CrossKD: 用于目标检测的交叉头式知识蒸馏

王家宝 [1*], 陈宇铭 [1*], 郑兆辉 [1], 李翔 [2,1], 程明明 [2,1], 侯淇彬 [2,1†]

[1] 媒体计算实验室, 计算机学院, 南开大学
[2] 南开国际先进研究院, 深圳福田

https://github.com/jbwang1997/CrossKD

摘要

知识蒸馏（*KD*）已被验证为一种有效的模型压缩技术，用于密集目标检测器的学习。现有的最先进目标检测知识蒸馏方法大多基于特征模仿。在本文中，我们提出了一种通用且有效的预测模仿蒸馏方案，称为 *CrossKD*，它将学生检测头部的中间特征传递给教师的检测头部。然后，产生的交叉头部预测被强制模仿教师的预测。这种方式减轻了学生检测头部接收来自标签和教师预测的矛盾监督信号的负担，极大地提高了学生的目标检测性能。此外，由于模仿教师的预测是知识蒸馏的目标，*CrossKD* 与特征模仿相比，提供了更多面向任务的信息。在 *MS COCO* 上，仅应用预测模仿损失，我们的 *CrossKD* 将 *GFL ResNet-50* 的平均精度从 *40.2* 提高到 *43.7*，超过了所有现有的 *KD* 方法。此外，我们的方法在蒸馏具有异构骨架的检测器时也表现良好。

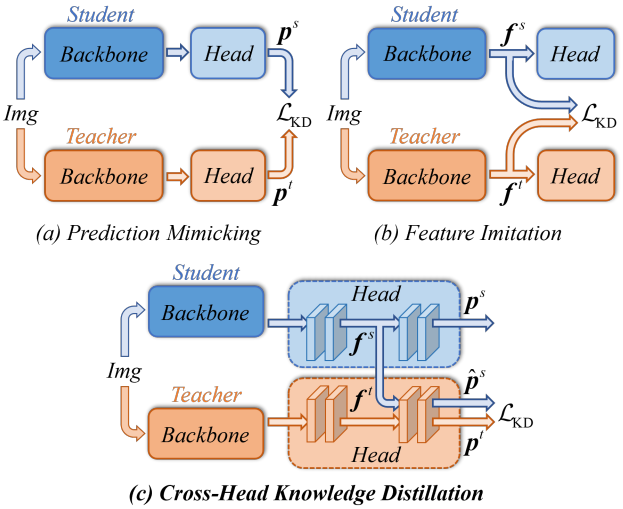

图 1. 传统知识蒸馏方法与我们的 CrossKD 之间的比较。与明确强制教师-学生对之间的中间特征图或预测的一致性不同，CrossKD 隐式地建立教师-学生对头部之间的联系，以提高蒸馏效率。

1. 引言

知识蒸馏(KD)，作为一种模型压缩技术，在目标检测领域已经被深入研究 [5, 13, 29, 31, 55, 59, 60, 73, 74]，并且最近取得了卓越的性能。根据检测器的蒸馏位置，现有的知识蒸馏方法大致可以分为两类：预测模仿和特征模仿。预测模仿（见图 Fig. 1(a)）最早在 [24] 中提出，它指出教师预测的平滑分布比真实标注的狄拉克分布更适合学生学习。换句话说，预测模仿迫使学生模仿教师的预测分布。不同地，特征模仿（见图 Fig. 1(b)）遵循 FitNet [53] 中提出的观点，该观点认为中间特征比教师的预测包含更多信息。它旨在强制教师-学生对之间的特征一致性。

预测模仿在蒸馏目标检测模型中扮演着至关重要的角色。然而，长期以来，人们观察到它不如特征模仿高效。最近，郑等人 [73] 提出了一种定位蒸馏（LD）方法，通过转移定位知识来改进预测模仿，这将预测模仿提升到了一个新的水平。尽管仍然落后于先进的特征模仿方法，例如 PKD [5]，LD 展示了预测模仿过程有能力传递特定任务的知识，这从与特征模仿正交的角

