层语义先验特征,并在训练过程中保持参数冻结. 主干检测分支基于YOLO-OBB架构进行训练,将融合后的特征进一步送入YOLO Neck与OBB Head,以完成定向目标预测. 除特别说明外,YOLOv11-OBB与YOLOv12-OBB在训练过程中采用相同的数据划分、输入设置与优化策略,以保证实验比较的公平性.

损失函数的权重系数设置为$\lambda_{\text{cls}}=0.5$,$\lambda_{\text{IoU}}=7.5$,$\lambda_{\text{dfl}}=1.5$,其中,λ_{cls}用于平衡分类损失项,λ_{IoU}用于控制旋转框回归损失项的贡献,λ_{dfl}用于调节分布焦点损失项的权重. 其余训练配置在不同模型之间保持一致.

3.2 消融实验 为了验证各关键设计的有效性,从数据增强、领域预训练、损失函数改进、语义辅助分支和检测主干配置等方面开展消融实验,重点分析CCPA对小目标检测的影响、PIDray预训练与改进损失函数的作用、SAM语义辅助分支及

CAF融合机制的有效性以及不同主干和SAM编码器配置下的性能变化．相关结果如表1、表2、图3和图4所示．

3.2.1　训练策略与关键模块贡献分析　表1展示了不同训练策略与模块组合下的消融结果．以YOLOv11-OBB为基线模型，在仅使用Mosaic与RandAugment时，模型取得了76.01的基线性能．随后，随着区域级数据增强、领域预训练、损失函数改进以及语义辅助分支的逐步引入，检测性能持续提升．

首先，引入CopyPaste，CutMix与MixUp后，模型性能由76.01提升至82.19，说明区域级与样本级增强能够有效改善安检图像中小目标稀疏、遮挡严重和样本分布不均的问题．进一步加入角度扰动后，性能提升至82.82，表明旋转增强对定向目标检测任务具有积极作用．

其次，引入PIDray数据集[2]预训练初始化后（PIDray：pre），模型性能由82.76提升至83.89；进一步采用高斯平滑分布焦点损失后（DFL：dflsm），性能提升至84.68；进一步利用PIDray与当前类别一样的数据训练（PIDray：pre＋use）时，性能继续提升至85.43．说明领域先验与平滑分布监督均有助于提升特征表征能力与旋转框回归稳定性．

在结构层面，引入冻结的SAM-B语义辅助分支后，模型性能由85.43提升至86.33，说明视觉基础模型提供的高层语义先验能够有效增强检测分支表示．随后，将自动增强策略替换为AugMix后，性能进一步提升至86.62．进一步引入本文CCPA后，性能达到87.10，说明该策略能够有效增强小目标样本的空间与姿态多样性．总体上，经过数据增强、领域预训练、改进损失函数、SAM语义辅助分支以及CCPA的逐步叠加后，模型性能较基线提升了11.09，验证了本文各项设计的有效性与互补性．

3.2.2　主干网络与辅助分支配置分析　表2展示了不同检测主干与SAM辅助分支配置下的对比结果，该组实验在统一训练策略下，仅对检测主干与辅助分支配置进行调整．结果表明，在YOLOv11-OBB主干上，将辅助分支由SAM-B更换为SAM-L后，性能由87.10小幅提升至87.21，说明更强的SAM编码器能带来一定增益，但提升幅度有限．将检测主干替换为YOLOv12-OBB后，模型取得了87.02的结果，没有

表1　不同数据增强策略、预训练方式及结构模块组合下的消融实验结果

Table 1　Ablation results for combinations of various data augmentation strategies, pre-training approaches, and structural modules

