定位蒸馏在稠密物体检测中的应用

Zhaohui Zheng[1*], Rongguang Ye[2*], Ping Wang[2], Dongwei Ren[3], Wangmeng Zuo[3],
Qibin Hou[1†], Ming-Ming Cheng[1]
[1]TMCC, CS, Nankai University [2]School of Mathematics, Tianjin University
[3]School of Computer Science and Technology, Harbin Institute of Technology

Abstract

知识蒸馏(KD)在目标检测中具有学习紧凑模型的强大能力。以往的目标检测KD由于对定位知识蒸馏效率低且改善效果小，多侧重于模仿特定区域内的深层特征，而不是模拟分类logit。本文通过对定位知识蒸馏过程的重新定义，提出了一种新的定位蒸馏方法，可以有效地将定位知识从教师转移到学生。此外，我们还启发式地引入了有价值的定位区域的概念，有助于有选择地提取特定区域的语义知识和定位知识。结合这两个新组成部分，我们首次表明logit模拟可以优于特征模仿，而定位知识蒸馏在提取目标检测器方面比语义知识更重要和更有效。我们的蒸馏方案简单而有效，可以很容易地应用于不同的稠密物体检测器。实验表明，该算法能在不影响推理速度的前提下，将单尺度1×训练计划下GFocal-ResNet-50的AP评分从COCO基准的40.1提高到42.1。我们的源代码和经过预先训练的模型可以在https://github.com/HikariTJU/LD上公开获取。

1. 引言

定位是目标检测中的一个基本问题 [15, 24, 32, 48, 49, 54–56, 60, 67]。边界框回归是迄今为止在目标检测中最流行的定位方法 [10, 31, 38, 41]，其中狄拉克 δ 分布表示是直观的，并已流行多年。 然而，定位模糊即物体不能通过其边缘确定位置仍然是一个常见的问题。如Fig. 1所示，“大象”的下边缘和“冲浪板”的右边缘是模糊定位的。这个问题对于轻量级检测器来说更加严重。缓解这一问题的一种方法是知识蒸馏(KD)，它作为一种模型压缩技术，已被广泛验证，可以通过转移大型教师网络捕获的广义知识来提高小型学生网络的性能。

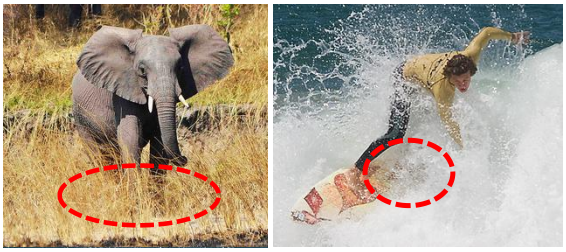

图 1. “大象”的下边缘和“冲浪板”的右边缘是模糊的。

对于KD在目标检测中的应用，已有的工作 [22, 51, 61]指出，原有的用于分类的logit模拟技术 [19]的效率较低，只转移了语义知识(如分类)，而忽略了定位知识蒸馏的重要性。因此，现有的目标检测KD方法大多侧重于加强师生对之间深度特征的一致性，并利用各个模仿区域进行蒸馏 [5, 8, 16, 25, 51]。Fig. 2展示了三种常用的用于目标检测的KD解决方案。然而，由于语义知识和定位知识在特征图上是混合的，很难判断对每个位置迁移混合知识是否有利于性能，以及哪些区域有利于哪一类型知识的迁移。

基于上述问题，本文提出了一种新的分治蒸馏策略，将语义和定位知识分别迁移，而不是简单地在特征图上蒸馏混合知识。对于语义知识，我们使用原始分类KD [19]。对于定位知识，我们重新制定了定位知识迁移过程，通过将边界框转换为概率分布，提出了一种简单而有效的定位蒸馏(LD)方法 [28, 36]。这与以

*Equal contribution.
†Corresponding author.

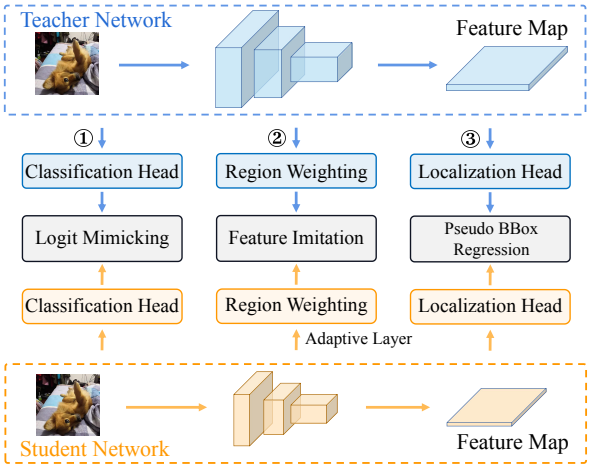

图 2. 用于目标检测的现有KD解决方案。①Logit模拟:分类KD [19]。②特征模仿:目前流行的方法基于不同的蒸馏区域提取中间特征，通常需要自适应层来对齐学生特征图的大小。③伪BBox回归:将教师预测的边界框作为附加回归目标。

往的研究 [5, 46]将教师的输出作为额外的回归目标(*i.e.*, Fig. 2中的伪BBox回归)有很大的不同。 得益于概率分布表示，我们的LD可以有效地将教师所学到的丰富的定位知识传递给学生。此外，在提出的分治蒸馏策略的基础上，我们进一步引入有价值的定位区域(VLR)，以帮助有效判断哪些区域有利于分类或定位学习。通过一系列的实验，我们首次证明了原始*logit*模拟比特征模仿更好，定位知识蒸馏比语义知识更重要和更有效。我们认为，将语义知识和定位知识根据各自的有利区域分别蒸馏，是训练更好的目标检测器的一种有前景的方法。

我们的方法简单，可以很容易地应用在任何的密集物体检测器，以提高其性能，而不引入任何推理开销。在MS COCO上的大量实验表明，在没有任何trick的情况下，我们可以用ResNet-50-FPN骨干网络将强基准GFocal [28]的AP评分从40.1提高到42.1，将AP_{75}从43.1提高到45.6。我们最好的模型使用ResNeXt-101-32x4d-DCN骨干网络可以在单尺度测试下达到50.5AP，在同样的骨干、颈部和测试设置下，它超过了所有现有的检测器。

