激光与光电子学进展

基于DETR和改进去噪训练的光学遥感图像多尺度旋转目标检测(特邀)

金睿蛟[1,2†], 王堃[1,2†], 刘敏豪[1,2], 腾锡超[1,2], 李璋[1,2*], 于起峰[1,2]

[1]国防科技大学空天科学学院, 湖南 长沙 410000;

[2]图像测量与视觉导航湖南省重点实验室, 湖南 长沙 410000

摘要 旋转目标检测是遥感图像解译的重要任务之一,存在目标方向任意、小目标密集排列、目标表示引起的角度周期性等典型问题。针对上述问题,提出一种基于DEtection Transformer(DETR)目标检测器和改进去噪训练的旋转目标检测方法,即arbitrary-oriented object detection Transformer with improved deNoising anchor boxes(AO^{2}DINO)。首先,该方法引入一种多尺度旋转可变形注意力模块,将角度信息以旋转矩阵的形式引入注意力权重的计算,提高了模型对旋转目标的适应能力。其次,针对小目标密集排列问题,提出一种自适应的样本分配器,引入旋转交并比和自适应阈值,实现对密集目标更加精确的采样,提升了模型对小目标的检测能力。最后,在模型中引入基于卡尔曼滤波的交并比(KFIoU)作为回归损失,以解决旋转目标表示引起的角度周期性问题。AO^{2}DINO在DOTAv1.0和DIOR-R两个公开数据集上与典型的旋转目标检测方法进行了比较,在DETR系列旋转目标检测方法中检测精度最高,且训练时收敛速度更快,在训练12个epochs时就几乎达到了其他旋转目标检测方法训练36个epochs时的检测效果。

关键词 旋转目标检测; DETR目标检测器; 多尺度旋转可变形注意力

中图分类号 O436 **文献标志码** A **DOI**: 10.3788/LOP240502

DETR with Improved DeNoising Training for Multi-Scale Oriented Object Detection in Optical Remote Sensing Images (Invited)

Jin Ruijiao[1,2†], Wang Kun[1,2†], Liu Minhao[1,2], Teng Xichao[1,2], Li Zhang[1,2*], Yu Qifeng[1,2]

[1]*College of Aerospace Science and Engineering, National University of Defense Technology, Changsha* 410000, *Hunan, China*;

[2]*Hunan Key Laboratory of Image Measurement and Vision Navigation, Changsha* 410000, *Hunan, China*

Abstract Oriented object detection is one of the important tasks in remote sensing image interpretation, which faces typical problems such as arbitrary object orientation, dense arrangement of small targets, and angular periodicity caused by target representation, thus, this paper proposes a method called arbitrary-oriented object detection Transformer with improved deNoising anchor boxes (AO^{2}DINO) which based on DEtection Transformer (DETR) and improved denoising training. First, a multi-scale rotated deformable attention (MS-RDA) module is proposed. The MS-RDA module introduces the angle information in the form of rotation matrix for the calculation of attention weights, which improves the adaptability of the model to the orientated objects. Second, this paper proposes a self-adaption assigner (SAA), which uses the rotated intersection over union (IoU) and adaptive threshold to accurately separate dense targets, to improve the small targets detection under the dense arrangement scenarios. Finally, the Kalman filtering IoU (KFIoU) is introduced as the regression loss to solve the angular periodicity problem caused by the representation of orientated objects. Our proposed method is compared with the typical oriented bounding box (OBB) methods on two public datasets, DOTAv1.0 and DIOR-R, and the detection accuracy is the highest among the DETR-based OBB methods, and the convergence speed is faster during training, which only needs 12 training epochs to achieve comparable detection accuracy as other methods using 36 training epochs.

Key words oriented object detection; DETR object detector; multi-scale rotated deformable attention

收稿日期: 2023-11-12; **修回日期**: 2023-12-18; **录用日期**: 2023-12-18; **网络首发日期**: 2023-12-25

基金项目: 国家自然科学基金(61801491)

通信作者: *lizhang08@nudt.edu.cn

† 共同第一作者

1 引 言

基于深度学习的旋转目标检测方法在遥感领域已得到广泛应用[1]，旨在预测图像中目标的边界框和类别。尽管它取得了显著进展，但是经典的旋转目标检测方法[2-8]主要是基于卷积神经网络（CNN）的，直到Carion等[9]将Transformer结构引入目标检测，提出一种新型目标检测器，即DEtection Transformer（DETR）。不同于基于CNN的旋转目标检测器，DETR将目标检测视为集合预测的问题，使用可学习的查询（Query）来预测目标的存在，并通过二分匹配来分配标签，摆脱了模型对先验知识（例如锚框设计）以及非极大值抑制（NMS）等后处理操作的依赖，实现了真正意义上的端到端目标检测。

基于DETR的目标检测因自主性成为目前的研究热点，但仍存在训练收敛慢、查询的含义不清晰、小目标检测效果不佳等问题[10]。Deformable DETR[10]引入参考点和采样点，每个查询对应一个参考点，采用稀疏空间的采样方法，在计算注意力权重时，只关注参考点周围的小部分关键采样点，从而提升模型训练时的收敛速度。DAB-DETR[11]借鉴了基于锚框的目标检测器的思路，引入锚框作为位置先验，使查询有了可解释性。针对模型训练时收敛慢的问题，DN-DETR[12]提出了去噪训练（deNoising training）模块，对真值框添加细微的扰动作为噪声，在训练过程中直接重构真值框以跳过匈牙利匹配过程，提升了训练时的收敛速度。DN-DETR在训练速度和稳定性上提升效果明显，在进行去噪训练时，基于附近有真值框的查询进行正样本的预测，但缺乏对附近没有真值框的负样本的预测能力，预测时容易造成目标冗余。DINO[13]提出了对比去噪训练（contrastive deNoising training）方法以解决上述问题，该方法对真值框添加了两种不同程度的噪声扰动，构成了正负样本，从而剔除无用的样本，避免重复输出带来的目标冗余。DINO在COCO数据集上达到了SOTA的效果[13]，但其使用的水平边界框不适用于遥感图像的旋转目标检测。

O^{2}DETR[14]首次将DETR结构用于旋转目标检测，通过使用深度可分离卷积代替注意力机制，显著降低在模型中使用多尺度特征的计算成本。AO^{2}DETR[15]提出旋转候选框生成机制和自适应的候选框优化模块，提取旋转不变的区域特征并消除区域特征和目标之间的错位。ARS-DETR[16]提出一种基于高宽比感知的圆平滑标签作为角度的回归分支，并将角度更新引入注意力计算模块，使得模型能够更加精准地适应旋转目标。尽管上述模型在旋转目标检测任务中取得了较高的检测精度，但在目标方向任意、小目标密集排列、旋转目标表示引起的角度周期性等典型问题上仍有改进空间[9-10, 16]。

为了解决上述DETR方法针对旋转目标检测的问题，本文提出基于DETR和改进去噪训练的模型，即arbitrary-oriented object detection Transformer with improved deNoising anchor boxes（AO^{2}DINO）。将DINO[13]作为模型基础引入旋转目标检测方法，提出一种多尺度旋转可变形注意力（MS-RDA）模块，该模块将5D的查询（x, y, w, h, θ）中的角度信息以旋转矩阵的形式引入注意力计算，提升了模型对任意方向旋转目标的适应能力。同时，提出一个自适应的正负样本分配器（SAA），根据旋转框之间的交并比和自适应阈值，实现对正负样本更加精确的判断，提升对密集小目标的检测效果。最后，借鉴单阶段旋转目标检测损失函数的设计思路，以基于卡尔曼滤波的交并比（KFIoU）[17]替换Smooth L1损失，将旋转框参数的分布转换为高斯分布，并利用类似卡尔曼滤波的方式计算两个高斯分布之间的重叠面积，以解决旋转目标表示引起的角度周期性问题，进一步提高AO^{2}DINO框架对旋转目标的检测能力。