Network	Mosaic	AutoAug	CopyPaste	MixUp	CutMix	Flip＋Deg	CCPA	PIDray	DFL	mAP_{50}
YOLOv11-OBB	1	RandAugment	0	0	0	0＋0	0	0	dfl	76.01
	1	RandAugment	1	0	0	0＋0	0	0	dfl	77.56
	1	RandAugment	0	0	0	0＋0	1	0	dfl	80.22
	1	RandAugment	1	0	1	0＋0	0	0	dfl	79.21
	1	RandAugment	1	0.5	1	0＋0	0	0	dfl	80.35
	1	RandAugment	1	1	1	0＋0	0	0	dfl	82.19
	1	RandAugment	1	1	1	0＋30	0	0	dfl	82.82
	1	RandAugment	1	1	1	0.5＋30	0	0	dfl	82.76
	1	RandAugment	1	1	1	0.5＋30	0	pre	dfl	83.89
	1	RandAugment	1	1	1	0.5＋30	0	pre	dflsm	84.68
	1	RandAugment	1	1	1	0.5＋30	0	pre＋use	dflsm	85.43
YOLOv11-OBB＋SAM-B	1	RandAugment	1	1	1	0.5＋30	0	pre＋use	dflsm	86.33
	1	AugMix	1	1	1	0.5＋30	0	pre＋use	dflsm	86.62
	1	AugMix	1	1	1	0.5＋30	1	pre＋use	dflsm	87.10

表 2 不同检测主干与 SAM 辅助分支配置下的消融实验结果

Table 2 Ablation results for various detection backbone and SAM auxiliary branch configurations

Network	Mosaic	AutoAug	CopyPaste	MixUp	CutMix	Flip+Deg	CCPA	PIDray	DFL	mAP_{50}
YOLOv11-OBB+SAM-B	1	AugMix	1	1	1	0.5+30	1	pre+use	dflsm	87.10
YOLOv11-OBB+SAM-L	1	AugMix	1	1	1	0.5+30	1	pre+use	dflsm	87.21
YOLOv12-OBB	1	AugMix	1	1	1	0.5+30	1	pre+use	dflsm	87.02
YOLOv12-OBB+SAM-B	1	AugMix	1	1	1	0.5+30	1	pre+use	dflsm	87.58

优于前两种配置. 进一步在 YOLOv12-OBB 基础上引入 SAM-B 辅助分支后,模型性能提升至 87.58,并取得该组实验中的最佳结果.

上述结果表明,SAM 语义辅助分支在不同检测主干上均能带来较稳定的性能增益,说明视觉基础模型提供的高层语义先验具有一定的主干适配性. 同时,更强检测主干与 SAM 语义辅助分支的联合使用能进一步提升模型的特征表征能力和目标判别性能. 可见本文构建的以检测主干为主体,冻结 SAM 语义分支为辅助,并通过 CAF 模块进行中间层语义融合的检测框架,能有效适应复杂 X 光安检场景下的定向目标检测任务.

3.2.3 CCPA 参数敏感性分析 图 3 展示了 CCPA 中小目标面积阈值与复制次数范围的消融实验结果. 当不使用小目标复制增强时,模型的 mAP_{50} 为 85.43. 随着小目标面积阈值和复制次数范围的合理增大,模型检测性能整体呈现先提升后下降的趋势. 其中,当复制次数范围设置为 $[5,15]$ 时,模型在不同面积阈值下整体表现最优;当小目标面积阈值设置为 0.025 时,模型的 mAP_{50} 最高,为 86.61,比不使用 CCPA 时提升 1.18. 相比之下,复制次数在 $[0,5]$ 时增强强度不足,性能提升较有限;复制次数在 $[10,20]$ 时,在阈值为 0.020 时结果较好,但继续增大阈值后性能出现下降,说明过多复制可能引入冗余样本并扰动原始数据分布. 上述结果表明,适度的小目标复制增强能有效改善小目标样本不足问题,但增强强度需要控制在合理范围内.

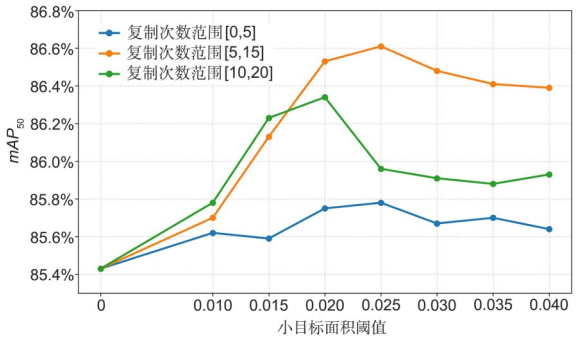


图 3 CCPA 中小目标面积阈值与复制次数范围的消融实验结果

Fig. 3 Ablation results for small object area threshold and copy count range in CCPA