2. 相关工作

在本节中，我们简要回顾了相关的工作，包括边界框回归、定位质量估计和知识蒸馏。

2.1. 边界框回归

边界框回归是目标检测中最常用的定位方法。R-CNN系列 [3, 34, 41, 59]采用多阶段回归细化检测结果，[2, 31, 38–40, 47]采用单阶段回归。在 [42, 57, 63, 64]中，提出了基于IoU的损失函数来提高边界框的定位质量。最近，边界框表示已经从狄拉克 δ 分布 [31, 38, 41]发展到高斯分布 [7, 18]，进一步发展到概率分布 [28, 36]。边界框的概率分布更全面地描述了边界框的不确定性，被证明是目前最先进的边界框表示。

2.2. 定位质量估计

顾名思义，定位质量估计(LQE)预测一个分数，该分数衡量检测器预测的边界框的定位质量。在训练 [27]时，通常使用LQE配合分类任务， *i.e.*，增强分类与定位的一致性。也可用于后处理过程中的联合决策 [21, 38, 47]， *i.e.*,在执行NMS时，同时考虑分类评分和LQE。早期的研究可以追溯到YOLOv1 [38]，其中预测的对象置信度用于惩罚分类得分。然后，分别提出框/掩码IoU [20, 21]和框/极中心度 [47, 52]对目标检测和实例分割的检测不确定度进行建模。从边界框表示的角度来看，Softer-NMS [18]和高斯YOLOv3 [7]对边界框的每条边进行方差预测。LQE是建模定位模糊性的一种初步方法。

2.3. 知识蒸馏

知识蒸馏 [1, 19, 33, 35, 44, 58]的目的是在优秀的教师网络指导下学习紧凑而高效的学生模型。FitNets [43]提出从教师模型的隐藏层中模拟中级隐含知识。在[5]中首次将知识蒸馏应用于目标检测，其中隐含知识学习和KD都用于多类目标检测。然后，Li等人 [25]提出模拟Faster R-CNN区域内的特征。Wang等人 [51]模拟了近锚框位置上的细粒度特征。最近，Dai等人 [8]引入了通用实例选择模块(General Instance Selection Module)来模拟师生对之间的判别块中的深层特征。DeFeat [16]在对目标区域和背景区域进行特征模仿时利用了不同的损失权重。与上述基于特征模仿的方法不同，本文引入了定位蒸馏，并提出基于有价值的定位区域分别传递分类和定位知识，使蒸馏效率更高。

3. 提出的方法

在本节中，我们将介绍本文提出的蒸馏方法。我

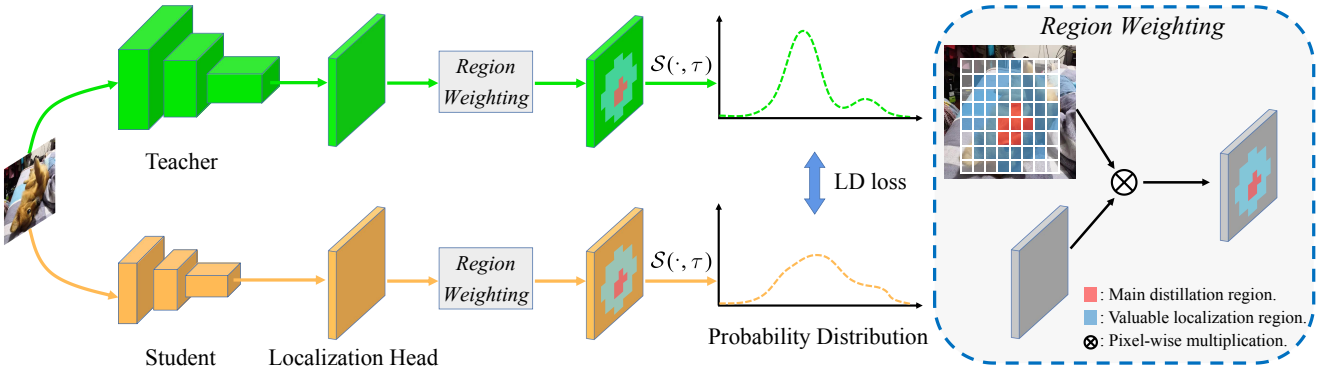

图 3. 边$e \in \mathcal{B} = \{t, b, l, r\}$的局部蒸馏(LD)说明。这里只显示了定位分支。$\mathcal{S}(\cdot, \tau)$是温度$\tau$的广义SoftMax函数。对于一个给定的检测器，我们首先将边界框表示转换为概率分布。然后，通过对主蒸馏区域和有价值的定位区域进行区域加权来确定蒸馏的位置。最后，我们计算了教师和学生预测的两个概率分布之间的LD损失。

们提出了一种新的分而治之的蒸馏策略，该策略根据各自的偏好区域分别提取语义和定位知识，而不是从特征图上提取混合知识。对于语义知识的迁移，我们简单地采用分类头上的分类KD [19]，而对于定位知识，我们提出了一种简单而有效的定位蒸馏(LD)。这两种技术都适用于单个头部的logits，而不是深层特征。然后，为了进一步提高蒸馏效率，我们引入了有价值的定位区域(VLR)，以帮助判断不同区域有益于哪类知识的迁移。在接下来的内容中，我们首先简要回顾边界框的概率分布表示，然后过渡到所提出的方法。

3.1. 基础知识

对于给定的边界框$\mathcal{B}$，常规表示有两种形式，*i.e.*, $\{x, y, w, h\}$(中心点坐标，宽度和高度) [31, 38, 41]和$\{t, b, l, r\}$(从采样点到上、下、左、右边缘的距离) [47]。这两种形式实际上遵循狄拉克 δ 分布，只关注真实物体位置，不能对边界框的模糊度建模，如Fig. 1所示。这在以往的著作中也得到了明确的证明 [7, 18, 28, 36]。