2 AO^{2}DINO方法

AO^{2}DINO方法以旋转DINO为基本框架，引入了多尺度旋转可变形注意力模块和自适应分配器，分别用于提升模型对旋转目标和密集小目标的检测能力。

2.1 检测网络整体架构

AO^{2}DINO模型由主干网络（Backbone）、编码器（Transformer Encoder）、解码器（Transformer Decoder）、预测头（Prediction Head）4部分组成。其中，主干网络用于多尺度特征提取；编码器用于对提取的特征进行多层处理，得到一组抽象的特征向量；解码器用于对编码器输出的特征与位置嵌入信息（position embedding）进行交互，实现目标位置和编码器特征的有效融合；最后，预测头用于输出最后的检测结果。

如图1所示，与通用的DETR系列检测器不同，AO^{2}DINO将解码器分为匹配部分和对比去噪部分，对比去噪部分仅在训练时进行，在推理时将会被移除。模型的训练过程为：图像输入主干网络后，将得到的特征铺平（flatten），加上位置嵌入信息后输入编码器，得到并行特征，将通过这些特征得到的可学习的查询作为解码器的输入；在解码器的匹配部分，对这些查询与真值框进行匈牙利匹配，并通过预测头的前馈神经网络（FFN）输出预测结果。匹配公式为

$$\hat{\sigma}=\underset{\sigma\in S_N}{\operatorname{argmin}}\sum_i^N \mathcal{L}_{\text{match}}\left(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_{\sigma(i)}\right), \tag{1}$$

式中：$\boldsymbol{y}_i$为真值框；$\hat{\boldsymbol{y}}_{\sigma(i)}$为待匹配的查询。同时，在解码器的对比去噪部分，除了输入可学习的查询外，还输入N个对比去噪组（CDN group）。每组中，对同一真值框添加两种不同大小均匀分布的噪声，经自适应的分配器筛选后被标记为正负样本，在训练时利用这些样本重构真值框，从而跳过匈牙利匹配的过程，加快模型训练时的收敛速度。AO^{2}DINO解码器的公式为

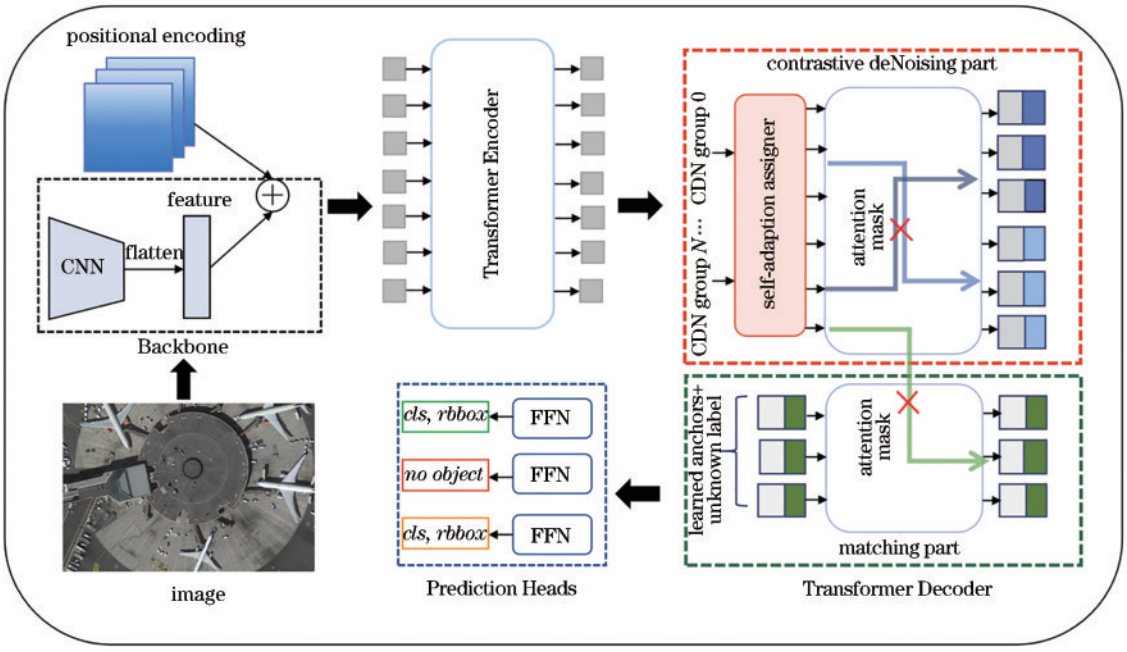

图 1　AO²DINO 整体架构

Fig. 1　Overall architecture of AO²DINO

$$\boldsymbol{o}=D(\boldsymbol{q},\boldsymbol{Q},\boldsymbol{F}|\boldsymbol{A}),\tag{2}$$

式中：$\boldsymbol{q}=\{\boldsymbol{q}_0,\boldsymbol{q}_1,\cdots,\boldsymbol{q}_{N-1}\}$表示解码器的对比去噪部分；$\boldsymbol{Q}=\{\boldsymbol{Q}_0,\boldsymbol{Q}_1,\cdots,\boldsymbol{Q}_{N-1}\}$表示解码器的匹配部分；$\boldsymbol{F}$和$\boldsymbol{A}$表示编码器输出的细化图像特征和注意力掩码(attention mask)；D表示解码器；$\boldsymbol{o}=\{\boldsymbol{o}_0,\boldsymbol{o}_1,\cdots,\boldsymbol{o}_{N-1}\}$表示解码器的输出。此外，由于噪声样本中包含大量真实信息，为防止信息泄漏，利用注意力掩码对不同去噪组间的样本以及去噪部分和匹配部分之间的样本进行隔离，保证训练的有效性。

2.2　多尺度旋转可变形注意力模块

为解决DETR[9]收敛慢的问题，deformable DETR[10]引入了可变形注意力模块，无须像Transformer[18-19]一样从全局开始学习逐渐过渡到局部。不考虑特征图的空间大小，该方法提出一个4D的参考点(x,y,w,h)，使每个查询只关注参考点周围的一小组关键采样点，即查询只与采样点进行局部稀疏的注意力权重计算，进而将得到的权重施加在对应的Value上，计算公式为

$$\mathrm{MultiHeadAttn}(\boldsymbol{z}_q,\boldsymbol{p}_q,\boldsymbol{x})=\sum_{m=1}^{M}W_m\left[\sum_{k=1}^{K}\boldsymbol{A}_{mqk}\cdot W_m'\boldsymbol{x}(\boldsymbol{p}_q+\Delta\boldsymbol{p}_{mqk})\right],\tag{3}$$

式中：$\boldsymbol{z}_q$表示查询；$\boldsymbol{p}_q$表示$\boldsymbol{z}_q$的坐标位置(即参考点)；W_m表示对注意力施加在Value后的结果进行线性变换；$\Delta\boldsymbol{p}_{mqk}$表示采样点相对于参考点的位置偏移；$\boldsymbol{A}_{mqk}$表示归一化的注意力权重；$W_m'\boldsymbol{x}(\boldsymbol{p}_q+\Delta\boldsymbol{p}_{mqk})$表示采样点位置插值处的Value。在此基础上，针对遥感图像，AO²DETR[15]将Transformer结构引入旋转目标检测，提出5D的参考点(x,y,w,h,θ)，但依旧使用上述多尺度旋转可变形注意力模块，因此当参考点输入模块前，会舍去角度维度，这种做法容易导致特征的错位。