3.2.4 SAM-B 对目标-背景混淆的影响分析 图 4 给出了引入 SAM-B 前后的归一化混淆矩阵对比. 由图可见,引入 SAM-B 后,部分类别的正确识别率有所提升,例如,knife 由 0.88 提升至 0.90,battery 由 0.84 提升至 0.85,OCbottle 由 0.88 提升至 0.89,metalbottle 由 0.94 提升至 0.95. 同时,部分目标被误判为 background 的比例有所下降,例如,knife 由 0.11 降低至 0.09,battery 由 0.16 降低至 0.15,metalbottle 由 0.03 降低至 0.02,pressure 由 0.06 降低至 0.04. 说明 SAM-B 提供的语义先验能够在一定程度上增强模型对前景目标的判别能力,降低部分类别被背景吸收或漏检的风险. 需要指出,SAM-B 对不同类别的影响存在差异,如 glassbottle 和 umbrella 的识别率略有下降,但整体上其对目标-背景混淆的缓解作用较明显,从检测结果层面支持了本文引入语义辅助分支的必要性.

此外,相关方案在 IJCAI 2025 的 The 3rd International Workshop on Generalizing from Limited Resources in the Open World (GLOW) 相关算法挑战赛中获得 Challenge Excellence Award-2nd Place(亚军),进一步说明本文方法在实际任务中的有效性.

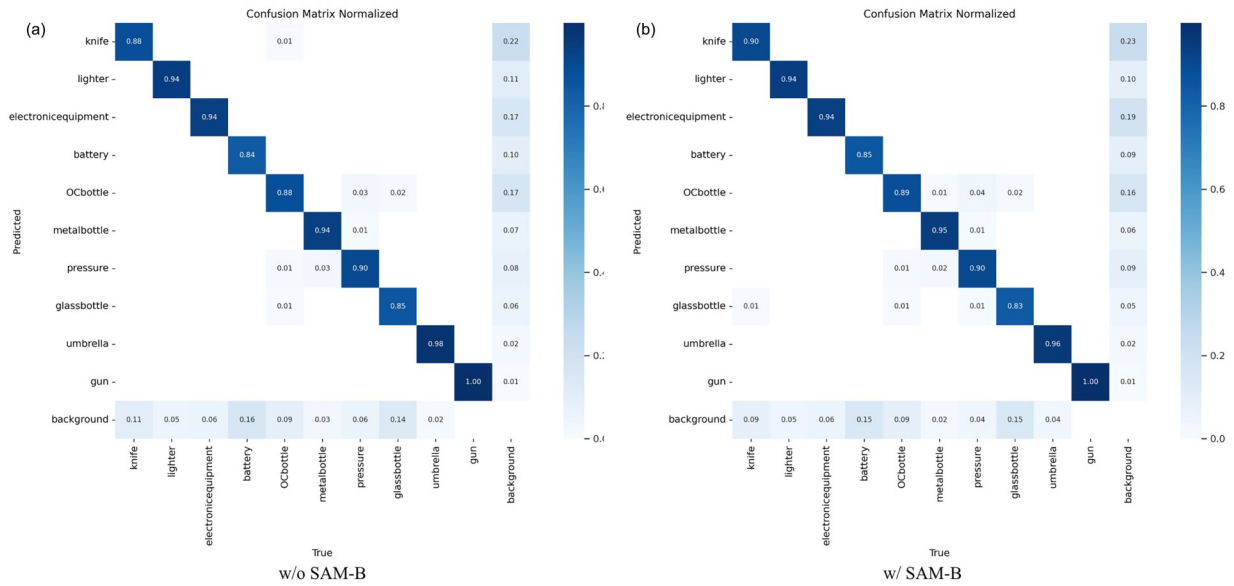


图4　引入SAM-B前后的归一化混淆矩阵对比：(a) w/o SAM-B；(b) w/ SAM-B

Fig. 4　Normalized confusion matrix comparison with and without SAM-B: (a) w/o SAM-B, (b) w/ SAM-B

4　结论

本文针对复杂安检X光场景下小目标检测困难、目标重叠严重以及语义表达不足等问题，提出一种基于语义与样本双重增强机制的定向目标检测框架. 在语义增强方面，利用冻结的SAM图像编码器引入高层语义先验，并通过通道注意力融合模块与YOLO-OBB中间层特征进行协同建模，提升模型对复杂结构与重叠目标的判别能力. 在样本增强方面，设计上下文约束复制粘贴增强策略，通过受约束的小目标复制、几何变换与无缝融合，缓解小目标样本匮乏和姿态分布不足的问题. 同时，结合高斯平滑分布焦点损失，进一步提升旋转框回归的稳定性. 实验与消融结果表明，语义增强与样本增强具有良好的互补性，能够有效提升复杂X光场景下的定向目标检测性能.