在我们的方法中，我们使用了最近的边界框概率分布表示 [28, 36]，它更全面地描述了边界框的定位不确定性。设$e \in \mathcal{B}$是边界框的一条边。

其取值一般可以表示为:

$$\hat{e} = \int_{e_{\min}}^{e_{\max}} x \Pr(x) dx, \quad e \in \mathcal{B}, \tag{1}$$

, 其中x 为区间 $[e_{\min}, e_{\max}]$的回归坐标，$\Pr(x)$为对应的概率。传统的狄拉克表示是Eqn. (1)的一个特例。其中 当 $x = e^{gt}, \Pr(x) = 1$，否则$\Pr(x) = 0$。将连续回归范围 $[e_{\min}, e_{\max}]$量化为n个子区间的均匀离散变量$\boldsymbol{e} = [e_1, e_2, \cdots, e_n]^{\mathrm{T}} \in \mathbb{R}^n$ ，其中$e_1 = e_{\min}$ and $e_n = e_{\max}$，给定边界框的每条边都可以用SoftMax函数表示为概率分布。

3.2. 定位蒸馏

在本小节中，我们介绍了定位蒸馏(LD)，一种提高蒸馏效率的新方法。我们的定位算法是从边界框的概率分布表示 [28]的观点发展而来的，它最初是为一般目标检测而设计的，具有丰富的定位信息。Fig. 1中的模糊边缘和清晰边缘将分别通过分布的平整度和锐度体现出来。

我们LD的工作原理如Fig. 3所示。给定一个任意稠密物体检测器，如 [28]，我们首先将边界框表示从四元表示切换到概率分布。我们选择$\mathcal{B} = \{t, b, l, r\}$作为边界框的基本形式。与$\{x, y, w, h\}$形式不同，$\{t, b, l, r\}$形式中每个变量的物理意义是一致的，便于我们将每条边的概率分布限制在同一区间范围内。根据 [62]，两种形式在性能上没有区别。因此，当$\{x, y, w, h\}$形式给定时，我们首先将其转换为$\{t, b, l, r\}$形式。

设$\boldsymbol{z}$为定位头预测的边e所有可能位置的n个logits，对于老师和学生分别表示为$\boldsymbol{z}_T$ 和 $\boldsymbol{z}_S$。与 [28, 36]不同的是，我们使用广义SoftMax函数$\mathcal{S}(\cdot, \tau) =$ SoftMax$(\cdot/\tau)$将$\boldsymbol{z}_T$ 和 $\boldsymbol{z}_S$转化为概率分布$\boldsymbol{p}_T$ 和 $\boldsymbol{p}_S$。注意，当$\tau = 1$时，它等价于原始的SoftMax函数。当$\tau \to 0$时，它趋向于狄拉克分布。当$\tau \to \infty$时，它将退化为均匀分布。根据经验，$\tau > 1$被设置为软化分布，使概率分布携带更多的信息。

测量两个概率 $\boldsymbol{p}_T, \boldsymbol{p}_S \in \mathbb{R}^n$之间相似度的局部蒸馏

Algorithm 1 Valuable Localization Region

Require: A set of anchor boxes $\boldsymbol{B}_l^a = \{\mathcal{B}_{i_l}^a\}$ and a set of ground truth boxes $\boldsymbol{B}^{gt} = \{\mathcal{B}_j^a\}$, $1 \leqslant i_l \leqslant I_l$, $1 \leqslant j \leqslant J$, $I_l = W_l \times H_l$. Positive threshold α_{pos} of label assignment. W_l and H_l are the sizes of l-th FPN level.

Ensure: $\boldsymbol{V}_l = \{v_{i_l j}\}_{I_l \times J}, v_{i_l j} \in \{0, 1\}$ encodes final location of VLR, where 1 denotes VLR and 0 indicates ignore.

1: Compute DIoU matrix $\boldsymbol{X}_l = \{x_{i_l j}\}_{I_l \times J}$ with $x_{i_l j} = DIoU(\mathcal{B}_{i_l}^a, \mathcal{B}_j^{gt})$.
2: $\alpha_{vl} = \gamma \alpha_{pos}$.
3: Select locations with $\boldsymbol{V}_l = \{\alpha_{vl} \leqslant \boldsymbol{X}_l \leqslant \alpha_{pos}\}$.
4: **return** $\boldsymbol{V}_l$

得到:

$$\mathcal{L}_{\text{LD}}^e = \mathcal{L}_{\text{KL}}(\boldsymbol{p}_S^\tau, \boldsymbol{p}_T^\tau) \tag{2}$$

$$= \mathcal{L}_{\text{KL}}\left(\mathcal{S}(\boldsymbol{z}_S, \tau), \mathcal{S}(\boldsymbol{z}_T, \tau)\right), \tag{3}$$

其中$\mathcal{L}_{\text{KL}}$表示KL-Divergence损失。那么，边界框$\mathcal{B}$所有四条边的LD可表示为:

$$\mathcal{L}_{\text{LD}}(\mathcal{B}_S, \mathcal{B}_T) = \sum_{e \in \mathcal{B}} \mathcal{L}_{\text{LD}}^e. \tag{4}$$

本文首次尝试采用logit模拟方法提取定位知识用于目标检测。虽然框的概率分布表示已被证明在通用目标检测任务 [28]中有用，但还没有人探索它在定位知识蒸馏中的性能。我们将框的概率分布表示和KL-Divergence损失相结合，证明了这种简单的logit模拟技术在提高目标检测器的蒸馏效率方面有很好的效果。这也使得我们的LD与以往的相关工作有所不同，相反，我们的LD强调特征模仿的重要性。在我们的实验部分，我们将展示更多的数值分析提出的LD的优势。

3.3. 有价值的定位区域

以往的工作大多是通过最小化l_2损失来促使学生模仿老师的深层特征。然而，一个直接的问题应该是:我们是否应该毫无区别地使用整个模仿区域来提取混合知识?根据我们的观察，答案是否定的。既往工作 [11, 13, 26, 45, 50]指出了分类和定位知识分布模式不同。因此，在本小节中，我们描述了有价值的定位区域(VLR)，以进一步提高蒸馏效率，我们相信这将是一个训练更好的学生检测器的有前景的方法。