为解决上述问题，本文提出多尺度的旋转可变形注意力模块。图2右图展示了该模块在解码器中的位置及输入，使用5D可学习的查询(x,y,w,h,θ)作为参考点，指示器用于判断模型正在进行匹配任务还是去噪任务。图2左图表示多尺度的旋转可变形注意力模块的具体计算过程，输入5D参考点后，将最后一维角度单独提取出来，通过变换形成旋转矩阵$\boldsymbol{\lambda}_{\mathrm{Ra}}$作用在采样点上，再进行注意力权重的计算，计算公式为

$$\boldsymbol{\lambda}_{\mathrm{Ra}}=\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix},\tag{4}$$

$$\mathrm{RotatedMSDeformAttn}(\boldsymbol{z}_q,\boldsymbol{p}_q,\boldsymbol{x})=\sum_{m=1}^{M}W_m\left[\sum_{k=1}^{K}\boldsymbol{A}_{mqk}\cdot W_m'\boldsymbol{x}(\boldsymbol{p}_{\mathrm{Rp}}+\boldsymbol{\lambda}_{\mathrm{Ra}}\Delta\boldsymbol{p}_{mqk})\right]。\tag{5}$$

此外，由于引入注意力机制，与基于CNN的经典旋转目标方法相比(以ReDet为例)，AO²DINO能够自动地学习并选择性地关注图像中的目标信息，对图像中的关键区域响应更强，赋予其更高的权重，目标更容易被检测和识别。两种方法关于Grad-CAM热力图可视化的对比如图3所示。

2.3　自适应分配器

在训练过程中，AO²DINO利用去噪任务重建真值框，从而跳过匈牙利匹配，加快模型收敛速度。如图4(a)所示，通常在加入均匀分布的噪声后，解码器直接根据样本角点的偏移量来判断样本正负[13]。然而，考虑到旋转目标检测中目标角度的任意性、背景的复杂性以及图像中目标占比小且密集排列等问题，本

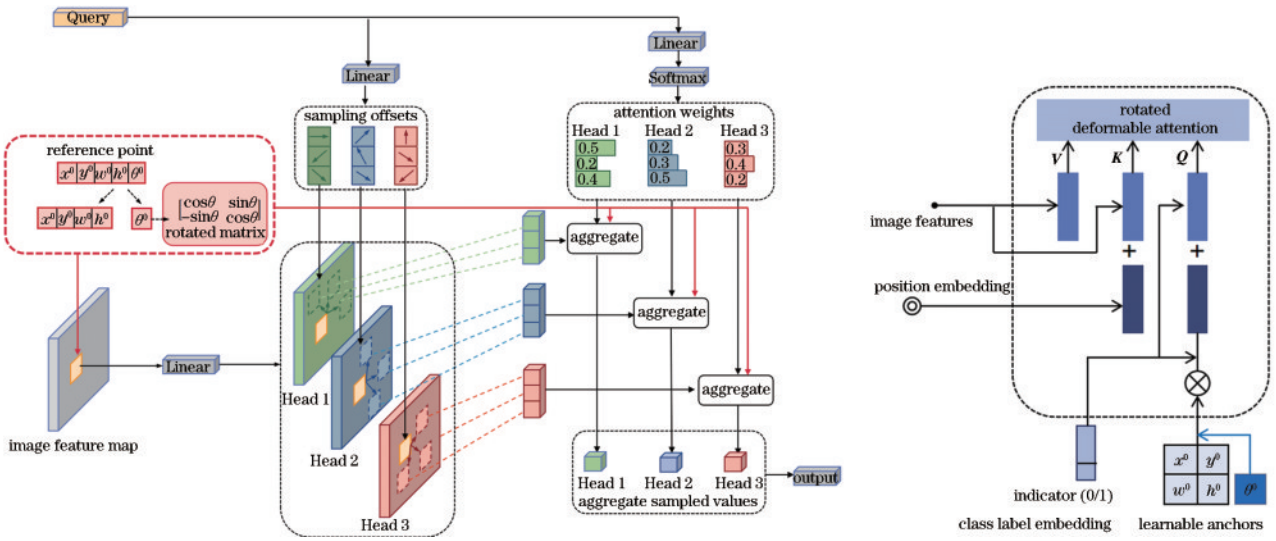

图 2　多尺度旋转可变形注意力模块

Fig. 2　Multi-scale rotated deformable attention module

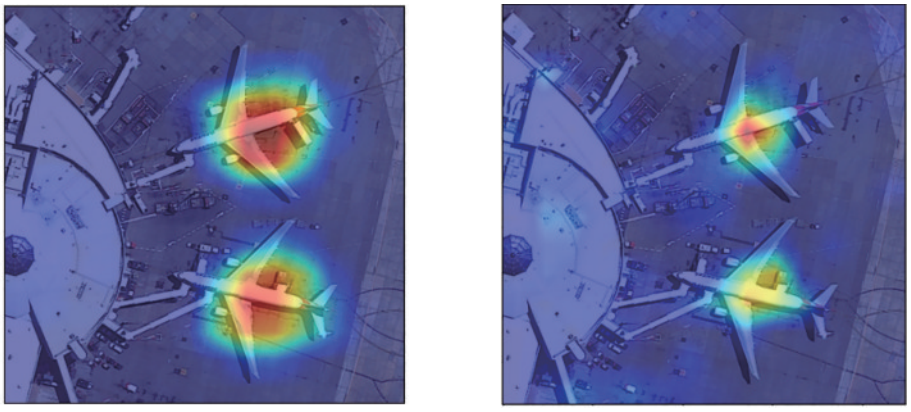

图 3　AO²DINO(左)与 ReDet(右)的注意力热力图对比

Fig.3　Comparison of attention heatmaps between AO²DINO (left) and ReDet (right)

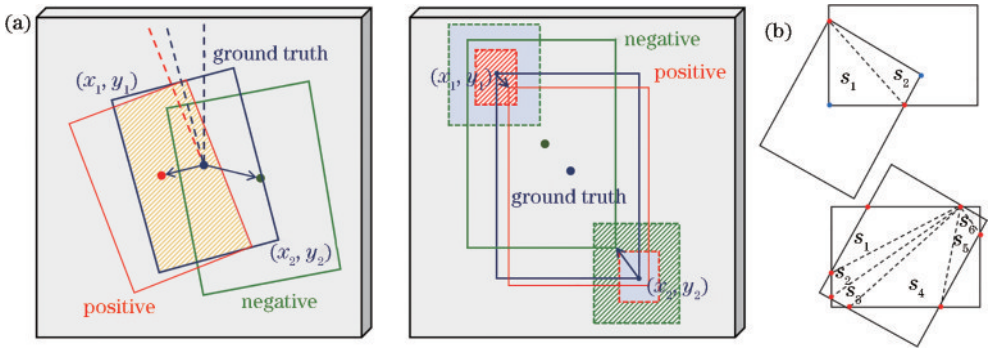

图 4　自适应的正负样本分配器。(a)AO²DINO(左)与 DINO(右)正负样本判别;(b)旋转框的重叠方式

Fig. 4　Self-adaption assigner. (a) Positive and negative assigner of AO²DINO (left) and DINO (right); (b) overlapping of rotated boxes

