通过粗到细的候选生成和模仿学习实现小目标检测

摘要

在过去几年中，目标检测技术取得了显著的进展。但是，尽管当前的优秀检测器表现出色，但在处理尺寸较小的目标时仍然面临许多挑战。具体来说，一个问题是先验框与目标区域之间的低重叠度限制了优化过程中可用的样本数量，而缺乏足够的判别性信息则加剧了这个问题。针对这些难点，我们提出了 CFINet，一种基于粗到细候选生成与特征模仿学习的两阶段小目标检测框架。我们在框架中引入了一个 CRPN，通过动态锚点选择策略和级联回归机制来确保小目标能够获得足够数量的高质量候选区域。接着，我们为传统的检测头设计了一个特征模仿分支，以便更好地对尺寸较小的目标进行区域特征表示建模。为了进一步优化这一分支，我们设计了一种辅助模仿损失，借助对比学习的监督范式，使模型能够更高效地学习小目标的判别特征。在与 Faster RCNN 集成后，CFINet 在大规模小目标检测上实现了显著提高，相较于其他主流检测方法表现出了显著的优势。这一成果表明，CFINet 不仅能够有效缓解小目标检测中的难点问题，同时也为相关领域的研究提供了一个可行的解决方案。

关键词：目标检测；锚点选择策略；特征模仿

1 介绍

小目标检测旨在对区域有限的实例进行分类，在行人检测、自动驾驶、智能监控理解等多种场景中发挥着重要作用。相较于通用目标检测的广泛研究，但是小目标检测受到的关注则相对较少且缺乏好的解决方案。此外，通用的检测器由于两个固有的挑战导致难以处理小目标对象。

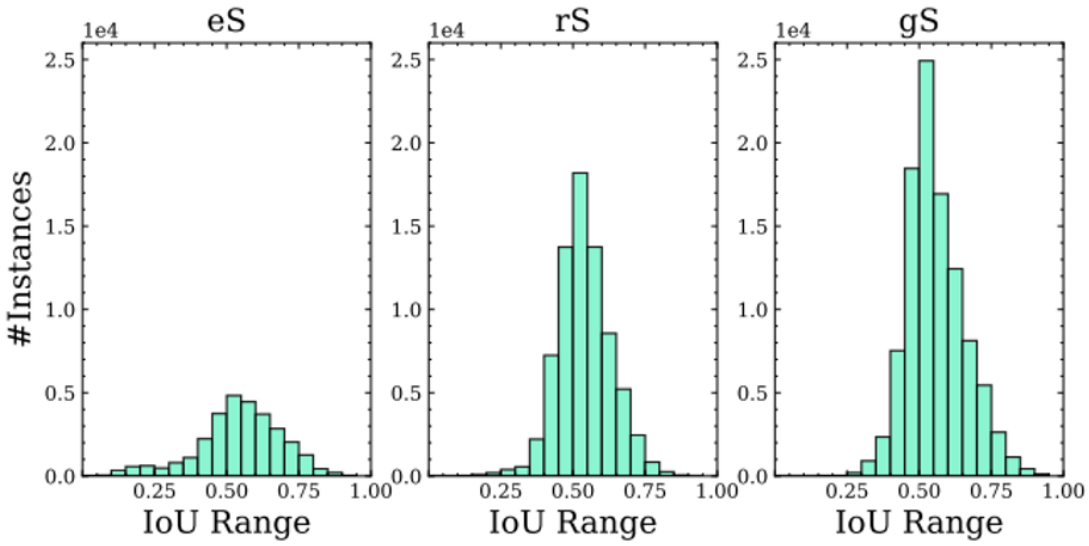

图 1. SODA-D 训练集中的锚点与每个真实目标匹配的最大 IoU 分布，其中极小目标 (eS)、相对较小目标 (rS) 和一般小目标 (gS) 对应 SODA 中的三个面积子集，范围分别为 (0, 144]、(144, 400] 和 (400, 1024]。目标越小，与之匹配的锚点 IoU 值越低，因此常用的正样本 IoU 阈值 (0.7) 对于小目标而言过于苛刻。

目前流行的检测器要么利用基于重叠的，或者利用基于距离的策略来选择用于训练的对象的正先验。然而小实例通常占据极其有限的区域，因此密集排列的锚点和真实框之间的区域重叠非常小，远低于常用的正样本 IoU 阈值，如图 1 所示。换句话说，现有的正样本准则在应用于小对象时过于严格，导致可用于优化的样本数量有限。一种直观的方法是降低定义样本的阈值。然而，虽然这你能够增加正样本数量，但它通常以牺牲整体样本质量为代价，其中的低质量样本会干扰优化并产生简单的回归解影响模型性能。更糟糕的是，这实际上与候选网络的目的是矛盾的，即保证召回率并减轻后续工作的负担。

ctr/ign ratio	AP	AP_{50}	AP_{75}	AP_{es}	AP_{rs}	AP_{gs}	AP_N
Baseline	28.9	59.4	24.1	13.8	25.7	34.5	43.0
0.2/0.5	29.1	56.5	25.9	12.5	25.5	35.4	44.7
0.5/0.8	29.5	57.8	26.0	13.5	26.2	35.8	45.0
0.8/1.0	27.5	54.1	24.3	11.2	23.8	33.7	42.5

表 1. Cascade RPN 与基线 (原始 Faster RCNN) 的性能比较。其中，ctr/ign 比例表示 Cascade RPN 第一回归阶段的采样区域大小比例。结果在 SODA-D 测试集上进行测试，并采用 ResNet-50 作为主干网络。

总而言之，当前流行的从先验到候选区域的范式（依赖于重叠或距离度量）在检测小目标时存在固有的局限性，而现如今设计的分配或采样方案对此问题的改善作用微乎其微。由于候选区域在两阶段检测器中起着至关重要的作用，那么改进的区域候选网络（RPN）变体是否能更好地应对小目标检测呢？沿着这一思路，我们选择了 Cascade RPN，这是一个在通用目标检测中表现优越的候选区域网络，进行初步实验，其结果如表 1 所示。尽管辅助回归阶段通过更好的初始化为后续回归提供了精细化的先验，但最终的结果仍不尽如人意。具体