具体来说，蒸馏分为两个部分，主要蒸馏区和有价值的定位区。通过标签分配直观地确定了主要蒸馏区域，*i.e.*,检测头的正位置。算法 1可以得到有价值的定位区域。首先，对于第l个FPN级别，我们计算所有锚框$\boldsymbol{B}_l^a$和真实物体$\boldsymbol{B}^{gt}$之间的DIoU [63]矩阵$\boldsymbol{X}_l$。设DIoU的下界为$\alpha_{vl} = \gamma \alpha_{pos}$，其中$\alpha_{pos}$为标签分配的正IoU阈值。VLR可以定义为$\boldsymbol{V}_l = \{\alpha_{vl} \leqslant \boldsymbol{X_l} \leqslant \alpha_{pos}\}$。我们的方法只有一个超参数$\gamma$，它控制着VLRs的范围。当$\gamma = 0$时，凡是在预设锚框与GT框之间的DIoUs满足$0 \leqslant x_{i_l j} \leqslant \alpha_{pos}$的位置都被确定为VLRs。当$\gamma \to 1$时，VLR将逐渐收缩为空。这里我们使用DIoU [63]，因为它给予靠近物体中心的位置更高的优先级。

与标签分配类似，我们的方法跨多级FPN为每个位置分配属性。这样的话，GT框以外的一些地点也会被考虑在内。所以，我们实际上可以把VLR看作是主蒸馏区向外的延伸。注意，对于没有锚点的检测器，比如FCOS，我们可以在特征图上使用预设锚点，并且不改变它的回归形式，从而使定位学习保持为无锚点类型。而对于基于锚点的检测器，通常在每个位置设置多个锚点，我们展开锚框计算DIoU矩阵，然后分配它们的属性。

3.4. 整个蒸馏过程

训练学生$\boldsymbol{S}$的总体损失可以表示为:

$$\begin{aligned}\mathcal{L} =& \lambda_0 \mathcal{L}_{\text{cls}}(\mathcal{C}_S, \mathcal{C}^{gt}) + \lambda_1 \mathcal{L}_{\text{reg}}(\mathcal{B}_S, \mathcal{B}^{gt}) + \lambda_2 \mathcal{L}_{\text{DFL}}(\mathcal{B}_S, \mathcal{B}^{gt}) \\ &+ \lambda_3 \mathbb{I}_{\text{Main}} \mathcal{L}_{\text{LD}}(\mathcal{B}_S, \mathcal{B}_T) + \lambda_4 \mathbb{I}_{\text{VL}} \mathcal{L}_{\text{LD}}(\mathcal{B}_S, \mathcal{B}_T) \\ &+ \lambda_5 \mathbb{I}_{\text{Main}} \mathcal{L}_{\text{KD}}(\mathcal{C}_S, \mathcal{C}_T) + \lambda_6 \mathbb{I}_{\text{VL}} \mathcal{L}_{\text{KD}}(\mathcal{C}_S, \mathcal{C}_T),\end{aligned} \tag{5}$$

前三项对任何基于回归的检测器的分类和边界框回归分支是完全相同的,*i.e.*, $\mathcal{L}_{\text{cls}}$是分类损失,$\mathcal{L}_{\text{reg}}$是边界框回归损失,$\mathcal{L}_{\text{DFL}}$是分布focal loss [28]。$\mathbb{I}_{\text{Main}}$ 和 $\mathbb{I}_{\text{VL}}$分别为主要蒸馏区域和有价值定位区域的蒸馏掩码，$\mathcal{L}_{\text{KD}}$为KD损失 [19]，$\mathcal{C}_S$ 和 $\mathcal{C}_T$分别表示学生和教师的分类头输出的logits，$\mathcal{C}^{gt}$为真实物体类标签。所有的蒸馏损失将根据其类型采用相同的权重因子进行加权，*e.g.*,LD损失遵循bbox回归法，KD损失遵循分类法。值得一提的是，DFL损失项可以被禁用，因为LD损失具有足够的引导能力。此外，我们可以启用或禁用四种类型的蒸馏损失，以便以单独的蒸馏区域方式对学生进行蒸馏。

τ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
–	40.1	58.2	43.1	23.3	44.4	52.5
1	40.3	58.2	43.4	22.4	44.0	52.4
5	40.9	58.2	44.3	23.2	**45.0**	53.2
10	**41.1**	**58.7**	**44.9**	**23.8**	44.9	**53.6**
15	40.7	58.5	44.2	23.5	44.3	53.3
20	40.5	58.3	43.7	**23.8**	44.1	53.5

(a) **LD中的温度**τ:大τ的广义Softmax函数带来相当大的增益。我们默认设置 τ=10。老师是ResNet-101，学生是ResNet-50。

ε	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
–	40.1	58.2	43.1	23.3	44.4	52.5
0.1	40.5	58.3	43.8	23.0	44.2	52.7
0.2	40.2	58.2	43.6	23.1	44.0	53.0
0.3	40.1	58.4	43.1	23.6	43.9	52.5
0.4	40.3	58.4	43.4	22.8	44.0	52.6
LD	**41.1**	**58.7**	**44.9**	**23.8**	**44.9**	**53.6**

(b) **LD *vs.* 伪BBox回归 [5]**:我们的LD可以更有效地传递定位知识，老师是ResNet-101，学生是ResNet-50。

γ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
–	40.1	58.2	43.1	23.3	44.4	52.5
1	41.1	58.7	44.9	23.8	44.9	53.6
0.75	41.2	58.8	44.9	23.6	45.4	53.5
0.5	41.7	59.4	45.3	24.2	45.6	54.2
0.25	**41.8**	**59.5**	**45.4**	24.2	45.8	**54.9**
0	41.7	**59.5**	**45.4**	**24.5**	**45.9**	54.0

(c) **γ在VLR中的作用**:在有价值的局部化区域进行LD对性能有积极影响。我们默认设置γ=0.25。老师是ResNet-101，学生是ResNet-50。