文提出一个自适应的正负样本分配方法。该方法采用旋转交并比(rotated IoU)计算样本与真值框的交叠程度,根据统计数据自适应地计算阈值,在不增加超参数的基础上,实现更加精确的正负样本判断。其中,在进行对比去噪训练时,针对不同组,每组正负样本 $\boldsymbol{D}_g$ 与其对应真值框 $\boldsymbol{g}$ 交并比均值与标准差之和即为该组适应结果下的阈值,大于该阈值且样本中心点落在真值框内的设置为正样本,其余为负样本。阈值计算公式为

$$V_{\mathrm{Th}}=\mathrm{Mean}\left[\mathrm{IoU}\left(\boldsymbol{D}_g,\boldsymbol{g}\right)\right]+\mathrm{Std}\left[\mathrm{IoU}\left(\boldsymbol{D}_g,\boldsymbol{g}\right)\right]。\tag{6}$$

与水平框相交重叠部分形状一定为矩形,旋转框相交的情况多变,重叠部分可能是四边形、八边形等,计算时情况更加复杂,图 4(b)显示了两种旋转框的重叠方式。具体的旋转框交并比计算方式为:假设旋转

框存在重叠部分，首先将两斜框的交点加入集合 S，即图 4(a)中红点，将集合 S 中的点按照逆时针顺序排序，再用三角剖分法将重叠的多边形分成多个三角形分别计算其面积并相加，最终计算旋转框交并比，具体计算流程如图 5 所示。

Rotated IoU

input：rotated bounding box $R_1, R_2, \cdots, R_N$

output：rotated IoU

for each group of rotated box $\{R_i, R_j\}(i<j)$ do：

add the points where R_i and R_j intersect to the non empty set S

add vertices R_i within R_j to set S

add vertices R_j within R_i to set S

sort the points in set S counterclockwise

calculate overlapping areas using triangulation method：$R_i \cap R_j$

$\mathrm{IoU}(R_i, R_j) = R_i \cap R_j \Big/ \left(R_i + R_j - R_i \cap R_j\right)$

end for

图 5　旋转交并比的计算

Fig. 5　Calculation of Rotated IoU

2.4　基于 KFIoU 的参数联合优化

旋转目标的表示一般有 OpenCV 表示法[20]和长边表示法[8]，通常会引起图 6 所示的角度的周期性(PoA)和图 7 所示的边界交换性(EoE)[8, 20-24]，从而导致预测不准确。

图 6(a)是预测框的理想表示方式，虚线旋转框和实线旋转框分别表示目标的预测框和真值，两者只存在角度和中心点的细微差别。若采用 OpenCV 表示法[20]，如图 6(b)所示，会存在长短边交换的问题，角度 θ_t 与角度 θ_p 相差 90°，且预测的长宽与真值相反，尤其对于大长宽比目标，该问题会更加突出；若采用长边表示法，如图 6(c)所示，角度 θ_t 与角度 θ_p 相差 π。与此同时，对于类正方形的目标，如图 7 所示，其长宽比接近 1∶1，若预测的长边刚好与真值相反，θ_t 与 θ_p 会相差 90°。两种方法都可能会使预测结果超出定义范围，导致模型需以更复杂的方式进行回归，产生很大的回归损失，导致训练不稳定。

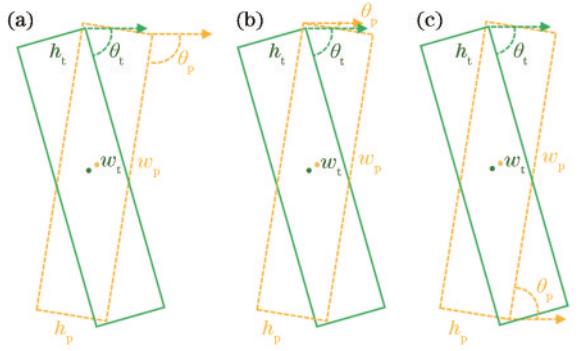

图 6　角度周期性。(a)预测框的理想表示方式；(b) 预测角度与理想角度相差 90°；(c)预测角度与理想角度相差 180°

Fig. 6　Periodicity of angle. (a) Ideal representation of bounding boxes; (b) the predicted angle differs from the ideal angle by 90°; (c) the predicted angle differs from the ideal angle by 180°

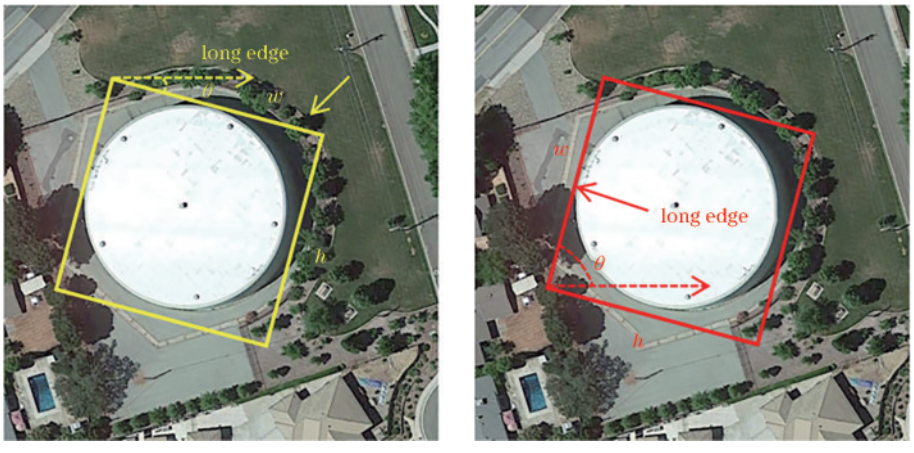

图 7　边界交换性

Fig. 7　Edge exchangeability

常用的旋转目标检测器通常使用 Smooth L1 损失作为回归损失，它对目标表示的各个参数是独立优化的，使得损失对任何参数的欠拟合都很敏感，影响旋转目标的检测精度。为解决上述问题以及 PoA、EoE 问题，GWD[22]和 KLD[25]采用高斯建模，使用高斯分布距离度量，在最终损失函数的设计上引入非线性变换和超参数，使得检测器免疫边界不连续问题，但其本质并不是旋转 IoU(SkewIoU)损失。本文引入 KFIoU[17]损失，即一个更简单高效的 SkewIoU 近似损失。该方法无需额外的超参数，可以使用深度学习框架利用现有的算子轻松实现，提高目标的检测精度。

KFIoU 原理如图 8 所示。首先将旋转矩形框转换为高斯分布，

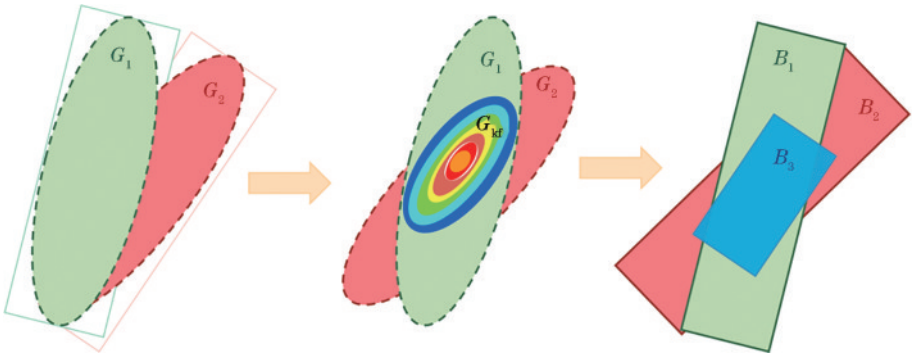

图 8　KFIoU 原理图

Fig. 8　Principle of KFIoU

$$\boldsymbol{\Sigma}^{1/2}=\boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^{\mathrm{T}}=\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\frac{w}{2} & 0\\ 0 & \frac{h}{2}\end{bmatrix}\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}=\begin{bmatrix}\frac{w}{2}\cos^2\theta+\frac{h}{2}\sin^2\theta & \frac{w-h}{2}\cos\theta\sin\theta\\ \frac{w-h}{2}\cos\theta\sin\theta & \frac{h}{2}\cos^2\theta+\frac{w}{2}\sin^2\theta\end{bmatrix},\tag{7}$$