*相同贡献。
†通讯作者。

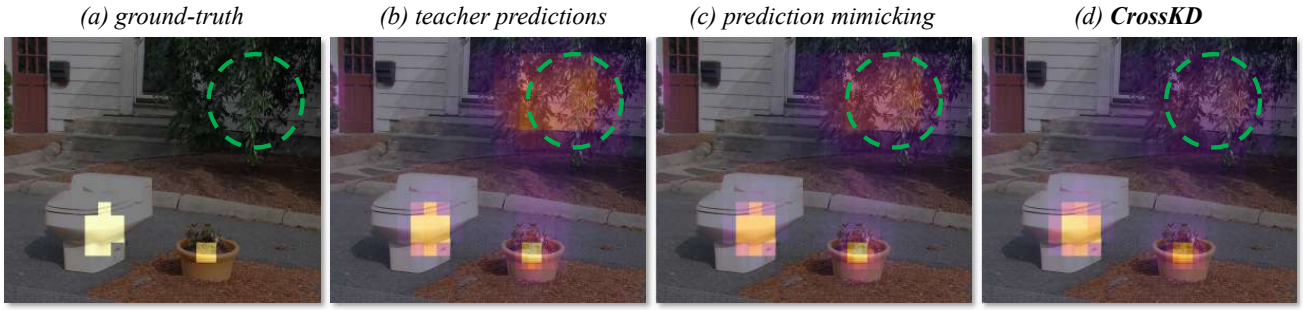

图 2. GFL [35] 分类预测的可视化。(a) 和 (b) 是真实标注和蒸馏目标。(c) 和 (d) 是使用传统预测模仿和提出的 CrossKD 训练的模型预测的分类输出。在绿色圆圈区域，教师预测的蒸馏目标与分配给学生的真值目标存在很大差异。预测模仿迫使学生模仿教师，而 CrossKD 可以平滑模仿过程。

度为学生模型带来了好处。这激励我们进一步探索和改进预测模仿。

通过调查，我们观察到传统的预测模仿可能会因为学生分配器的真实标注目标和教师预测的蒸馏目标之间的冲突而受到影响。在使用预测模仿训练检测器时，学生模型的预测被迫同时模仿真实标注目标和教师的预测。然而，教师预测的蒸馏目标通常与分配给学生的真实标注目标存在很大差异。如图 Fig. 2(a) 和 Fig. 2(b) 所示，教师在绿色圆圈区域产生了类别概率，这与分配给学生的真实标注目标相冲突。因此，学生检测器在蒸馏过程中经历了一个矛盾的学习过程，这严重干扰了优化。

为了缓解上述冲突，先前的预测模仿方法 [13, 19, 73] 倾向于在包含中等教师-学生差异的区域内进行蒸馏。然而，我们认为高度不确定的区域通常包含更多对学生有益的信息。在本文中，我们提出了一种新颖的交叉头部知识蒸馏流程，简称为 *CrossKD*。如 Fig. 1(c) 所示，我们提议将学生头部的中间特征输入到教师头部，产生交叉头部预测。然后，可以在新的交叉头部预测和教师的预测之间进行知识蒸馏操作。

尽管 CrossKD 简单，但它提供了以下两个主要优势。首先，由于交叉头部预测和教师的预测都是通过共享教师检测头部的一部分产生的，交叉头部预测与教师的预测相对一致。这减轻了教师-学生对之间的差异，并增强了预测模仿的训练稳定性。此外，由于模仿教师的预测是知识蒸馏的目标，CrossKD 在理论上是最优的，并且与特征模仿相比，可以提供更多面向任务的信息。这两个优势使我们的 CrossKD 能够高效地从教师的预测中蒸馏知识，因此比以往的 state-of-the-art 特征模仿方法表现得更好。

无需任何花哨的装饰，我们的方法可以显著提升学生检测器的性能，实现更快的训练收敛。本文在 COCO [40] 数据集上进行了全面的实验，以阐述 CrossKD 的有效性。具体来说，仅应用预测模仿损失，CrossKD 在 1× 训练计划下的 GFL 上达到了 43.7 AP，比基线高出 3.5 AP，超过了所有以前的最先进的目标检测知识蒸馏方法。此外，实验还表明我们的 CrossKD 与特征模仿方法是正交的。通过将 CrossKD 与最先进的特征模仿方法结合，如 PKD [5]，我们在 GFL 上进一步达到了 43.9 AP。此外，我们还展示了我们的方法可以用于蒸馏具有异构骨架的检测器，并且比其他方法表现得更好。