而言，性能的提升主要来源于较大目标，而 APeS（极小目标平均精度）和 APrS（相对较小目标平均精度）实际上显著下降，这表明基于区域的采样策略倾向于较大的目标，进而使其主导了候选区域网络。同时，扩大采样区域对此问题的改善作用甚微（甚至可能产生负面影响），如表 1 底部所示。因此，粗到细的管道方法有潜力克服传统从先验到候选区域范式的障碍，但关键在于对小目标给予足够的关注。

其次，小目标通常缺乏判别性信息且结构容易失真，这使得模型倾向于给出模糊甚至错误的预测。与此同时，某些数量较大的目标则包含清晰的视觉线索和更强的判别能力。基于这一观察，一些研究提出了缩小小目标与大目标之间表征差距的方法，其中大多数方法借助生成对抗网络或相似性学习，在较大目标（被视为视觉上可信的实例）的指导下，对尺寸受限实例的特征进行超分辨/恢复。然而，这些方法忽视了一个事实：高质量不等于大尺寸，同时小尺寸不等于低质量。换句话说，人类和模型判断样本是否合格的标准是不同的。对于模型而言，这一标准是动态的，应根据检测器当前的优化状态进行调整。此外，这些方法通常需要借助复杂的训练策略或额外的模型，这既耗时，又打破了传统的端到端训练范式。

综合上述部分，我们提出了一种基于粗到细管道和特征模仿学习的两阶段小目标检测器 CFINet。具体而言，受 Cascade RPN 中的多阶段候选区域生成方案的启发，我们设计了粗到细 RPN（CRPN）。该模块首先采用动态锚点选择策略来挖掘潜在先验并进行粗略回归，随后，这些精细化的锚点将通过区域候选网络进行分类和回归。此外，我们在传统的分类与回归设置中扩展出一个辅助特征模仿（FI）分支，该分支可以利用高质量实例的区域特征，指导那些存在不确定/错误预测目标的学习。同时，我们设计了一种基于监督对比学习（SCL）的损失函数来优化整个过程。本文的主要贡献总结如下：

(1) 提出了一种粗到细的候选区域生成管道，即 CRPN，用于执行从锚点到候选区域的过程。该管道通过基于区域的锚点挖掘策略和级联回归机制，为小目标生成高质量的候选区域。

(2) 引入了一个辅助特征模仿（FI）分支，该分支通过高质量实例的监督，丰富了那些困扰模型的低质量实例的特征表示。该新颖分支通过基于 SCL 的定制化损失函数进行优化。

(3) 在 SODA-D 和 SODA-A 数据集上的实验结果展示了我们提出的 CFINet 在检测极限尺寸目标方面的优越性。

2 相关工作

锚点精细化与候选区域生成。两阶段的基于锚点的方法在很大程度上依赖于高质量的候选区域。为此，RPN 首先在 Faster RCNN 中被引入，用于在全卷积网络中生成候选区域，这种简单而有效的设计促进了端到端模型的优化。在 RPN 的基础上，提出了迭代回归预定义锚点的方法。而 GA-RPN 摒弃了传统的统一锚点策略，采用了两步式的锚点生成流程：首先确定可能包含目标的位置，然后在这些位置预测锚点的尺度。通过引入多阶段锚点到候选区域的策略并缓解精细化锚点与图像特征之间的不匹配问题，Cascade RPN 实现了高质量候选区域的生成。然而，目前的候选区域导向框架在生成针对区域有限实例的高质量候选区域方面仍然表现不佳，其根本原因在于目标和先验框之间普遍存在的低重叠问题。与上述方法不同，我们的粗到细候选生成管道能够充分挖掘多阶段精细化范式的潜力，从而在保证候选区域数量的同时，提升尺寸极其有限目标的候选区域质量。

小目标检测的特征模仿。检测小目标的主要挑战之一在于低质量的特征表示，而大目标通

常具有清晰的结构和显著的特征。因此，一系列研究致力于通过挖掘小目标与大目标之间的内在关联来增强小/微小目标的语义表示。基于生成对抗网络（GAN）范式，Perceptual GAN 设计了一种生成器，其目标是生成高质量的小目标特征表示以欺骗后续的判别器。Bai 等人提出了一种新颖的管道，可以从输入的模糊图像中恢复出清晰的人脸。Noh 等人进一步引入了精确的监督机制来改进小目标的超分辨处理。此外，Wu 等人和 Kim 等人都利用相似性学习强制将小规模行人的特征与通过额外模型获得的大规模行人特征对齐。然而，这些方法中存在的超分辨分支或离线特征库阻碍了端到端优化，而我们的方法采用在线方式更新示例特征，确保了高质量特征集合的多样性，从而避免了特征崩塌问题。

目标检测中的对比学习。近年来，自监督学习的兴起主要得益于对比学习范式，并且一些研究已将这种范式扩展到目标检测领域。DetCo 是一个有效的自监督目标检测框架，它利用图像及其局部区域进行对比学习。Wu 等人将对比学习应用于烟雾条件下的目标检测。尽管对比学习最近引起了广泛关注，但利用对比学习来改进小目标特征表示的潜力迄今尚未被深入研究。

3 我们的方法

本节详细介绍了 CFINet 的相关内容。我们首先讨论了 Cascade RPN 在面对小目标时存在的固有局限性，随后介绍了我们针对尺寸受限实例设计的粗到细高质量候选区域生成管道。接着，我们阐述了新设计的特征模仿分支的架构，以及其优化和训练过程。CFINet 的整体架构如图 2 所示。