式中：$\boldsymbol{\Sigma}$为旋转矩形框(x, y, w, h, θ)变换高斯分布$G(u, \boldsymbol{\Sigma})$后的方差；$\boldsymbol{R}$表示旋转矩阵；$\boldsymbol{\Lambda}$表示特征值的对角矩阵。

然后，引入中心损失L_c使得两个分布具有相同的中心，通过类似卡尔曼滤波的方式将高斯分布G_1和G_2相乘得到相交区域的高斯分布G_{kf}，但此时高斯分布G_{kf}的协方差只和G_1和G_2的协方差有关，无论高斯分布如何移动，只要它们的协方差固定，计算出的面积就不会改变。由于在第一步中已经引出中心点损失，让两个高斯分布同中心，因此整个损失函数也能优化没有相交的情况。最后，将三个高斯分布反转为旋转矩形框来计算近似的 SkewIoU 损失，即本文引入的 KFIoU 损失，公式为

$$V_{\mathrm{KFIoU}}=\frac{v_{B_3}(\boldsymbol{\Sigma})}{v_{B_1}(\boldsymbol{\Sigma})+v_{B_2}(\boldsymbol{\Sigma})-v_{B_3}(\boldsymbol{\Sigma})},\tag{8}$$

式中：B_1，B_2和B_3为图 8 中的 3 个不同的旋转矩形框。$v_B(\boldsymbol{\Sigma})$为高斯分布计算对应的旋转矩形框的面积，即协方差特征值的累计相乘，具体公式为

$$v_B(\boldsymbol{\Sigma})=2^n\sqrt{\prod\mathrm{eig}(\boldsymbol{\Sigma})}=2^n\cdot\left|\boldsymbol{\Sigma}^{\frac{1}{2}}\right|=2^n\cdot|\boldsymbol{\Sigma}|^{\frac{1}{2}}。\tag{9}$$

上述分析表明，得益于两个高斯分布之间的重叠面积和趋势级别的调整策略，KFIoU 可以独立优化参数，理论上可以得到更高的目标检测精度。

3　实验结果与分析

3.1　数据集介绍与实验设置

使用带有旋转目标标签的 DOTAv1.0[26] 和 DIOR-R[27] 数据集对模型的有效性和适应性进行测试。DOTAv1.0 包含来自不同传感器和航空平台的 2806 张大尺寸航拍图像，188282 个带标注的目标，这些目标分为飞机（PL）、桥梁（BR）、操场（GTF）等 15 类。图像大小从 800×800 像素到 20000×20000 像素不等，包含各类具有不同方向、尺度和外观的目标，是迄今为止最具挑战性的旋转目标检测数据集。DIOR-R 数据集是 DIOR 数据集的扩展，与 DIOR 共享相同的图像，主要用于旋转目标检测，包含 23463 张图像，192518 个标注目标，涵盖飞机（APL）、机场（APO）、篮球场（BC）等 20 个类别。在 DOTAv1.0 数据集的实验中，使用训练集和验证集对模型进行训练，提交至 DOTA 官网进行测试。对于 DIOR-R 数据集，使用训练集验证集进行训练，测试集进行测试。实验配置如表 1 所示。

表 1　实验软硬件配置

Table 1　Experimental software and hardware configuration

Configuration	Model
Operating system	Ununtu 20.0.4
GPU	NVIDIA GeForce RTX-4080Ti GPU
Hardware configuration	i9-10920X
Environment	Python 3.8, PyTorch1.7.1, CUDA11.2

实验训练迭代周期为 12 个 epochs，初始学习率设置为 0.0005，使用动量为 0.9 的随机梯度下降（SGD）作为优化器。对于 DOTAv1.0 数据集，将原始图像按照 200 像素的重叠度分割为多个子图像，子图像大小为 1024×1024。对于 DIOR-R 数据集，图像大小保持 800×800 的原始大小。为避免过拟合，采用随机旋转、随机水平翻转、随机垂直翻转等方法进行数据增强。

3.2　评估标准

目标检测结果主要采用精度（P）、召回率（R）、平均精度均值（mAP）作为评价标准。精度和召回率公式为

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}},R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}},\tag{10}$$

式中：TP是真正例(true positive)，表示被正确预测为正类的正样本；FP是假正例(false positive)，表示被错误预测为正类的负样本；FN是假反例(false negative)，表示被错误预测为负类的正样本。mAP是一种综合了准确率和召回率的评价标准，每个类别都可得到一条PR(precision-recall)曲线。计算曲线下的面积可得到该类别的平均精度(AP)，mAP则为各个类别AP的均值，是评判目标检测网络整体性能最常用的指标之一。

3.3 实验结果分析

不同旋转目标检测器在DOTAv1.0和DIOR-R数据集上的检测结果如表2和表3所示。实验中将模型分为单阶段目标检测方法、双阶段目标检测方法和DETR系列的目标检测方法。其中加粗和下划线字体分别表示所有检测方法中检测精度最高和次高的结果。

表2 DOTAv1.0数据集上各模型对比

Table 2 Comparison of different models on the DOTAv1.0 dataset

Category	One-stage		Two-stage			DETR-like				
	Rotated RetinaNet (3×)	R3Det (3×)	Rotated FCOS (3×)	Rotated Faster R-CNN (3×)	ReDet (3×)	Rotated D-DETR(3×)	AO²DETR (3×)	ARS-DETR (3×)	AO²DINO (1×)	AO²DINO (3×)
PL	87.33	89.24	88.52	**89.09**	88.94	78.95	87.99	86.61	86.33	86.87
BD	78.91	**83.32**	77.54	78.28	78.07	68.64	79.46	77.26	76.79	81.91
BR	46.45	48.03	47.06	48.93	**51.19**	32.57	45.74	48.84	49.52	48.25
GTF	69.81	72.52	63.78	71.54	72.76	55.17	66.64	66.76	63.43	**72.90**
SV	67.72	77.52	**80.42**	74.01	74.26	72.53	78.90	78.38	77.43	79.92
LV	62.34	76.72	80.50	74.99	78.08	57.77	73.90	**78.96**	62.83	63.24
SH	73.59	86.48	87.34	85.90	**87.44**	73.71	73.30	87.40	84.54	85.87
TC	90.85	**90.89**	90.39	90.84	90.84	88.36	90.40	90.61	90.12	88.23
BC	82.79	82.33	77.83	**86.87**	80.79	75.46	80.55	82.76	83.92	82.89
ST	79.37	83.51	84.13	85.03	78.59	79.34	85.89	82.19	84.82	**86.87**
SBF	59.62	60.96	55.45	57.97	60.85	45.36	55.19	54.02	55.94	**61.17**
RA	61.89	63.09	65.84	**69.74**	64.22	53.78	63.62	62.61	67.22	65.97
HA	65.01	67.58	65.02	68.10	**76.84**	52.94	51.83	72.64	68.11	65.60
SP	67.76	69.27	72.77	71.28	72.79	66.35	70.15	72.80	71.89	**77.39**
HC	44.95	49.50	49.17	56.88	54.85	50.38	60.04	64.96	75.40	**77.13**
mAP	69.23	73.40	72.45	73.96	74.03	63.42	70.91	73.79	72.16	**74.07**