2. 相关工作

2.1. 目标检测

目标检测是计算机视觉中最基础的任务之一，它要求同时识别和定位对象。现代目标检测器可以简要地分为两类：单阶段 [3, 10, 11, 35, 39, 50, 56, 69] 检测器和双阶段 [8, 17, 18, 20, 21, 38, 51, 57, 72] 检测器。其中，单阶段检测器，也称为密集检测器，由于其出色的速度-准确性权衡，已成为检测领域的主流趋势。

自 YOLOv1 [48] 以来，密集目标检测器受到了极大的关注。通常，YOLO 系列检测器 [2, 16, 43, 48–50] 试图平衡模型大小和准确性，以满足实际应用的要求。无锚点检测器 [27, 56, 75] 试图摒弃锚点框的设计，以避免耗时的框操作和繁琐的超参数调整。动态标签分配方法 [15, 46, 69] 被提出，以更好地定义模型学习的正负样本。GFL [34, 35] 引入了质量焦点损失（QFL）和分布引导的质量预测器，以增加分类分数和

定位质量之间的一致性。它还将边界框表示为概率分布，以便能够捕捉框边缘的定位模糊性。最近，由于 Transformer 编码表达性特征的强大能力，DETR 家族 [4, 6, 30, 41, 44, 67, 76] 已成为目标检测社区的新趋势。

2.2. 目标检测中的知识蒸馏

知识蒸馏（KD）是一种有效的技术，可以将大规模教师模型的知识转移到小规模学生模型中。它在分类任务中已经被广泛研究 [12, 23, 26, 36, 37, 45, 47, 53, 62, 66, 70, 71]，但由于极端的背景比例，蒸馏检测模型仍然具有挑战性。先驱工作 [7] 提出了第一个目标检测的蒸馏框架，通过简单地结合特征模仿和预测模仿。从那时起，特征模仿吸引了越来越多的研究关注。通常，一些工作 [13, 25, 33, 60] 专注于选择有效的蒸馏区域以更好地进行特征模仿，而其他工作 [19, 31, 74] 旨在更好地加权模仿损失。还有一些方法 [5, 64, 65, 68] 尝试设计新的教师-学生一致性函数，旨在探索更多的一致性信息或解除 MSE 损失的严格限制。

作为在 [24] 中提出的最早的蒸馏策略，预测模仿在分类蒸馏中扮演着至关重要的角色。最近，一些改进的预测模仿方法被提出以适应目标检测。例如，Rank Mimicking [31] 将教师的分数排名视为一种知识，并旨在迫使学生像教师一样对实例进行排名。LD [73] 提出蒸馏边界框的定位分布 [35] 以传递定位知识。在本文中，我们构建了一个 CrossKD 流程，将检测和蒸馏分离到不同的头部，以缓解预测模仿的目标冲突问题。值得注意的是，HEAD [58] 将学生特征传递给一个独立的助手头部，以弥合异构教师-学生对之间的差距。相比之下，我们观察到简单地将学生特征传递给教师就足以实现 SOTA 结果。这使得我们的方法非常简洁，与 HEAD 不同。我们的方法也与 [1, 28, 32, 63] 相关，但它们都旨在蒸馏分类模型，并不针对目标检测。

3. 方法

3.1. 目标冲突问题分析

目标冲突是传统预测模仿方法中面临的一个常见问题。与为每张图片分配特定类别的分类任务不同，高级检测器中的标签通常是动态分配的，并且不是确定性的。通常，检测器依赖于一个手工制定的原则，即分配器，来确定每个位置的标签。在大多数情况下，检测器无法精确复制分配器的标签，这导致在知识蒸馏（KD）中教师-学生目标之间产生冲突。此外，在现实世界场景中，学生和教师的分配器不一致性进一步扩大了真实标注和蒸馏目标之间的距离。

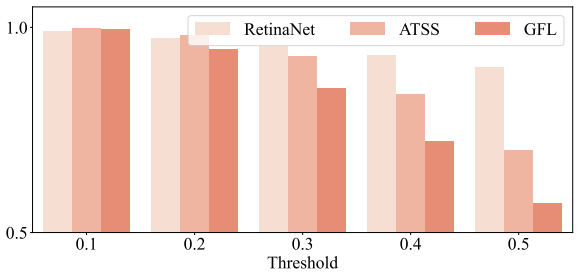

图 3. 学生（GFL-R50）与教师（GFL-R101, ATSS-R101, RetinaNet-R101）之间目标冲突程度的统计。X 轴是用于冲突区域的教师-学生差异阈值。Y 轴表示目标冲突区域与正区域的比例。