表 1. **消融实验**。我们在MS COCO val2017上进行了LD和VLR消融实验。

4. 实验

在本节中，我们进行了全面的消融研究和分析以证明提出的LD和蒸馏方案在具有挑战性的大规模MS COCO[31]基准上的优越性。

4.1. 实验设置

train2017 (118K图像)用于训练，val2017 (5K图像)用于验证。我们还通过提交到COCO服务器获得MS COCO test-dev 2019 (20K图像)的评估结果。实验在mmDetection [6]框架下进行。除非另有说明，我们使用ResNet [17]和FPN [29]作为我们的骨干和颈部网络，以及FCOS风格的 [47]无锚头用于分类和定位。消融实验训练计划设置为单尺度1×模式(12代)。对于其他训练和测试超参数，我们完全遵循GFocal [28]协议，包括用于分类的QFL损失和用于bbox回归的GIoU损失等。我们使用标准COCO风格的测量方法，*i.e.*,平均精度(AP)进行评估。所有基准模型都采用相同的设置进行重新训练，以便与我们的LD进行比较。更多的实现细节和更多关于PASCAL VOC [9]的实验结果可以在补充材料中找到。

4.2. 消融实验与分析

在LD里的温度τ。我们的LD引入了一个超参数，即温度τ。表1a报告了不同温度下LD的结果，其中教师模型为AP 44.7的ResNet-101，学生模型为ResNet-50。这里只采用主蒸馏区。与表1a中的第一行相比，不同的温度始终能产生更好的结果。在本文中，我们简单地将LD中的温度设为$\tau = 10$，这在其他所有实验中都是固定的。

LD *vs.*伪BBox回归。教师有界回归(TBR)损失 [5]是在定位头上增强学生的初步尝试，即Fig. 2中的伪bbox回归，其表示为:

$$\mathcal{L}_{\text{TBR}} = \lambda \mathcal{L}_{\text{reg}}(\mathcal{B}^s, \mathcal{B}^{gt}), \text{if } \ell_2(\mathcal{B}^s, \mathcal{B}^{gt}) + \varepsilon > \ell_2(\mathcal{B}^t, \mathcal{B}^{gt}), \tag{6}$$

其中$\mathcal{B}^s$ 和 $\mathcal{B}^t$分别表示学生和教师的预测框，$\mathcal{B}^{gt}$表示真实物体框，ε 是预定义的边界，$\mathcal{L}_{\text{reg}}$表示GIoU损失[42]。这里只采用主蒸馏区。从表1b可以看出，在Eqn. (6)中取适当的阈值$\varepsilon = 0.1$ 时，TBR损失确实会带来性能收益(+0.4 AP和+ 0.7 AP_{75})，但采用粗糙的框表示，不包含检测器的任何定位模糊性信息，导致次优结果。相反，我们的LD直接生产了41.1AP和44.9 AP_{75}，因为它利用了bbox的概率分布，包含丰富的定位知识。

VLR中的各种γ。新引入的VLR具有控制VLR范围的参数γ。如表1c所示，当γ在0到0.5范围内时，AP是稳定的。AP在这一范围内的变化约为0.1。随着γ的增加，VLR逐渐缩小至空。性能也逐渐下降到41.1，即只在主蒸馏区进行LD。对参数γ的灵敏度分析实验表明，在VLR上进行LD对性能有积极的影响。在其余的实验中，为了简单起见，我们将γ设为0.25。

分离蒸馏区方式。关于KD和LD及其偏好区域的作用有一些有趣的观察。我们在表2中报告了相关的消融实验结果，其中“Main”表示logit模拟在主要蒸馏区域，*i.e.*,标签分配的阳性位置，“VLR”表示有价值的定位区域。可以看出，进行“Main KD”、“Main LD”以及两者的结合都可以提高学生的成绩+0.1、+1.0、+1.3 AP。这说明蒸馏的主要区域包含对分类和定位有价值的知识，分类KD的收益低于LD。因此，我们对更大的范围，即VLR进行蒸馏。我们可以看到，“VLR LD”(表2的第5行)可以在“Main LD”(第3行)的基础上

表 2. KD和我们LD的**分离蒸馏区域方式**的比较，老师是ResNet-101，学生是ResNet-50。“Main”表示主要蒸馏区域，即标签分配的正位置。“VLR”表示有价值的定位区域。在MS COCO val2017上进行了结果的展示。

Main KD	Main LD	VLR KD	VLR LD	AP	AP_{50}	AP_{75}
				40.1	58.2	43.1
✓				40.2	58.6	43.4
	✓			41.1	58.7	44.9
✓	✓			41.4	59.2	45.0
✓		✓		40.4	58.9	43.4
	✓		✓	41.8	59.5	45.4
✓	✓		✓	**42.1**	**60.3**	**45.6**
✓	✓	✓	✓	42.0	60.0	45.4

表 3. **Logit模拟*vs*.特征模仿**。“Ours”是指我们采用单独的蒸馏区域方式，即在主蒸馏区域进行KD和LD，在VLR上进行LD。老师是ResNet-101，学生是ResNet50 [17]，在MS COCO val2017上进行了结果的展示。

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline (GFocal [28])	40.1	58.2	43.1	23.3	44.4	52.5
FitNets [43]	40.7	58.6	44.0	23.7	44.4	53.2
Inside GT Box	40.7	58.6	44.2	23.1	44.5	53.5
Main Region	41.1	58.7	44.4	24.1	44.6	53.6
Fine-Grained [51]	41.1	58.8	44.8	23.3	45.4	53.1
DeFeat [16]	40.8	58.6	44.2	24.3	44.6	53.7
GI Imitation [8]	41.5	59.6	45.2	24.3	45.7	53.6
Ours	42.1	60.3	45.6	24.5	46.2	54.8
Ours + FitNets	42.1	59.9	45.7	**25.0**	46.3	54.4
Ours + Inside GT Box	42.2	60.0	45.9	24.3	46.3	55.0
Ours + Main Region	42.1	60.0	45.7	24.6	46.3	54.7
Ours + Fine-Grained	**42.4**	**60.3**	45.9	24.7	46.5	**55.4**
Ours + DeFeat	42.2	60.0	45.8	24.7	46.1	54.4
Ours + GI Imitation	**42.4**	**60.3**	**46.2**	**25.0**	**46.6**	54.5