旋转目标检测器中，双阶段方法能够在第一阶段提取特定候选框的基础上进行旋转框的精细化回归，因此检测精度一般高于单阶段方法和DETR系列的方法。实验发现，AO²DINO模型的检测精度已逐渐达到甚至超越了包含ReDet在内的一些双阶段检测方法。在DOTAv1.0数据集上，AO²DINO仅训练12个epochs就几乎达到了其他旋转目标检测方法训练36个epochs的检测效果，在DOTAv1.0上训练12个epochs时，mAP为72.16%，训练36个epochs时，mAP为74.07%。训练同等轮次时，在DETR系列模型中AO²DINO的检测精度最高，相比于目前主流的DETR方法AO²DETR[15]和ARS-DETR[16]，分别提升了3.16百分点和0.28百分点。

AO²DINO与AO²DETR和ARS-DETR在DOTAv1.0上的可视化对比如图9所示，AO²DINO对旋转目标的适应性有明显提升，旋转框更加贴合实际目标，同时，有效地规避了角度周期性带来的问题，对大长宽比的目标提升效果尤为明显。

本文采用DIOR-R数据集评估AO²DINO模型的适应性。根据DIOR-R数据集的特性，将输入图像大小调整为800×800，检测目标类别调整为20，使用DIOR-R数据集重新训练和测试模型。检测结果如表3所示，所提AO²DINO是DETR系列模型中性能最佳的，同样训练36个epochs后mAP达到65.94%，比ARS-DETR高0.04百分点。典型场景的可视化检测结果如图10所示。

AO²DINO在小目标检测效果上远优于其他目标检测方法，如表4所示，选取DOTAv1.0数据集中船舶(SH)、小型车辆(SV)，DIOR-R数据集中车辆(VE)、船舶(SH)、风车(WM)作为典型小目标进行比较，平均面积小于32×32像素[28]。通用的旋转目标检测器如R3Det[6]、ReDet[4]、ARS-DETR[16]通常致力于解决旋转目标中的角度问题，而缺乏对遥感图像中大量占比的小目标的考虑，当小目标密集排列时，检测效果不佳。所提多尺度的旋转可变形注意力模块利用包含高级语义信息的多尺度特征图计算注意力权重，提

表 3 DIOR-R 数据集上各模型对比

Table 3 Comparison of different models on the DIOR-R dataset

Category	One-stage			Two-stage				DETR-like		
	Rotated RetinaNet(3×)	R3Det (3×)	Rotated FCOS(3×)	GWD (3×)	KLD (3×)	Rotated Faster R-CNN (3×)	ReDet (3×)	ARS-DETR (3×)	AO^{2}DINO (1×)	AO^{2}DINO (3×)
APL	59.54	62.55	62.31	**69.68**	66.52	63.07	63.22	65.82	63.93	68.78
APO	25.03	43.44	42.18	28.83	46.80	40.22	44.18	**53.40**	42.21	48.83
BF	70.08	71.72	**75.34**	74.32	71.76	71.89	72.11	74.22	73.24	74.32
BC	81.01	81.84	81.32	81.49	81.43	81.36	81.26	81.11	83.57	**84.49**
BR	28.26	36.49	39.26	29.62	40.81	39.67	**43.83**	42.13	40.39	41.62
CH	72.02	72.63	74.89	72.67	**78.25**	72.51	72.72	76.23	63.65	72.67
ESA	55.35	79.50	77.42	76.45	79.23	79.19	79.10	**82.24**	64.91	76.45
ETS	56.77	64.41	68.67	63.14	66.63	69.45	69.78	**71.52**	68.98	69.14
DAM	21.26	27.02	26.00	27.13	29.01	26.00	28.45	**38.90**	33.45	34.13
GF	65.70	77.36	73.94	77.19	78.68	77.93	**78.69**	75.91	71.24	71.19
GTF	70.28	77.17	78.73	78.94	80.19	**82.28**	77.18	77.91	77.03	78.94
HA	30.52	40.53	41.28	39.11	44.88	46.91	**48.24**	33.03	42.67	43.11
OP	44.37	53.33	54.19	42.18	57.23	53.90	56.81	57.02	66.65	**66.18**
SH	77.02	79.66	80.61	79.10	80.91	81.03	81.17	84.82	85.43	**86.10**
STA	59.01	69.22	66.92	70.41	74.17	**75.77**	69.17	69.71	69.80	70.41
STO	59.39	61.10	69.17	58.69	68.02	62.54	62.73	**72.20**	62.34	62.69
TC	81.18	81.54	**87.20**	81.52	81.48	81.42	81.42	80.33	72.98	81.66
TS	38.43	52.18	52.31	47.78	54.63	54.50	54.90	**58.91**	54.55	55.78
VE	39.10	43.57	47.08	44.47	47.80	43.17	44.04	**51.52**	49.80	50.47
WM	61.58	64.13	65.21	62.36	64.41	65.73	66.37	**70.73**	68.21	69.36
mAP	54.83	61.91	63.21	60.31	64.63	63.41	63.81	65.90	60.54	**65.94**