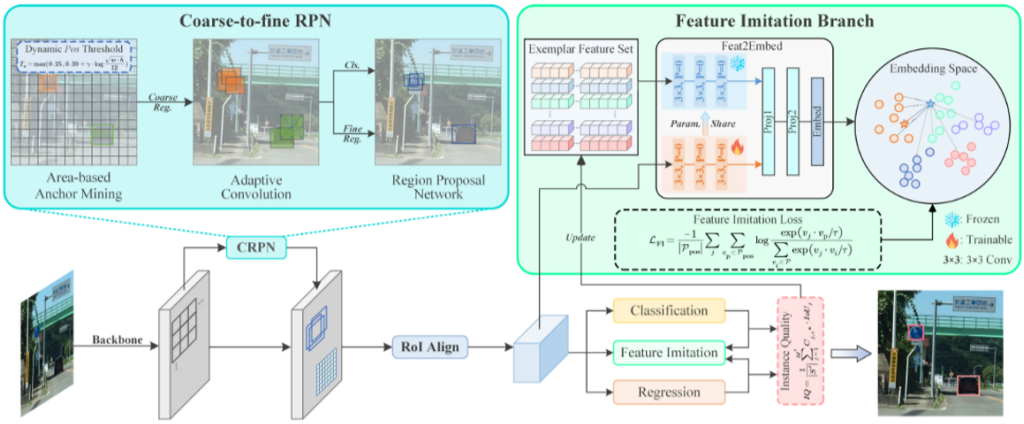

图 2. CFINet 的整体架构。在 CRPN 中，基于区域的锚点挖掘策略基于动态正样本阈值确保了不同尺寸实例（小目标：橙色框，大目标：绿色框）的足够候选区域，这些候选区域将被用于生成粗候选区域。随后，通过自适应卷积实现粗候选区域与其对应特征之间的对齐，然后输入到 RPN 中以生成高质量候选区域。特征模仿分支旨在增强小目标的特征表示。在该分支中，不确定或错误预测的 RoI 特征会在嵌入空间中通过 Feat2Embed 模块被拉近至其类别对应的示例特征集合，同时远离其他类别和背景的示例特征。而示例特征的收集基于模型预测结果，并使用提出的质量指标 IQ 进行筛选。我们为这一辅助分支设计了特征模仿损失函数 LFI 来优化其表现。需要注意的是，为了清晰展示，此处仅展示了单层特征金字塔网络的特征。

3.1 更优的候选区域

Cascade RPN 的局限性。高质量的候选区域在两阶段检测器中起着关键作用，但传统方法依赖于启发式的锚点设置。Cascade RPN 摒弃了这种传统方法，在每个特征点上仅放置一个锚点，并通过多阶段精细化进行优化。尽管在一般尺度目标上表现优越，但 Cascade RPN 在处理极小目标时却存在固有的局限性。具体而言，第一阶段回归中使用的距离度量无法为小目标（中心区域显著较小的目标）提供足够的潜在锚点。此外，Cascade RPN 仅在单一特征金字塔层中将符合条件的锚点标记为正样本，而这种启发式方法简单地丢弃了其他层可能存在的锚点，这些锚点仍然可以提供小目标的存在信息和粗略的位置信息。

粗到细 RPN 为了解决 Cascade RPN 在处理小目标时的上述问题，我们提出了粗到细 RPN，其详细结构如图 2 所示。首先，我们设计了一种基于区域的锚点选择策略，使得不同尺寸的目标都可以拥有（相对）足够的潜在锚点。具体而言，对于一个宽度为 w、高度为 h 的目标框，任何与其 IoU 大于阈值 Ta 的锚点都被视为正样本用于粗回归。阈值 Ta 的定义如下：

$$T_a = \max\left(0.25, 0.20 + \gamma \cdot \log\left(\frac{\sqrt{w \cdot h}}{12}\right)\right), \tag{1}$$

其中，y 表示一个比例因子，在我们的实验中默认设置为 0.15，而其中的 12 实际上对应于 SODA 数据集中极小目标的最小面积定义。这种设置能够为极限尺寸的目标提供足够的样本，同时可以根据不同的数据集进行调整。此外， 和 max 操作可以避免优化过程被低质量的先验锚点所干扰。以 IoU 作为挖掘潜在锚点的标准，可以避免 Cascade RPN 多阶段回归中的优化不一致问题。同时，基于提出的连续阈值，模型能够以更平滑的方式判定正样本。

与 Cascade RPN 不同的是，我们保留了所有特征金字塔网络层级 P2, P3, P4, P5 中的锚点以执行第一阶段的回归。通过这种方式，我们能够为极小目标挖掘足够的潜在锚点，同时，较大的目标由于其匹配锚点的 IoU 值自然较高（如图 1 所示），依然能够获得适当的关注。在第一阶段回归之后，我们捕捉回归框内部的偏移量，并将其与特征图一起输入到 RPN 中，在此过程中利用 Adaptive Convolution 对特征进行对齐，并执行第二阶段的回归和前景-背景分类。

损失函数。我们的训练目标是：

$$L_{\text{CRPN}} = \alpha_1 \left(L^c_{\text{reg}} + L^f_{\text{reg}}\right) + \alpha_2 L_{\text{cls}}, \tag{2}$$

其中，Lcls 和 Lreg 分别采用交叉熵损失和 IoU 损失。公式中的 c 和 f 分别表示 CRPN 中的粗阶段和精阶段，需要注意的是我们仅在精阶段进行分类。损失权重 1 和 2 分别设置为 9.0 和 0.9 。

3.2 小目标检测的特征模仿

近年来，许多研究试图挖掘不同尺度目标之间的内在关联以增强小目标的特征表示，但大多在有效性和多样性方面未取得令人满意的效果。具体而言，大多数早期方法借助 GAN 技术对小目标的特征进行超分辨，这需要复杂的训练方案，并且容易生成伪纹理和伪影。另一类方法则转向相似性学习，但这类方法要么需要以繁琐的方式构建离线特征库，要么直接使

用 L2 范数进行不同 RoI 特征间的相似性度量，这可能导致特征崩塌问题：修正后的区域特征可能具有高相似性，但却丢失了自身的特性。这种特征空间的同质化实际上削弱了模型的泛化性和鲁棒性。