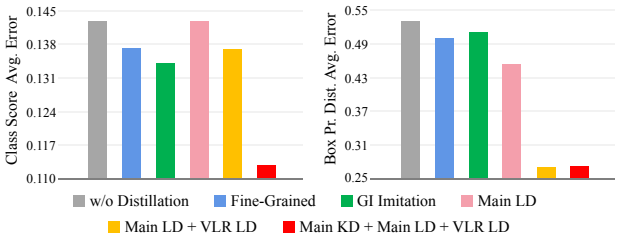

图 4. SOTA特征模仿和我们的LD的可视化比较。我们展示了P4、P5、P6和P7 FPN水平上教师和学生之间分类得分和框概率分布的平均L1误差。老师是ResNet-101，学生是ResNet-50。在MS COCO val2017上对结果进行了测评。

进一步提高0.7 AP。然而，我们观察到，涉及“VLR KD”的改进有限(表2的第2行和第5行)，甚至没有改进(表2的最后两行)。这再次表明，定位知识蒸馏比语义知识蒸馏更重要和有效，我们的分治蒸馏方案，即“主KD”+“主LD”+“VLR LD”，是对VLR的补充。

Logit模拟*vs*.特征模仿。我们将提出的LD与几种最先进的特征模仿方法进行了比较。我们采用分离蒸馏区方式，即在主蒸馏区进行KD和LD，在VLR上进行LD。由于现在的检测器通常具有FPN [29]结构，在前人的工作 [8, 16, 51]的基础上，我们重新实现了他们的方法，并将所有的特征模仿加到多级FPN上，以进行公平的比较。在这里，“FitNets” [43]蒸馏了整个特征图。“DeFeat” [16]表示GT框外的特征模仿损失权重大于GT框内的。“Fine-Grained” [51]蒸馏了接近锚框位置的深层特征。“GI Imitation” [8]根据学生和老师的判别预测选择蒸馏区域。“Inside GT Box”意味着我们在FPN层上使用与特征模仿区域相同步幅的GT框。“Main Region”指的是我们模拟主蒸馏区域内的特征。

从表3中我们可以看到，整个特征图内部的蒸馏达到+0.6AP的增益。通过在GT框外设置一个更大的损失(DeFeat [16])，性能略好于在所有位置使用相同的损失权重。Fine-Grained [51]集中在GT框附近的位置，产生41.1AP，其结果与使用主区域的特征模拟结果相当。GI imitation [8]搜索特征模仿的判别块，得到41.5 AP。由于学生和教师的预测差距很大，模仿区域可能出现在任何地方。

尽管这些特征模仿方法有了显著的改进，但它们没有明确考虑知识分布模式。相反，我们的方法可以通过一个单独的蒸馏区域的方式迁移知识，直接产生42.1 AP。值得注意的是，我们的方法操作的是logits而不是特征，这表明只要采用适当的蒸馏策略，logit模拟并不亚于特征模仿，就像我们的LD。此外，我们的方法与上述特征模仿方法正交。表 3显示，使用这些特征模仿方法，我们的性能可以进一步提高。特别是，通过GI imitation，我们将强基准GFocal提高了2.3 AP和3.1AP_{75}。

我们进一步进行了实验，检验分类得分的平均误差和框概率分布，如Fig. 4所示。可以看到，Fine-Grained特征模仿 [51]和GI imitation [8]如预期的那样减少了两个错误，因为语义知识和定位知识混合在特征图上。我们的“Main LD”和“Main LD + VLR LD”的分类得分平均误差与Fine-Grained [51]和GI imitation [8]的分类得分平均误差相当或更大，但框概率分布平均误差较低。这表明这两个只有LD的设置可以显著减小教师与学生之间的框概率分布距离，而不能减小分类头的误差是合理的。如果将分类KD加到主蒸

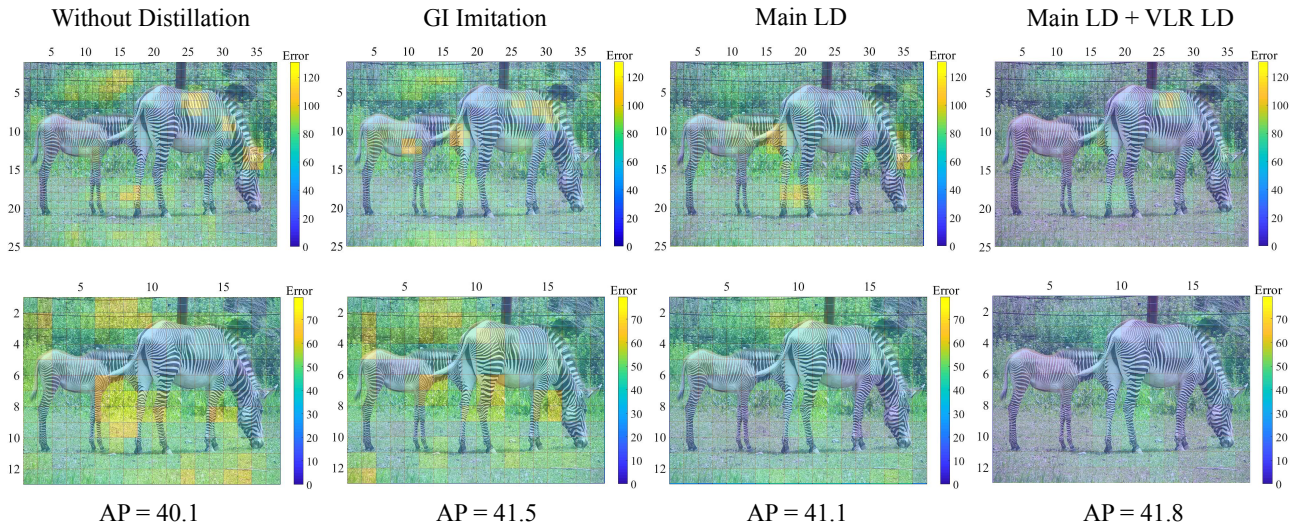

图 5. 最先进的特征模仿和我们的LD之间的可视化比较。我们展示了在P5(第一行)和P6(第二行)FPN级别上，教师和学生之间的定位头部logits的每个位置L1误差总和。老师是ResNet-101，学生是ResNet-50。我们可以看到，与GI imitation [8]相比，我们的方法(主LD + VLR LD)可以显著降低几乎所有位置的误差。越黑越好。最好用彩色观看。