为了定量测量目标冲突的程度，我们统计了在 COCO *minival* 数据集中不同教师-学生差异下冲突区域与正区域的比例，并在 Fig. 3中报告了结果。正如我们所见，即使教师（ATSS [69] 和 GFL [35]）和学生（GFL）具有相同的标签分配策略，仍然有许多位置在真实标注和蒸馏目标之间存在大于 0.5 的差异。当我们使用具有不同分配器（RetinaNet）的教师来蒸馏学生（GFL）时，冲突区域大幅增加。Sec. 4.5中的更多实验也证明，目标冲突问题严重阻碍了预测模仿的性能。

尽管目标冲突有很大的影响，但这个问题在以前的预测模仿方法中 [24, 31] 长时间被忽视。这些方法旨在直接最小化教师-学生预测之间的差异。其目标可以描述为：

$$\mathcal{L}_{\text{KD}} = \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{R}} \mathcal{S}(r) \mathcal{D}_{\text{pred}}(\boldsymbol{p}^s(r), \boldsymbol{p}^t(r)), \qquad (1)$$

其中 $\boldsymbol{p}^s$ 和 $\boldsymbol{p}^t$ 分别是由学生和教师的检测头部生成的预测向量。$\mathcal{D}_{\text{pred}}(\cdot)$ 指的是计算 $\boldsymbol{p}^s$ 和 $\boldsymbol{p}^t$ 之间差异的损失函数，例如，分类任务中的 KL 散度 [24]，回归任务中的 L1 损失 [7] 和 LD [73]。$\mathcal{S}(\cdot)$ 是区域选择原则，它在整张图像区域 $\mathcal{R}$ 中的每个位置 r 生产一个权重。

值得注意的是，$\mathcal{S}(\cdot)$ 在一定程度上可以通过减少教师-学生差异较大的区域的权重来缓解目标冲突问题。然而，高度不确定的区域通常包含比无争议区域更多的对学生有益的信息。忽略这些区域可能对预测模仿方法的有效性产生很大影响。因此，为了推动预测模仿

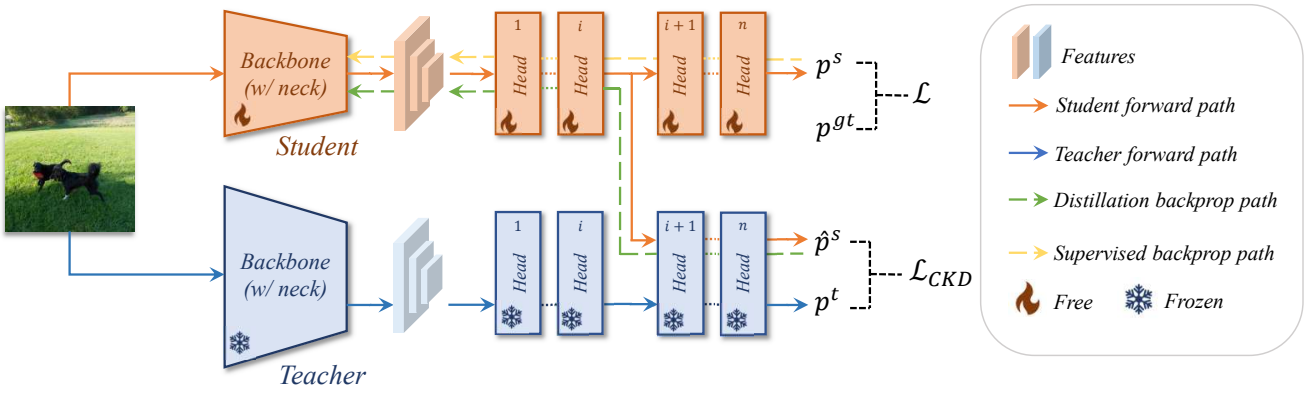

图 4. 提出的 CrossKD 的总体框架。对于给定的教师-学生对，CrossKD 首先将学生的中间特征传递到教师层，并生成交叉头部预测 $\hat{\boldsymbol{p}}^s$。然后，计算原始教师预测和学生交叉头部预测之间的蒸馏损失。在反向传播中，与检测损失相关的梯度通常通过学生检测头部传递，而蒸馏梯度通过冻结的教师层传播。