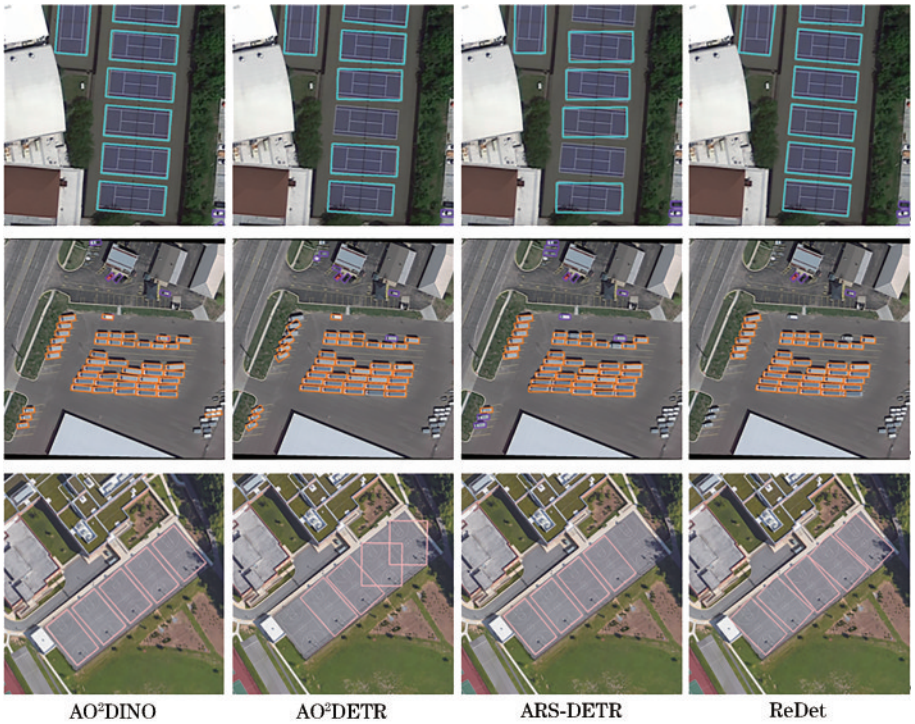

图 9 DOTAv1.0 数据集上不同方法的检测结果对比

Fig. 9 Comparison of test results of different methods on DOTAv1.0 dataset

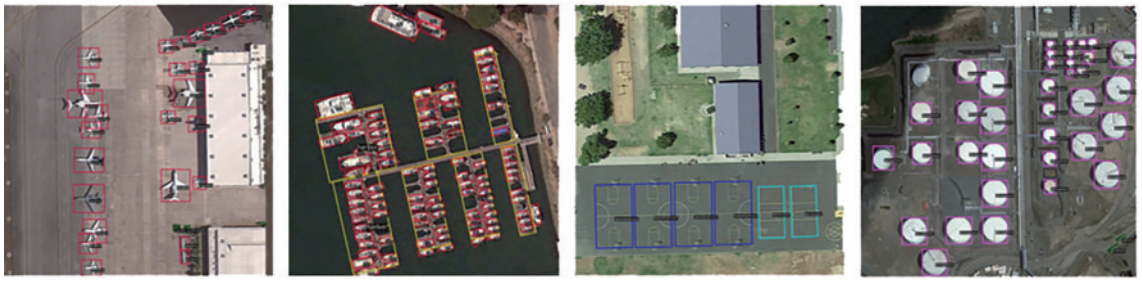

图 10　DIOR-R 上 AO²DINO 的适应性

Fig. 10　Adaptability of AO²DINO on DIOR-R dataset

表 4　DOTAv1.0 和 DIOR-R 数据集中密集小目标检测的 mAP

Table 4　mAP of dense small target detection on DOTAv1.0 and DIOR-R datasets　　unit: %

Method	Epoch	DOTAv1.0		DIOR-R		
		SH	SV	SH	VE	WM
R3Det	3×	86.48	77.52	79.66	43.57	64.13
ReDet	3×	87.44	74.26	81.17	44.04	66.37
ARS-DETR	3×	87.40	78.38	84.82	51.52	70.73
AO²DINO	3×	85.87	79.92	86.10	50.47	69.36
AO²DINO	1×	84.54	77.43	85.43	49.80	68.21
AO²DINO-ms	1×	88.90	79.98	87.57	50.66	70.68

升了对小目标的检测效果。同时，根据旋转交并比和自适应的阈值，实现对密集目标更加精确的采样。本文模型在小目标检测问题上有明显改善，如图 11 所示，主要体现在两方面：首先，当图像中出现小目标密集混乱排列时，能够精准框选出每一个密集排列的小目标（SV），检测框贴合目标实际旋转角度，无误检和漏检等情况，如图 11 第 1 行所示；其次，在准确检测小目标（SH）的同时，不会影响对场景中其他目标（HA）的检测，即使是在小目标密集排列、相互交叠造成背景紊乱的情况下也能保持较好的检测效果，如图 11 第 2 行所示。此外，还采用多尺度训练和测试，进一步提高模型对小目标的检测能力。

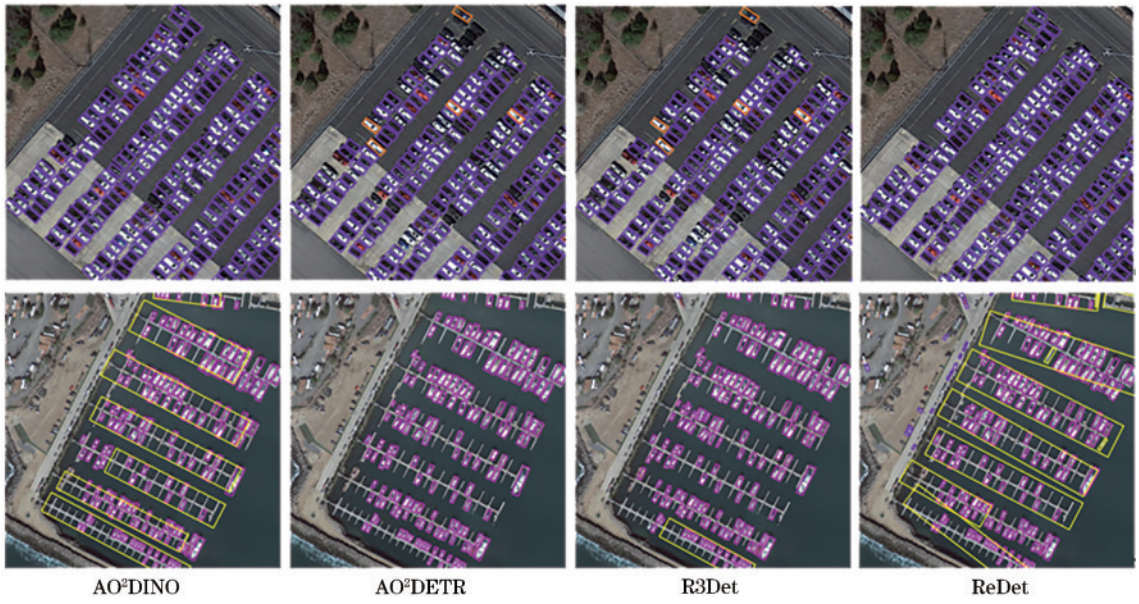

图 11　DOTAv1.0 上密集小目标检测效果对比

Fig. 11　Comparison of dense small object detection performance on DOTAv1.0 dataset

3.4　消融实验

在仅使用去噪训练的 Rotated DINO 的基础上训练 12 个 epochs，利用消融实验分别测试多尺度旋转可变形注意力模块、自适应的样本分配器和 KFIoU 损失函数对模型性能的影响，并对比了 GWD、KLD 和 KFIoU 三种用于旋转目标检测的损失函数的性能。在 DOTAv1.0 数据集上，采用去噪训练的 Rotated DINO 的 AP_{50} 为 67.12%，在此基础上分别添加 RDA、SAA 以及 KFIoU 组件进行实验，结果如表 5 所示，其中 AP_{50} 和 AP_{75} 分别为 IoU 阈值为 0.5 和 0.75 的平均精度。结果显示，各个组件对最终的检测结果均有明显贡献，相比原始模型，分别添加 RDA、SAA 以及 KFIoU 组件的模型的 AP_{50} 分别提升了 1.78 百分点、3.94 百分点和 3.17 百分点。综合使用所有组件时，最终模型的 AP_{50} 进一步提升至 72.16%，充分验证了组件在提升检测性能方面的有效性。与此同时，最终模

表 5 AO²DINO 组件在 DOTAv1.0 数据集上的消融实验

Table 5 Ablation experiments of AO²DINO's component on DOTAv1.0 dataset

CDN	MS-RDA	SAA	KFIoU	AP_{50} /%	AP_{75} /%
√				67.12	33.35
√	√			68.90 (+1.78)	38.70 (+5.35)
√		√		71.06 (+3.94)	40.15 (+6.80)
√			√	70.29 (+3.17)	36.65 (+3.30)
√	√	√	√	72.16 (+5.04)	41.80 (+8.45)

型的 AP_{75} 也得到了显著提升，达到了 41.80%，较原始模型提升了 8.45 百分点。由于在旋转目标检测中，模型在面对图像中一些边界框中心点和宽高相似但角度偏差很大的目标时，更倾向于采用更加严格的度量方式，例如采用 AP_{75} 来评估模型性能，因此 AP_{75} 的显著提升进一步证明了模型在复杂场景下的高精度检测能力。此外，多尺度训练和测试对模型效果提升也十分显著，如表 6 所示，经多尺度训练和测试的模型在 DOTAv1.0 数据集上的 AP_{50} 达到 75.77%，AP_{75} 达到 44.29%。