表 4. 轻量化检测器LD的定量结果。老师是ResNet-101。在MS COCO val2017上进行了结果的测评。

Student	LD	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-18		35.8	53.1	38.2	18.9	38.9	47.9
	✓	37.5	54.7	40.4	20.2	41.2	49.4
ResNet-34		38.9	56.6	42.2	21.5	42.8	51.4
	✓	41.0	58.6	44.6	23.2	45.0	54.2
ResNet-50		40.1	58.2	43.1	23.3	44.4	52.5
	✓	42.1	60.3	45.6	24.5	46.2	54.8

表 5. LD在各种常用稠密物体检测器上的定量结果。老师是ResNet-101，学生是ResNet50。在MS COCO val2017上进行了结果的测评。

Student	LD	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [30]		36.9	54.3	39.8	21.2	40.8	48.4
	✓	39.0	56.4	42.4	23.1	43.2	51.1
FCOS [47]		38.6	57.2	41.5	22.4	42.2	49.8
	✓	40.6	58.4	44.1	24.3	44.1	52.3
ATSS [62]		39.2	57.3	42.4	22.7	43.1	51.5
	✓	41.6	59.3	45.3	25.2	45.2	53.3

馏区，得到“主KD +主LD + VLR LD”，分类得分平均误差和框概率分布平均误差均可降低。我们还可视化了在P5和P6 FPN级别的每个位置上，学生和老师之间的定位头部logits的L1误差总和。在Fig. 5中，与“无蒸馏”相比，我们可以看到GI imitation [8]确实减少了老师和学生之间的定位差异。注意，我们特别选择了一个模型(主LD + VLR LD)进行可视化，其AP性能略优于GI imitation。我们的方法可以更明显地减少这一误差，并减轻定位模糊性。

轻量级检测器的LD。接下来，我们用分离蒸馏区域的方式验证我们的LD，即主KD +主LD + VLR LD，用于轻量级检测器。我们选择由mmDetection [6]提供的AP44.7的ResNet-101作为老师蒸馏出一系列轻量级的学生。如表 4所示，我们的LD能够稳定地提高学生ResNet-18、ResNet-34、ResNet-50的AP水平，分别为+1.7、+2.1、+2.0,AP_{75}水平分别为+2.2、+2.4、+2.4。从这些结果中，我们可以得出结论，我们的LD可以稳定地提高所有学生的定位精度。

扩展到其他稠密物体检测器。我们的LD可以灵活地合并到其他稠密物体检测器中，无论是基于锚的还是无锚的类型。我们将LD与分离蒸馏区域方式应用于最近流行的几种检测器，如RetinaNet [30](基于锚)、FCOS [47](无锚)和ATSS [62](基于锚)。根据表5的结果，LD对于这些致密探测器可以持续提高~2AP。

4.3. 与最先进方法的比较

我们将我们的LD与最先进的稠密物体检测器进行比较，使用我们的LD进一步增强GFocalV2 [27]。对于COCO val2017，由于之前的大多数工作都使用ResNet-50-FPN骨干和单尺度1×训练计划(12代)进行验证，我们也展现了在这种设置下的结果，以进行公平的比较。对于COCO test-dev 2019，在之前的工作 [27]之后，包含了1333 × [480 : 960]多尺度2×训练计划(24代)的LD模型。训练在一个有8个GPU的机

器节点上进行，每个GPU的批处理大小为2，初始学习率为0.01，以便进行公平的比较。在推理过程中，采用单尺度测试([1333 × 800]分辨率)。对于不同的学生ResNet-50、ResNet-101和ResNeXt-101-32x4d-DCN [53, 68]，我们也分别选择不同的网络ResNet-101、ResNet-101- DCN和Res2Net101-DCN [12]作为他们的老师。

表 6. 与COCO *val2017*和*test-dev2019*上最先进的方法进行比较。**TS**:训练计划。“1×”:单尺度训练12代。“2×”:多尺度训练24代。

Method	TS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50 backbone on val2017							
RetinaNet [30]	1×	36.9	54.3	39.8	21.2	40.8	48.4
FCOS [47]	1×	38.6	57.2	41.5	22.4	42.2	49.8
SAPD [66]	1×	38.8	58.7	41.3	22.5	42.6	50.8
ATSS [62]	1×	39.2	57.3	42.4	22.7	43.1	51.5
BorderDet [37]	1×	41.4	59.4	44.5	23.6	45.1	54.6
AutoAssign [65]	1×	40.5	59.8	43.9	23.1	44.7	52.9
PAA [23]	1×	40.4	58.4	43.9	22.9	44.3	54.0
OTA [14]	1×	40.7	58.4	44.3	23.2	45.0	53.6
GFocal [28]	1×	40.1	58.2	43.1	23.3	44.4	52.5
GFocalV2 [27]	1×	41.1	58.8	44.9	23.5	44.9	53.3
LD (ours)	1×	**42.7**	**60.2**	**46.7**	**25.0**	**46.4**	**55.1**
ResNet-101 backbone on test-dev 2019							
RetinaNet [30]	2×	39.1	59.1	42.3	21.8	42.7	50.2
FCOS [47]	2×	41.5	60.7	45.0	24.4	44.8	51.6
SAPD [66]	2×	43.5	63.6	46.5	24.9	46.8	54.6
ATSS [62]	2×	43.6	62.1	47.4	26.1	47.0	53.6
BorderDet [37]	2×	45.4	64.1	48.8	26.7	48.3	56.5
AutoAssign [65]	2×	44.5	64.3	48.4	25.9	47.4	55.0
PAA [23]	2×	44.8	63.3	48.7	26.5	48.8	56.3
OTA [14]	2×	45.3	63.5	49.3	26.9	48.8	56.1
GFocal [28]	2×	45.0	63.7	48.9	27.2	48.8	54.5
GFocalV2 [27]	2×	46.0	64.1	50.2	27.6	49.6	56.5
LD (ours)	2×	**47.1**	**65.0**	**51.4**	**28.3**	**50.9**	**58.5**
ResNeXt-101-32x4d-DCN backbone on test-dev 2019							
SAPD [66]	2×	46.6	66.6	50.0	27.3	49.7	60.7
GFocal [28]	2×	48.2	67.4	52.6	29.2	51.7	60.2
GFocalV2 [27]	2×	49.0	67.6	53.4	29.8	52.3	61.8
LD (ours)	2×	**50.5**	**69.0**	**55.3**	**30.9**	**54.4**	**63.4**