为了缓解特征崩塌风险、避免内存开销并实现端到端优化，我们设计了一个特征模仿头（见图 2）。最重要的是，我们不再仅仅使用大尺度目标作为指导，而是结合当前模型对每个实例的响应，动态构建一个在线更新的、经过优化的高质量示例特征库。FI 分支主要由一个示例特征集合和一个特征到嵌入模块组成，其中前者保存高质量示例的 RoI 特征，后者将输入特征映射到嵌入空间。接下来我们详细介绍特征模仿分支的设计。正如我们之前讨论的，示例在模仿学习中至关重要。为了确定能够为当前模型困惑的小目标提供有效指导/监督的最具代表性、合适且高质量的示例，我们首先为每个实例引入了一个简单的质量指标。假设一个真实目标 g = (c* , b*)，其中 c* 和 b* 分别表示其类别标签和边界框坐标。假设检测头对目标 g 的预测输出为一个集合 S = {Ci , IoUi}i=1,2,...,M，其中 $Ci \in RN + 1$ 表示预测的分类向量，IoUi 表示预测框与真实框之间的 IoU 值，N 是前景类别的数量。我们可以得到潜在的高质量预测子集：

$$S' = \{(C_j, \mathrm{IoU}_j) \mid \arg\max(C_j) = c^*\}_{j=1,2,\ldots,M'}, \tag{3}$$

其中 M' $\leq M$

现在，我们将目标 g 的实例质量定义为：

$$IQ = \frac{1}{|\mathcal{S}'|} \sum_{j=1}^{M'} C_{j,c^*} \cdot \mathrm{IoU}_j \tag{4}$$

GT 的 IQ 可以作为当前模型检测能力的指示器，使我们能够捕捉到具有精准定位和高置信分类分数的高质量样本，而那些让模型困惑的实例通常无法满足这两项要求。通过设定适当的阈值，我们可以选择合适的实例来构建教师特征集，并执行模仿过程。

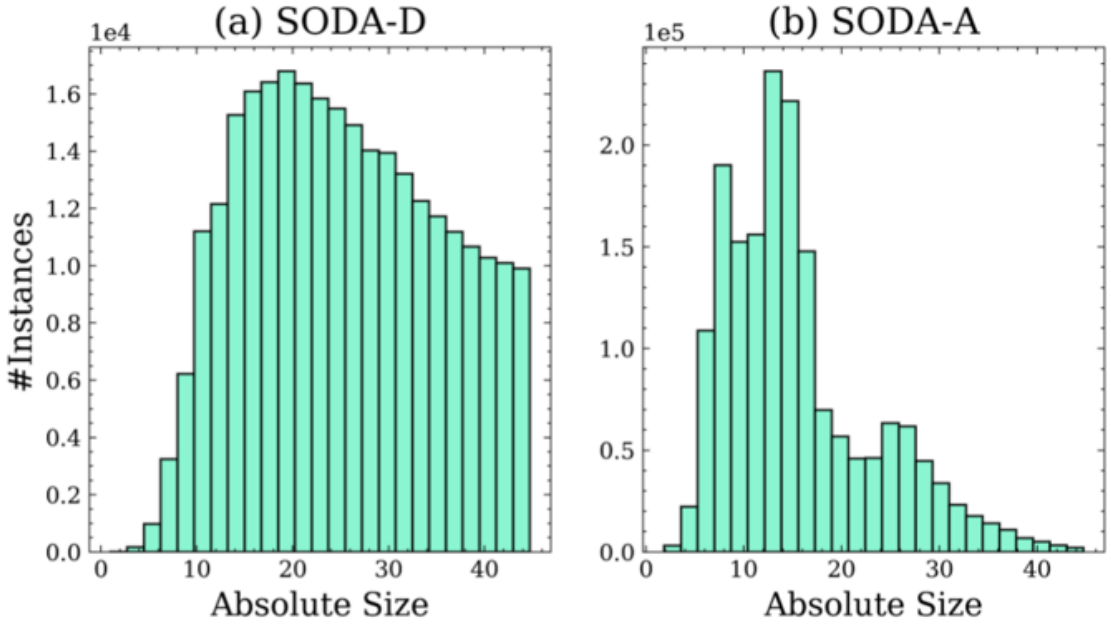

图 3. (a) SODA-D 和 (b) SODA-A 中实例的大小分布，其中绝对大小对应于目标区域的平方根。Feat2Embed 模块。与直接测量不同 RoI 特征之间的相似性相比，我们首先通过简单的

Feat2Embed 模块。与直接测量不同 RoI 特征之间的相似性相比，我们首先通过简单的 Feat2Embed 模块对这些特征进行嵌入处理。FI 分支的输入是通过 RoI 操作获得的区域特征 xi RH×W×C，该特征首先经过三个连续的 3×3 卷积层处理以提取紧凑的表示。值得注意的是，我们在提取当前区域特征时会更新参数，而在提取示例特征时会冻结参数，从而提高性能的稳定性。随后，中间特征将在两层感知器和嵌入层（嵌入层维度为 128，隐藏层维度为 512）的基础上被映射到嵌入空间。我们还探索了各种设计选择和结构用于 Feat2Embed 模块，详细信息可以在补充材料中找到。

最终，特征模仿分支的输出定义为：

$$v_i = \Theta_{\mathrm{FI}}(x_i), \tag{5}$$

其中，ΘFI 表示特征模仿（Feature Imitation）分支中需要优化的参数。损失函数。FI 头的目标非常简单：计算候选区域的 RoI 特征与嵌入空间中存储的高质量实例特征之间的相似性，从而将那些使模型困惑的实例的特征拉近到其所属类别的示例特征，同时远离其他类别和背景的特征。为此，我们提出了一种基于监督对比学习的损失函数，该方法扩展了对比学习的设置，允许通过利用可用的标签信息为一个锚点对象提供多个正样本。我们为 FI 分支量身定制的损失函数如下：

$$\mathcal{L}_{\mathrm{FI}} = \frac{-1}{|\mathcal{P}_{\mathrm{pos}}|} \sum_{j} \sum_{v_p \in \mathcal{P}_{\mathrm{pos}}} \log \frac{\exp\left(v_j \cdot v_p / \tau\right)}{\sum_{v_i \in \mathcal{P}} \exp\left(v_j \cdot v_i / \tau\right)} \tag{6}$$