的极限，有必要优雅地处理目标冲突问题，而不是直接减少权重。

3.2. 交叉头式知识蒸馏

如 Sec. 3.1中所述，我们观察到直接模仿教师的预测会面临目标冲突问题，这阻碍了预测模仿达到有希望的性能。为了缓解这个问题，我们在本节提出了一种新颖的交叉头部知识蒸馏（CrossKD）。整体框架在 Fig. 4中进行了说明。像许多以前的预测模仿方法一样，我们的 CrossKD 在预测上执行蒸馏过程。不同的是，CrossKD 将学生的中间特征传递给教师的检测头部，并生成交叉头部预测以进行蒸馏。

给定一个密集检测器，如 RetinaNet [39]，每个检测头部通常由一系列卷积层组成，表示为 $\{C_i\}$。为简化起见，我们假设每个检测头部总共有 n 个卷积层（例如，在具有 4 个隐藏层和 1 个预测层的 RetinaNet 中为 5）。我们使用 $\boldsymbol{f}_i, i \in \{1, 2, \cdots, n-1\}$ 来表示由 C_i 产生的特征图，而 $\boldsymbol{f}_0$ 表示 C_1 的输入特征图。预测 $\boldsymbol{p}$ 由最后一个卷积层 C_n 生成。因此，对于给定的教师-学生对，教师和学生的预测可以分别表示为 $\boldsymbol{p}^t$ 和 $\boldsymbol{p}^s$。

除了教师和学生原有的预测之外，CrossKD 还额外将学生的中间特征 $\boldsymbol{f}_i^s, i \in \{1, 2, \cdots, n-1\}$ 传递给教师检测头部的第 $(i+1)$ 个卷积层 C_{i+1}^t，从而产生交叉头部预测 $\hat{\boldsymbol{p}}^s$。给定 $\hat{\boldsymbol{p}}^s$，我们不是计算 $\boldsymbol{p}^s$ 和 $\boldsymbol{p}^t$ 之间的知识蒸馏（KD）损失，而是提议使用交叉头部预测 $\hat{\boldsymbol{p}}^s$ 和教师原始预测 $\boldsymbol{p}^t$ 之间的知识蒸馏损失作为我们 CrossKD 的目标，描述如下：

$$\mathcal{L}_{\text{CrossKD}} = \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{R}} \mathcal{S}(r) \mathcal{D}_{\text{pred}}(\hat{\boldsymbol{p}}^s(r), \boldsymbol{p}^t(r)), \tag{2}$$

其中 $\mathcal{S}(\cdot)$ 和 $|\mathcal{S}|$ 分别是区域选择原则和归一化因子。我们没有设计复杂的 $\mathcal{S}(\cdot)$，而是在整个预测图上平等地进行 $\hat{\boldsymbol{p}}^s$ 和 $\boldsymbol{p}^t$ 之间的蒸馏。具体来说，在我们的 CrossKD 中，$\mathcal{S}(\cdot)$ 是一个常数函数，其值为 1。根据每个分支的不同任务（例如，分类或回归），我们执行不同类型的 $\mathcal{D}_{\text{pred}}(\cdot)$，以有效地向学生传递特定任务的知识。

通过执行 CrossKD，检测损失和蒸馏损失分别应用于不同的分支。如图 Fig. 4所示，检测损失的梯度通过学生的整个头部传递，而蒸馏损失的梯度通过冻结的教师层传播到学生的潜在特征，这启发式地增加了教师和学生之间的一致性。与直接缩小教师-学生对之间的预测差异相比，CrossKD 允许学生检测头部的一部分仅与检测损失相关，从而更好地优化以接近真实标注目标。我们在实验部分提供了定量分析。

3.3. 优化目标

训练的总损失可以表示为检测损失和蒸馏损失的加权和，写成如下形式：

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{\text{cls}}(\boldsymbol{p}^s_{\text{cls}}, \boldsymbol{p}^{gt}_{\text{cls}}) + \mathcal{L}_{\text{reg}}(\boldsymbol{p}^s_{\text{reg}}, \boldsymbol{p}^{gt}_{\text{reg}}) \\ & + \mathcal{L}^{\text{cls}}_{\text{CrossKD}}(\hat{\boldsymbol{p}}^s_{\text{cls}}, \boldsymbol{p}^t_{\text{cls}}) + \mathcal{L}^{\text{reg}}_{\text{CrossKD}}(\hat{\boldsymbol{p}}^s_{\text{reg}}, \boldsymbol{p}^t_{\text{reg}}), \end{aligned} \tag{3}$$