如表 6所示，在使用ResNet-50-FPN骨干时，我们的LD将SOTA GFocalV2的AP评分提高了1.6，将AP_{75}评分提高了1.8。当使用ResNet-101-FPN和ResNext-101-32x4d-DCN进行多尺度2×训练时，我们获得了最高的AP得分，47.1和50.5，优于所有现有的稠密物体检测器在相同的主干、颈部和测试设置下的表现。更重要的是，我们的LD不会引入任何额外的网络参数或计算开销，因此可以保证与GFocalV2完全相同的推理速度。

5. 结论

本文提出了一种用于稠密物体检测的灵活定位蒸馏方法，并设计了一个有价值的定位区域，以独立蒸馏区域的方式对学生检测器进行蒸馏。结果表明:1)logit模拟优于特征模仿;2)在蒸馏目标检测器时，分离迁移分类和定位知识的蒸馏区域方式很重要。我们希望我们的方法可以为目标检测领域提供新的研究直觉并制定更好的蒸馏策略。此外，LD在稀疏目标检测器(DETR [4]系列)和其他相关领域的应用，如实例分割、目标跟踪和三维目标检测，值得进一步研究。

致谢 本研究由国家重点研究与发展计划(NO.2018AAA0100400)和国家自然科学基金资助项目(NO.61922046，NO.62172127)资助。

参考文献

[1] Ji-Hoon Bae, Doyeob Yeo, Junho Yim, Nae-Soo Kim, Cheol-Sig Pyo, and Junmo Kim. Densely distilled flow-based knowledge transfer in teacher-student framework for image classification. *IEEE Transactions on Image Processing*, 29:5698–5710, 2020.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NeurIPS*, 2017.

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[7] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian YOLOv3: An accurate and fast object de-

tector using localization uncertainty for autonomous driving. In *ICCV*, 2019.

[8] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *CVPR*, 2021.

[9] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2009.

[11] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, 2021.

[12] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2021.

[13] Ziteng Gao, Limin Wang, and Gangshan Wu. Mutual supervision for dense object detection. In *ICCV*, 2021.

[14] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. OTA: Optimal transport assignment for object detection. In *CVPR*, 2021.

[15] Spyros Gidaris and Nikos Komodakis. Locnet: Improving localization accuracy for object detection. In *CVPR*, 2016.

[16] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *CVPR*, 2021.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[18] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *CVPR*, 2019.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring R-CNN. In *CVPR*, 2019.

[21] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, 2018.

[22] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. In *NeurIPS*, 2021.

[23] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, 2020.

[24] Tao Kong, Fuchun Sun, Chuanqi Tan, Huaping Liu, and Wenbing Huang. Deep feature pyramid reconfiguration for object detection. In *ECCV*, 2018.

[25] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *CVPR*, 2017.

[26] Wuyang Li, Zhen Chen, Baopu Li, Dingwen Zhang, and Yixuan Yuan. Htd: Heterogeneous task decoupling for two-stage object detection. *IEEE Transactions on Image Processing*, 30:9456–9469, 2021.

[27] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *CVPR*, 2021.

[28] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized Focal Loss: learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, 2020.

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[32] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid R-CNN. In *CVPR*, 2019.

[33] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, 2020.

[34] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra R-CNN: Towards balanced learning for object detection. In *CVPR*, 2019.

[35] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

[36] Heqian Qiu, Hongliang Li, Qingbo Wu, and Hengcan Shi. Offset bin classification network for accurate object detection. In *CVPR*, 2020.

[37] Han Qiu, Yuchen Ma, Zeming Li, Songtao Liu, and Jian Sun. Borderdet: Border feature for dense object detection. In *ECCV*, 2020.

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[39] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017.

[40] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[42] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized Intersection over Union: A metric and a loss for bounding box regression. In *CVPR*, 2019.

[43] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

[44] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *ICCV*, 2021.

[45] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *CVPR*, 2020.

[46] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *arXiv preprint arXiv:2006.13108*, 2020.

[47] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019.

[48] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *CVPR*, 2019.

[49] Jiaqi Wang, Wenwei Zhang, Yuhang Cao, Kai Chen, Jiangmiao Pang, Tao Gong, Jianping Shi, Chen Change Loy, and Dahua Lin. Side-aware boundary localization for more precise object detection. In *ECCV*, 2020.

[50] Keyang Wang and Lei Zhang. Reconcile prediction consistency for balanced object detection. In *ICCV*, 2021.

[51] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019.

[52] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020.

[53] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[54] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *ICML*, 2021.

[55] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *ICCV*, 2019.

[56] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. In *NeurIPS*, 2021.

[57] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. UnitBox: an advanced object detection network. In *ACM MM*, 2016.

[58] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[59] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *ECCV*, 2020.

[60] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, 2021.

[61] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *ICLR*, 2020.

[62] 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 *CVPR*, 2020.

[63] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU Loss: Faster and better learning for bounding box regression. In *AAAI*, 2020.

[64] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. Enhancing ge-

ometric factors in model learning and inference for object detection and instance segmentation. *IEEE Transactions on Cybernetics*, 2021.

[65] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020.

[66] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides. Soft anchor-point object detection. In *CVPR*, 2020.

[67] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *CVPR*, 2019.

[68] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019.