其中，P = Ppos Pneg 表示样本集合，Ppos 和 Pneg 分别表示正样本集和负样本集，理想情况下它们的基数（样本数量）相同，vp 和 vn 分别是来自 Ppos 和 Pneg 的正样本和负样本。此外，j 表示当前候选框）的索引，T 表示温度参数，在对比学习中起着至关重要的作用，需要精心设计。我们通过消融实验（见表 9）来确定在我们的框架中最优的设置。总的损失函数如下所示：

$$\mathcal{L} = \mathcal{L}_{\mathrm{CRPN}} + \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{reg}} + \alpha_3 \mathcal{L}_{\mathrm{FI}}$$

其中，Lcls 和 Lreg 是检测头的原始损失函数，3 用于调整特征模仿部分的权重。在对比学习的设置下，我们不仅能够完成模仿学习，还能够防止特征崩塌问题，从而有效提升小目标的特征表示能力。此外，模仿过程仅在训练阶段进行，不会影响推理阶段的速度。

训练。接下来我们阐述 FI 分支的训练细节。示例集 E = Eic=1,2,⋯,N 包含了 N 个前景类别的高质量特征，其中 Ei = xi,jj=1,2,⋯,Ni 对应于第 i 类的示例特征，Ni 表示该类别示例的数量。我们使用 Thq 挑选出那些适合作为优质示例的高质量实例。在实践中，我们为实例的高质量预测数量设置一个上限值，以过滤掉网络波动的影响。函数 Γ 用于对高质量实例的特征进行增强，即高质量实例的正样本特征是其自身的变换。FI 分支的整体训练过程如算法 1 所示，更多细节请参阅补充材料。

Algorithm 1 Training of Feature Imitation branch.

Input: The set of GT boxes $\mathcal{G} = \{c_i^*, b_i^*\}_{i=1,2,...,T}$ and corresponding RoI features $\{x_i^g\}_{i=1,2,...,T}$ in current batch; The set of exemplar features $\mathcal{E} = \{\mathcal{E}_i\}_{c=1,2,...,N}$; The set of background RoI features in current batch χ_{bg}; The threshold of high-quality T_{hq}; The number of pos/neg samples N_{pos} and N_{neg}; The transformation function Γ;

Output: The set of positive embeddings $\mathcal{P}_{\text{pos}}$ and negative embeddings $\mathcal{P}_{\text{neg}}$;

Initialize the set of positive features $\mathcal{X}_{\text{pos}}$ and negative features $\mathcal{X}_{\text{neg}}$ with $\emptyset$;
for g in $\mathcal{G}$ **do**
    Compute the IQ of current g according to Eq.(3);
    $\mathcal{X}_{\text{neg}}^g \leftarrow$ sample N_{neg} features from $\mathcal{X}_{\text{bg}} \cup \mathcal{E} \setminus \mathcal{E}_{c^*}$;
    if $IQ \geq T_{hq}$ **then**
        $\mathcal{E}_{c^*} \leftarrow x_i^g$;
        $\mathcal{X}_{\text{pos}}^g \leftarrow \Gamma(x_i^g)$;
    else
        $\mathcal{X}_{\text{pos}}^g \leftarrow$ sample N_{pos} features from $\mathcal{E}_{c^*}$;
    end if
    $\mathcal{X}_{\text{pos}} = \mathcal{X}_{\text{pos}} \cup \mathcal{X}_{\text{pos}}^g, \mathcal{X}_{\text{neg}} = \mathcal{X}_{\text{neg}} \cup \mathcal{X}_{\text{neg}}^g$;
    Apply Eq.(4) to $\mathcal{X}_{\text{pos}}$ and $\mathcal{X}_{\text{neg}}^g$ to obtain $\mathcal{P}_{\text{pos}}$ and $\mathcal{P}_{\text{neg}}$;
end for
return $\mathcal{P}_{\text{pos}}$ and $\mathcal{P}_{\text{neg}}$;

4 实验结果

4.1 数据集

为了评估我们方法的有效性，我们在最近发布的专门针对小目标检测的大规模基准数据集 SODA 上进行了广泛实验，包括 SODA-D 和 SODA-A 两个子集。

SODA-D 聚焦于驾驶场景，包含 24,828 张高质量图像和 278,433 个实例，分布在以下九个类别中：行人、骑手、自行车、摩托车、车辆、交通标志、交通信号灯、交通摄像头和警示锥。SODA-D 的一个显著优势是其多样性，包括时间段、地理位置、天气条件和摄像头视角等方面的丰富变化。

SODA-A 包含 2513 张航拍图像和 872,069 个带方向框标注的目标，涵盖九个类别：飞机、直升机、小型车辆、大型车辆、船只、集装箱、储罐、游泳池和风车。SODA-A 中的目标可以以任意方向出现，并伴有显著的密度变化。具体来说，SODA-A 每张图像的平均实例数量约为 350 个。

作为一个专门针对小目标检测的基准数据集，SODA 中的目标非常小，大多数目标的平均大小在 10 到 30 像素之间（见图 3）。与传统的目标检测数据集不同，SODA 包含了大量的忽略标注，用于过滤掉那些过大或由于严重遮挡或镜头眩光而难以被明确识别的实例。这一策略有助于模型更专注于有价值的小目标。

在 SODA 中，目标根据面积被划分为小目标和正常目标两类，其中小目标进一步细分为三个子集：极小目标、相对较小目标和一般小目标。SODA 的评价指标与 COCO 的评价方式

一致，即在 10 个 IoU 阈值（从 0.5 到 0.95，间隔为 0.05）上计算平均精度。不过，SODA 的评价特别注重对小目标的检测性能。

4.2 实现细节

在以下实验中，除非另有说明，我们使用训练集进行训练，并将测试集用于性能比较和消融研究。考虑到 SODA 中的图像具有非常高的分辨率（约 4000 × 3000），我们首先将原始图像拆分成一系列 800×800 的图块，步长为 650，并且类似于，这些图块将在训练和测试过程中调整为 1200×1200。本文中的所有实验均在单个 RTX 3090 上进行，批量大小为 4。数据增强仅涉及随机翻转。我们使用 1× 的训练计划（共 12 个 epoch），学习率设置为 0.01，并在第 8 和第 11 个 epoch 时分别衰减 0.1。默认的优化器是 SGD，动量为 0.9，权重衰减为 0.0001。我们使用带有 FPN 的 ResNet-50 作为所有模型的基础。