其中 $\mathcal{L}_{\text{cls}}$ 和 $\mathcal{L}_{\text{reg}}$ 分别代表检测损失，这些损失是根据学生预测 $\boldsymbol{p}^s_{\text{cls}}$, $\boldsymbol{p}^s_{\text{reg}}$ 与它们相应的真实目标 $\boldsymbol{p}^{gt}_{\text{cls}}$, $\boldsymbol{p}^{gt}_{\text{reg}}$

表 1. 在不同位置应用 CrossKD 的有效性。索引 i 表示用作交叉头部分支输入的中间特征。'LD' 表示在学生的头部直接应用预测模仿，使用 LD [73]。教师-学生对是具有 ResNet-50 和 ResNet-18 骨架的 GFL。我们可以看到，在这次实验中 $i = 3$ 产生了最佳性能。

i	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
-	35.8	53.1	38.2	18.9	38.9	47.9
0	38.2	55.6	41.3	20.2	41.9	50.9
1	38.3	55.8	41.1	20.8	42.1	49.8
2	38.6	56.2	41.5	20.8	42.7	50.7
3	38.7	56.3	41.6	21.1	42.2	51.1
4	38.2	55.7	41.2	20.3	41.9	50.2
LD	37.8	55.5	40.5	20.0	41.4	49.5

表 2. 特征模仿与 CrossKD 之间的比较。我们选择先进的 PKD 来代表特征模仿，并将 PKD 应用于不同位置，以公平地与 CrossKD 进行比较。在这里，'neck' 表示在 FPN 颈部执行 PKD。'cls' 和'reg' 分别表示将 PKD 应用于分类分支和回归。教师-学生对是具有 ResNet-50 和 ResNet-18 骨架的 GFL。

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
-	35.8	53.1	38.2	18.9	38.9	47.9
PKD:neck	38.0	55.0	41.2	19.6	41.5	50.2
PKD:cls	37.5	54.9	40.5	19.5	41.1	50.5
PKD:reg	37.2	54.0	40.2	19.0	40.9	50.0
PKD:cls+reg	37.3	54.3	40.0	19.2	41.1	49.8
CrossKD	38.7	56.3	41.6	21.1	42.2	51.1

计算得出的。额外的 CrossKD 损失表示为 $\mathcal{L}^{\text{cls}}_{\text{CrossKD}}$ 和 $\mathcal{L}^{\text{reg}}_{\text{CrossKD}}$，它们是在交叉头部预测 $\hat{\boldsymbol{p}}^s_{\text{cls}}$, $\hat{\boldsymbol{p}}^s_{\text{reg}}$ 与教师预测 $\boldsymbol{p}^t_{\text{cls}}$, $\boldsymbol{p}^t_{\text{reg}}$ 之间进行的。

我们应用不同的距离函数 $\mathcal{D}_{\text{pred}}$ 来在不同的分支中传递特定任务的信息。在分类分支中，我们将教师预测的分类分数视为软标签，并直接使用 GFL [35] 中提出的 Quality Focal Loss (QFL) 来缩小教师-学生之间的距离。至于回归，密集检测器中主要有两种回归形式。第一种回归形式直接从锚点框（例如，RetinaNet [39]，ATSS [69]）或点（例如，FCOS [56]）回归边界框。在这种情况下，我们直接使用 GIoU [52] 作为 $\mathcal{D}_{\text{pred}}$。在另一种情况下，回归形式预测一个向量来表示框位置的分布（例如，GFL [35]），这比边界框表示的狄拉克分布包含更丰富的信息。为了有效地蒸馏位置分布的知识，我们采用 KL 散度，如 LD [73]，来传递定位知识。有关损失函数的更多细节在补充材料中给出。

4. 实验

4.1. 实现细节

根据大多数先前工作所做的方式，我们在大规模 MS COCO [40] 基准测试上评估所提出的方法。为确保与标准实践一致，我们使用 *trainval135k* 集（$115K$ 图像）进行训练，使用 *minival* 集（$5K$ 图像）进行验证。在评估中，使用标准的 COCO 风格测量，即平均精度（AP）。我们还报告了 IoU 阈值为 0.5 和 0.75 的 mAP，以及小型、中型和大型物体的 AP。所提出的方法，CrossKD，在 Python 中的 MMDetection [9] 框架下实现。为了公平比较，所有实验都使用 8 个 Nvidia V100 GPU 开发，每个 GPU 的小批量为两张图像。除非另有说明，所有超参数在训练和测试中均遵循相应学生模型的默认设置。