总体上，本文验证了将视觉基础模型语义先验与面向小目标的样本增强机制协同引入定向检测框架的有效性，为复杂安检场景下高精度目标检测提供了一种可行思路.

未来将围绕更轻量的语义增强结构、更精细的小目标增强策略以及更强的跨场景泛化能力开展研究.

参考文献

[1] Miao C J, Xie L X, Wan F, et al. SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 2114－2123.

[2] Zhang L B, Jiang L T, Ji R Y, et al. PIDray: A large-scale X-ray benchmark for real-world prohibited item detection. International Journal of Computer Vision, 2023, 131(12): 3170－3192.

[3] Tao R S, Wang H Y, Guo Y Z, et al. Dual-view X-ray detection: Can AI detect prohibited items from dual-view X-ray Images like humans?//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2025: 10338－10347.

[4] Khanam R, Hussain M. YOLOv11: An overview of the key architectural enhancements. https://arxiv.org/abs/2410.17725, 2024－10－23.

[5] Tian Y J, Ye Q X, Doermann D. YOLO12: Attention-centric real-time object detectors. https://arxiv.org/abs/2502.12524, 2025－02－18.

[6] Ghiasi G, Cui Y, Srinivas A, et al. Simple copy-paste is a strong data augmentation method for instance

segmentation//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2021: 2917－2927.

[7] Li X, Wang W H, Wu L J, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection//Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc, 2020: 21002－21012.

[8] Kirillov A, Mintun E, Ravi N, et al. Segment anything//2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023: 3992－4003.

[9] Ravi N, Gabeur V, Hu Y T, et al. SAM 2: Segment anything in images and videos. https://arxiv. org/abs/2408.00714, 2024－10－28.

[10] Lv C H, Zhao D, Wang S, et al. Generalizable knowledge distillation from vision foundation models for semantic segmentation. https://arxiv. org/abs/2603.02554, 2026－03－03.

[11] He C M, Zhang R H, Tang L X, et al. SCALER: SAM - enhanced collaborative learning for label - deficient concealed object segmentation. https://arxiv.org/abs/2511.18136, 2025－11－22.

[12] Shu H, Li W S, Tang Y H, et al. TinySAM: Pushing the envelope for efficient segment anything model. Proceedings of the AAAI Conference on Artificial Intelligence, 2025: 39(19): 20470－20478.

[13] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning. Journal of Big Data, 2019, 6(1): 60.

[14] Zhang H Y, Cisse M, Dauphin Y N, et al. Mixup: Beyond empirical risk minimization//ICLR 2018 Conference Track. Vancouver, Canada: ICLR, 2018: 1－13.

[15] Yun S, Han D, Chun S, et al. CutMix: Regularization strategy to train strong classifiers with localizable features//2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 6022－6031.

[16] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection. https://arxiv.org/abs/2004.10934, 2020－04－23.

[17] Cubuk E D, Zoph B, Shlens J, et al. Randaugment: Practical automated data augmentation with a reduced search space//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, WA, USA: IEEE, 2020: 3008－3017.

[18] Hendrycks D, Mu N, Cubuk E D, et al. AugMix: A simple data processing method to improve robustness and uncertainty//International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020: 1－15.

[19] Murrugarra-Llerena J, Kirsten L N, Zeni L F, et al. Probabilistic intersection-over-union for training and evaluation of oriented object detectors. IEEE Transactions on Image Processing, 2024, 33: 671－681.

[20] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press, 2016.

[21] Kullback S, Leibler R A. On information and sufficiency. The Annals of Mathematical Statistics, 1951, 22(1): 79－86.

(责任编辑 杨可盛)