Method	**Publication**	**Schedule**	**AP**	**AP**$_{50}$	**AP**$_{75}$	**AP**$_{es}$	**AP**$_{rs}$	**AP**$_{gs}$	**AP**$_{N}$
One-stage									
Rotated RetinaNet[26]	ICCV'17	1X	26.8	63.4	16.2	9.1	22.0	35.4	28.2
S^2A-Net[15]	TGRS'22	1X	28.3	69.6	13.1	10.2	22.8	35.8	29.5
Oriented RepPoints[23]	CVPR'22	1x1x1x	26.3	58.8	19.0	9.4	22.6	32.4	28.5
DHRec[27]	TPAMI'22	1x	30.1	68.8	19.8	10.6	24.6	40.3	34.6
Two-stage									
Baseline[29]	NeurIPS'15	1X	32.5	70.1	24.3	11.9	27.3	42.2	34.4
Gliding Vertex[41]	TPAMI'21	1x	31.7	70.8	22.6	11.7	27.0	41.1	33.8
Oriented RCNN[39]	ICCV'21	1X	34.4	70.7	28.6	12.5	28.6	44.5	36.7
DODet[8]	TGRS'22	1x	31.6	68.1	23.4	11.3	26.3	41.0	33.5
CFINet(ours)	-	1x	34.4	73.1	26.1	13.5	29.3	44.0	35.9

表 2. 与最先进检测方法在 SODA-D 测试集上的比较，其中 "Baseline" 指的是 Faster RCNN，作为表中两阶段方法的基准。除 YOLOX (CSP-Darknet) 和 CornerNet (HourglassNet-104) 外，所有方法都在 ResNet-50 上训练。'Schedule' 表示训练的 epoch 数量，其中 '1×' 表示 12 个 epoch，'50e' 表示 50 个 epoch。

4.3 主要结果

为了展示我们方法的有效性，我们在 SODA-D 和 SODA-A 数据集上，与当前具有代表性的方法进行了全面对比。表 2 展示了我们的方法与几种主流方法在 SODA-D 测试集上的结果。结合 Faster RCNN，我们 CFINet 实现了最先进的性能，整体 AP 达到 30.7 %，比基线模型高出 1.8 个百分点。在具体的评估指标上，我们的方法展现出了显著的优势，特别是在最具挑战性的 APeS 和 APrS 指标上。此外，与专门为小目标检测设计的 RFLA 方法相比，在同样基于 Faster RCNN 的前提下，CFINet 在 AP、APeS 和 APrS 上分别领先 1.0%、1.5% 和 0.9%。实际上，RFLA 在极小目标检测上牺牲了性能（相比 Faster RCNN，APeS 指标下降了 0.6 个百分点）。

在 SODA-A 测试集上，CFINet 同样取得了最佳结果，特别是在 APeS 指标上展现出了显著的优势（见表 3），进一步证明了方法的优越性和泛化能力。此外，尽管 Oriented RCNN 在

Method	Publication	Schedule	AP	AP_{50}	AP_{75}	AP_{es}	AP_{rs}	AP_{gs}	AP_N
One-stage									
Rotated RetinaNet[26]	ICCV'17	1x	26.8	63.4	16.2	9.1	22.0	35.4	28.2
S^2A-Net[15]	TGRS'22	1X	28.3	69.6	13.1	10.2	22.8	35.8	29.5
Oriented RepPoints[23]	CVPR'22	1X	26.3	58.8	19.0	9.4	22.6	32.4	28.5
DHRec[27]	TPAMI'22	1x	30.1	68.8	19.8	10.6	24.6	40.3	34.6
Two-stage									
Baseline[29]	NeurIPS'15	1X	32.5	70.1	24.3	11.9	27.3	42.2	34.4
Gliding Vertex[41]	TPAMI'21	1X	31.7	70.8	22.6	11.7	27.0	41.1	33.8
Oriented RCNN[39]	ICCV'21	1x	34.4	70.7	28.6	12.5	28.6	44.5	36.7
DODet[8]	TGRS'22	1x	31.6	68.1	23.4	11.3	26.3	41.0	33.5
CFINet(ours)	-	1x	34.4	73.1	26.1	13.5	29.3	44.0	35.9

表 3. 与最先进检测方法在 SODA-A 测试集上的比较，其中 "Baseline" 指的是 Rotated Faster RCNN [29]，作为表中两阶段方法的基准。其他设置与表 2 一致。

AP75 指标和较大目标实例上表现出了一定的优势，但在 AP50（73.1% vs. 70.7%）和 APeS（13.5% vs. 12.5%）指标上远远落后于我们的方法，这表明了我们方法的强大性能。

4.4 CRPN 的有效性

本文的主要设计之一是 CRPN 。其设计基于以下观察：当前基于固定重叠的采样范式由于内在矛盾，不适合小目标实例；虽然 RPN 的优化设计可以部分减轻这一问题，但仍无法获得令人满意的结果。我们在此通过全面分析，展示了 CRPN 在生成高质量小目标提议上的能力。

我们首先在表 4 中展示了 CRPN 和其对比方法的召回性能。可以看到，适当降低正样本阈值略微提高了平均召回率，但同时牺牲了较大目标实例的性能（ARN 的召回率从 57.1% 降至 54.1%）。如我们之前讨论的，GA-RPN 和 Cascade RPN 都未能取得更好的结果，因为它们的模式倾向于较大目标。与 RPN 及其变体相比，CRPN 在小目标上的表现优于其他方法，同时在普通目标上也表现出可比的性能。这验证了我们关于优化提议网络更倾向于较大目标的假设。

Proposal Method	AR	AR_{es}	AR_{rs}	AR_{gs}	AR_N
RPN [?]	41.2	24.0	38.3	47.3	57.1
RPN-0.5	41.3	24.2	38.5	47.3	54.1
GA-RPN [?]	42.1	24.1	39.2	48.9	56.2
Cascade RPN [?]	41.8	22.8	38.2	48.7	57.1
CRPN	**42.6**	**24.6**	38.9	**49.1**	56.9