表 6 AO²DINO 在 DOTAv1.0 数据集上不同尺度的对比实验

Table 6 Comparative experiment of AO²DINO with different scales on DOTAv1.0 dataset

Baseline	Scale	ResNet50	Swin-T	AP_{50} /%	AP_{75} /%
AO²DINO	4 scale	√		72.16	41.80
			√	72.50	42.10
	5 scale	√		72.54	41.73
			√	72.68	42.21
	multi-scale	√		75.77	44.29

GWD、KLD 和 KFIoU 三种损失函数的性能对比如表 7 所示。以 Rotated DINO 为基础框架，采用 GWD、KLD 和 KFIoU 三种损失函数分别训练模型，得到不同模型的 mAP。可以看到无论是在 DOTAv1.0 数据集上还是 DIOR-R 数据集上，KFIoU 损失的 mAP 都明显高于 GWD 和 KLD 损失，能够更好地解决旋转目标的角度周期性问题。

表 7 不同损失函数在 DOTAv1.0 数据集和 DIOR-R 数据集上的 AP_{50}

Table 7 AP_{50} of different loss functions on DOTAv1.0 dataset and DIOR-R dataset unit: %

Loss function	DOTAv1.0	DIOR-R
L1 loss	67.12	53.50
GWD	70.01	55.56
KLD	69.82	55.91
KFIoU	70.29	56.02

4 结 论

提出了一种基于 DETR 的旋转目标检测方法 AO²DINO，显著提高了 DETR 类模型对旋转目标检测的精度。除了使用 5D 的查询(x, y, w, h, θ)外，提出多尺度的旋转可变形注意力模块，将角度信息以旋转矩阵的形式引入注意力计算，使得模型对旋转目标具有更强的适应能力，同时多尺度信息也进一步提升了对小目标的检测效果。针对遥感图像中小目标密集排列问题，提出自适应的样本分配器，以获得更加精确的正负样本，从而提高模型对小目标的检测精度。与此同时，使用 KFIoU 损失代替 Smooth L1 损失，解决了旋转目标检测中的角度周期性问题。AO²DINO 与当前主流的旋转目标检测方法在两个公开数据集 DOTAv1.0 和 DIOR-R 上进行了比较，实验结果表明，AO²DINO 在 DETR 系列旋转目标检测方法中的检测精度最高，同时，在小目标检测精度上超过了当前主流的旋转目标检测方法。AO²DINO 训练收敛速度更快，在训练 12 个 epochs 时就几乎达到了其他旋转目标检测方法训练 36 个 epochs 的检测效果。本文的研究主要针对光学卫星遥感图像数据集，主要面向正下视力图像的目标检测任务，而对斜前视及其他不同视角拍摄的图像检测效果不佳，后续需要将该方法拓展至多视角成像条件下(如无人机载光学影像)的旋转目标检测任务。可将不同视角下的目标检测问题转换为域适应问题，引入对抗训练模块，即 Nuisance Disentangled Feature Transform(NDFT)[29]，提高模型在不同视角以及其他复杂条件下的目标检测能力。同时，为实现机载平台实时目标检测的要求，可结合知识蒸馏的方法[30]，实现适合多视角光学遥感影像的目标检测轻量化模型。

参 考 文 献

[1] Wang K, Wang Z, Li Z, et al. Oriented object detection in optical remote sensing images using deep learning: a survey[EB/OL]. (2023-02-21)[2023-11-09]. http://arxiv.org/abs/1909.00133.

[2] Ding J, Xue N, Long Y, et al. Learning RoI Transformer for oriented object detection in aerial images [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: 2844-2853.

[3] Ma J Q, Shao W Y, Ye H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.

[4] Han J M, Ding J, Xue N, et al. ReDet: a rotation-equivariant detector for aerial object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 2785-2794.

[5] Xie X X, Cheng G, Wang J B, et al. Oriented R-CNN for object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 3500-3509.

[6] Yang X, Yan J C, Feng Z M, et al. R3Det: refined single-stage detector with feature refinement for rotating object[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3163-3171.

[7] 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 (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[8] Han J M, Ding J, Li J, et al. Align deep features for oriented object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2048, 60: 5602511.

[9] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with Transformers[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12346: 213-229.

[10] Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: deformable Transformers for end-to-end object detection [EB/OL]. (2020-10-08) [2023-11-09]. http://arxiv.org/abs/2010.04159.

[11] Liu S L, Li F, Zhang H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR[EB/OL]. (2022-01-28)[2023-11-09]. http://arxiv.org/abs/2201.12329.

[12] Li F, Zhang H, Liu S L, et al. DN-DETR: accelerate DETR training by introducing query DeNoising[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 13609-13617.

[13] Zhang H, Li F, Liu S L, et al. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection[EB/OL]. (2022-03-07) [2023-11-09]. http://arxiv.org/abs/2203.03605.

[14] Ma T L, Mao M Y, Zheng H H, et al. Oriented object detection with Transformer[EB/OL]. (2021-06-06)[2023-11-09]. http://arxiv.org/abs/2106.03146.

[15] Dai L H, Liu H, Tang H, et al. AO_2-DETR: arbitrary-oriented object detection Transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(5): 2342-2356.

[16] Zeng Y, Yang X, Li Q Y, et al. ARS-DETR: aspect ratio sensitive oriented object detection with Transformer [EB/OL]. (2023-03-09) [2023-11-09]. http://arxiv.org/abs/2303.04989.

[17] Yang X, Zhou Y, Zhang G F, et al. The KFIoU loss for rotated object detection[EB/OL]. (2022-01-29)[2023-11-09]. http://arxiv.org/abs/2201.12558.

[18] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22)[2023-11-09]. http://arxiv.org/abs/2010.11929.

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM Press, 2017: 6000-6010.

[20] Yang X, Yan J C, Liao W L, et al. SCRDet++: detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2384-2399.

[21] Qian W, Yang X, Peng S L, et al. Learning modulated loss for rotated object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2458-2466.

[22] Yang X, Yan J C, Ming Q, et al. Rethinking rotated object detection with Gaussian Wasserstein distance loss [EB/OL]. (2021-01-28) [2023-11-09]. http://arxiv.org/abs/2101.11952.

[23] Yang X, Zhang G F, Yang X J, et al. Detecting rotated objects as Gaussian distributions and its 3-D generalization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4335-4354.

[24] Qian W, Yang X, Peng S L, et al. RSDet: point-based modulated loss for more accurate rotated object detection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7869-7879.

[25] Yang X, Yang X J, Yang J R, et al. Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence[EB/OL]. (2021-06-03) [2023-11-09]. http://arxiv.org/abs/2106.01883.

[26] Xia G S, Bai X, Ding J, et al. DOTA: a large-scale dataset for object detection in aerial images[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: 3974-3983.

[27] Cheng G, Wang J B, Li K, et al. Anchor-free oriented proposal generator for object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 1809, 60: 5625411.

[28] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[29] Wu Z Y, Suresh K, Narayanan P, et al. Delving into robust object detection from unmanned aerial vehicles: a deep nuisance disentanglement approach[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 1201-1210.

[30] Li Z, Li X, Yang L F, et al. Curriculum temperature for knowledge distillation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 1504-1512.