表 4. 我们的 CRPN 及其对比方法在 SODA-D 测试集上的平均召回率（AR）表现。所有方法均以 Faster RCNN（ResNet-50）作为基准，并采用 1× 训练计划，其中 RPN 表示 Faster RCNN 的原始版本，RPN 阶段的正样本阈值设置为 0.7，而 RPN-0.5 表示正样本阈值为 0.5 的版本。结果是在每张图像上测试 300 个提议。

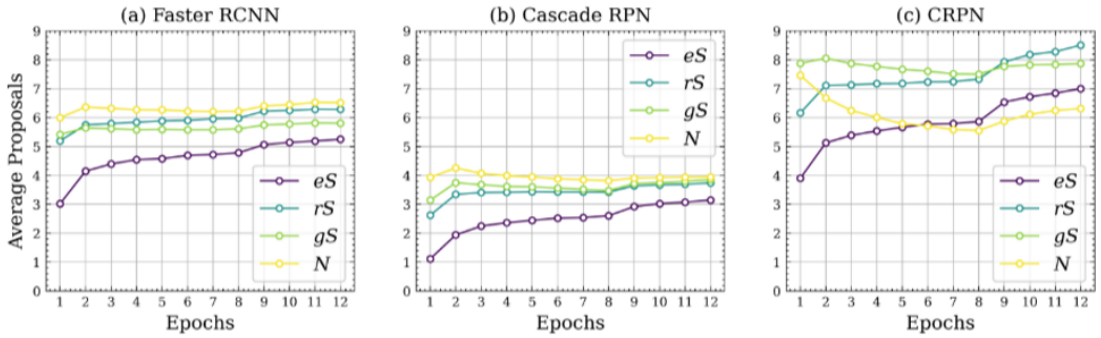

图 4. (a) RPN、(b) Cascade RPN 和 (c) CRPN 在极小（eS）、相对小（rS）、一般小（gS）和正常（N）子集中的平均高质量提议数量。值得注意的是，如果某个提议与任何真实框的 IoU 大于 0.5，则该提议会被视为高质量提议。

我们推测，实现准确的小目标检测的一个主要挑战是缺乏高质量的样本，这也是 CRPN 设计的主要动机。因此，我们直观地比较了基线 RPN、Cascade RPN 和我们的 CRPN 在高质量样本数量上的表现。如图 4 所示，CRPN 生成的高质量提议明显多于其他方法。更有趣的是，CRPN 可以在训练过程中动态调整焦点：在训练初期，模型更关注有利于早期优化的大目标，而随着训练的进行，模型逐渐将注意力转向之前难以处理的小目标。这可以解释为，由于极小目标的不确定性较高，在训练初期拟合这些目标不是检测器的最优选择。

4.5 消融实验与讨论

在本部分，我们进行了消融实验以及全面的讨论，以验证 CRPN 和 FI 分支的重要性，并确定我们方法的最佳设置。所有实验均在 SODA-D 测试集上进行。

设计模块的研究。我们首先进行消融实验，以验证两个模块的有效性。如表 5 所示，CRPN 和 FI 分支均能够稳定提升性能，而引入特征模仿（FI）分支有助于小目标实例的识别。整合 FI 和 CRPN 后，性能达到最佳，因为 CRPN 能够生成足够多的高质量提议（如图 4 所示），从而更准确地指示目标质量并具有作为样本的潜力。

固定或动态。一个自然的想法是直接设置固定的正样本阈值，以为 CRPN 的单阶段回归获取更多锚框。然而，我们的方法采用的基于区域的锚框挖掘策略简单而有效，并能达到最

Baseline	**CRPN**	**FI**	**AP**	AP_{es}	AP_{rs}	AP_{gs}
✓			28.9	13.8	25.7	34.5
✓	✓		30.3	14.3	27.3	36.1
✓		✓	29.5	14.4	26.3	35.1
✓	✓	✓	30.7	14.7	27.8	36.4

表 5. Performance comparison with different configurations.

Strategy	AP	AP_{es}	AP_{rs}	AP_{gs}
0.20	29.9	13.7	26.8	35.9
0.40	30.1	14.1	26.9	**36.2**
Ours	**30.3**	**14.3**	**27.3**	36.1

表 6. CRPN 中正锚点的不同定义，其中“我们的”表示提出的动态策略。

佳性能。如表 6 所示，当第一阶段回归的正样本 IoU 阈值降至 0.20 时，APeS 仅为 13.7

Strategy	AP	AP_{es}	AP_{rs}	AP_{gs}
0.20	29.9	13.7	26.8	35.9
0.40	30.1	14.1	26.9	**36.2**
Ours	**30.3**	**14.3**	**27.3**	36.1

表 7. 特征模仿分支的损失权重影响。

特征模仿损失的权重。在本节中，我们分析了 FI 分支权重参数（即超参数 3）对模型的影响。如表 7 所示，过少或过多地关注 FI 分支都会降低最终性能。因此，我们在实验中将 3 设置为 0.5，以确保整体精度。

作为样本的标准。样本质量在模仿过程中起着至关重要的作用。接下来我们讨论捕获高质量样本用于构建特征集的选择。如表 8 所示，降低 Thq（质量阈值）会增加样本数量并更频繁地更新教师集，而提高 Thq 则相反。可以看到，当 Thq 从 0.5 增加到 0.65 时，有利于模仿过程；但当 Thq 达到 0.70 时，总体 AP 反而下降。这可能是由于早期存储在特征集中的样本不适合当前状态，因为优化是动态的，模型在不断演变，因此作为样本的标准也已发生改变。

温度参数。温度参数在对比学习中至关重要。我们通过一系列实验验证了 的最佳选择。如表 9 所示，当 范围从 0.10 到 0.80 时，总体性能先增加后下降至 30.2

T_{hq}	AP	AP_{es}	AP_{rs}	AP_{gs}
0.50	30.3	14.0	27.3	36.0
0.55	30.6	14.3	27.3	36.5
0.65	30.7	14.7	27.8	36.4
0.70	30.5	14.6	27.3	36.3

表 8. 作为示例实例的标准调查。

τ	AP	AP_{es}	AP_{rs}	AP_{gs}
0.10	30.3	14.1	27.3	36.1
0.50	30.4	14.4	27.2	36.2
0.60	30.7	14.7	27.8	36.4
0.80	30.2	14.0	27.3	36.0

表 9. 温度　对最终性能的选择。

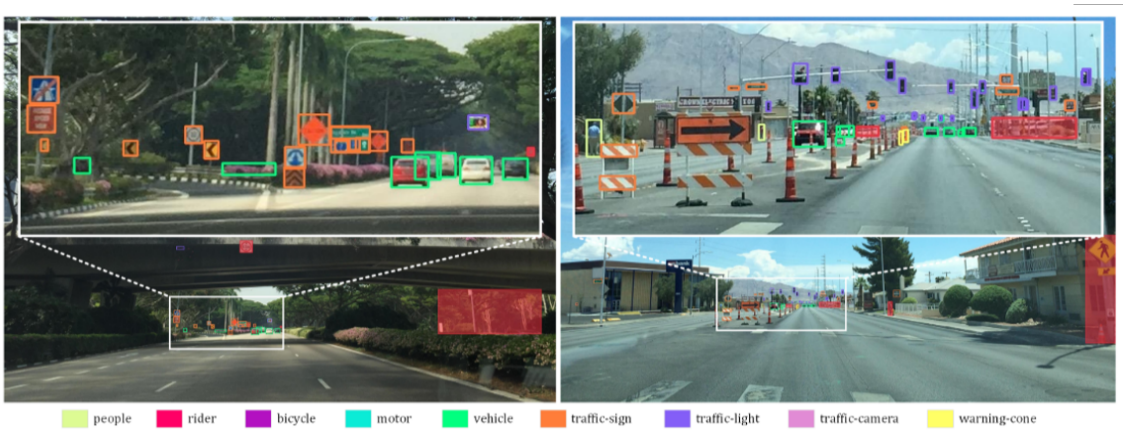

图 5. 我们方法在 SODA-D 测试集上的定性结果。仅展示置信度大于 0.3 的预测，遮罩框表示忽略区域。最佳效果请在彩色和放大窗口中查看。

可视化。为了展示检测器在检测小目标时的能力，我们在图 5 中展示了 SODA-D 测试集的示例图像的可视化结果。

5　结论

在本文中，我们提出了 CFINet，这是一种基于粗到细区域候选网络和特征模仿设置的两阶段检测器。其中，前者能够为小目标生成充足的高质量候选区域。而基于特征模仿分支的新型检测头则通过对比学习范式，有效提升了对模型具有挑战性的小目标的特征表示能力。实验结果表明，我们的方法在大规模小目标检测数据集 SODA-D 和 SODA-A 上实现了当前最先进的性能。在未来，我们认为研究一种更灵活且通用的实例质量指标将具有重要意义。

参考文献

[1] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Finding tiny faces in the wild with generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–30, 2018.

[2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European Conference on Computer Vision*, pages 206–221, 2018.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229, 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020.

[5] Gong Cheng, Junwei Han, Peicheng Zhou, and Dong Xu. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. *IEEE Transactions on Image Processing*, 28(1):265–278, 2019.

[6] Gong Cheng, Qingyang Li, Guangxing Wang, Xingxing Xie, Lingtong Min, and Junwei Han. Sfrnet: Fine-grained oriented object recognition via separate feature refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–10, 2023.

[7] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[8] Gong Cheng, Yanqing Yao, Shengyang Li, Ke Li, Xingxing Xie, Jiabao Wang, Xiwen Yao, and Junwei Han. Dual-aligned oriented detector. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

[9] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2023.

[10] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2021.

[11] Chunfang Deng, Mengmeng Wang, Liang Liu, Yong Liu, and Yunliang Jiang. Extended feature pyramid network for small object detection. *IEEE Transactions on Multimedia*, 24:1968–1979, 2021.

[12] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[13] Spyros Gidaris and Nikos Komodakis. Attend refine repeat: Active box proposal generation via in-out localization. *arXiv preprint arXiv:1606.04446*, 2016.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[15] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[19] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Robust small-scale pedestrian detection with cued recall via memory learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3050–3059, 2021.

[20] Chunbo Lang, Gong Cheng, Binfei Tu, Chao Li, and Jun-wei Han. Base and meta: A new perspective on few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2023.

[21] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[22] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1222–1230, 2017.

[23] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1829–1838, 2022.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014.

[25] Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Doll´ar. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017.

[27] Guangtao Nie and Hua Huang. Multi-oriented object detection in aerial images with double horizontal rectangles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4932–4944, 2022.

[28] Junhyug Noh, Wonho Bae, Wonhee Lee, Jinhwan Seo, and Gunhee Kim. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9725–9734, 2019.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[30] Peize Sun et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14449–14458, 2021.

[31] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9627–9636, 2019.

[32] Thang Vu, Hyunjun Jang, Trung X. Pham, and Chang Yoo. Cascade RPN: Delving into high-quality region proposal network with adaptive convolution. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2974, 2019.

[34] Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-Song Xia. Tiny object detection in aerial images. In *Proceedings of the IEEE International Conference on Pattern Recognition*, pages 3791–3798, 2021.

[35] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7303–7313, 2021.

[36] Jialian Wu, Chunluan Zhou, Qian Zhang, Ming Yang, and Junsong Yuan. Self-mimic learning for small-scale pedestrian detection. In *Proceedings of the ACM International Conference on Multimedia*, pages 2012–2020, 2020.

[37] Wei Wu, Hao Chang, Yonghua Zheng, Zhu Li, Zhiwen Chen, and Ziheng Zhang. Contrastive learning-based robust object detection under smoky conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4302, 2022.

[38] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. DetCo: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8392–8401, 2021.

[39] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3520–3529, 2021.

[40] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. RFLA: Gaussian receptive-based label assignment for tiny object detection. In *Proceedings of the European Conference on Computer Vision*, 2022.

[41] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1452–1459, 2020.

[42] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13668–13677, 2022.

[43] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. RepPoints: Point set representation for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9657–9666, 2019.

[44] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. UnitBox: An advanced object detection network. In *Proceedings of the ACM International Conference on Multimedia*, pages 516–520, 2016.

[45] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1257–1265, 2020.

[46] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020.

[47] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S3FD: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 192–201, 2017.

[48] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020.