4.2. 方法分析

为了研究我们方法的有效性，我们基于 GFL [35] 进行了广泛的消融实验。如果没有特别说明，我们使用具有 ResNet-50 骨架 [22] 的 GFL 作为教师检测器，并在学生检测器中使用 ResNet-18 骨架。教师和学生模型的准确率分别为 40.2 AP 和 35.8 AP。所有实验遵循默认的 1× 训练计划（12 个周期）。

应用 CrossKD 的位置。如 Sec. 3.2所述，CrossKD 将学生的第 i 个中间特征传递给教师头部的一部分。在这里，我们在分类和边界框回归分支上都进行蒸馏。当 $i = 0$ 时，CrossKD 直接将学生的 FPN 特征输入到教师的头部。在这种情况下，学生的整个头部只受到检测损失的监督，不涉及蒸馏损失。随着 i 逐渐增加，学生头部的更多层同时受到检测损失和蒸馏损失的影响。当 $i = n$ 时，我们的方法退化为原始的预测模仿，蒸馏损失将直接在教师-学生对的两个预测之间执行。

在 Tab. 1中，我们报告了在不同中间特征上执行 CrossKD 的结果。可以看到，我们的 CrossKD 可以提高所有 i 选择的蒸馏性能。这一发现意味着交叉头部策略可以有效增强预测模仿的性能。如表所示，无论位置

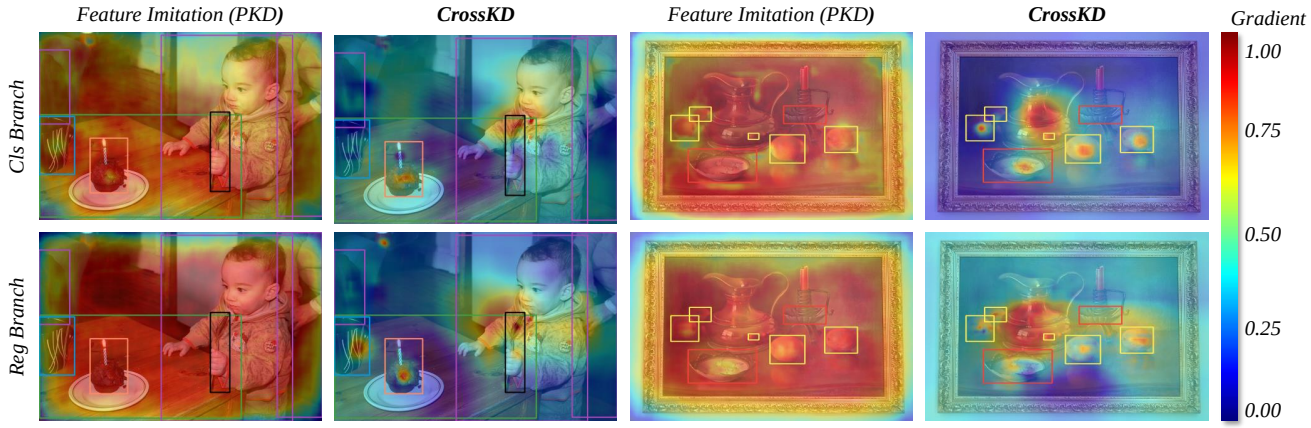

图 5. 特征模仿和 CrossKD 梯度的可视化。可视化表明，我们由预测模仿引导的 CrossKD 可以有效地专注于潜在有价值的区域。

表 3. CrossKD 在不同分支上的有效性。我们分别在分类（cls）和回归（reg）分支上应用 CrossKD。教师-学生对是具有 ResNet-50 和 ResNet-18 骨架的 GFL。

cls	reg	LD			CrossKD		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
✓		37.3	55.2	40.0	37.7	55.6	40.2
	✓	36.8	53.8	39.6	37.2	54.0	40.0
✓	✓	37.8	55.4	40.5	38.7	56.3	41.6

如何，CrossKD 至少可以将结果提高 0.4% mAP，与预测模仿相比，这意味着交叉头部策略可以有效改善预测模仿的性能。值得注意的是，当使用第 3 个中间特征时，CrossKD 达到了 38.7 AP 的最佳性能，比最近最先进的预测模仿方法 LD [73] 高出 0.9 AP。这表明，学生头部的所有层并不都需要隔离蒸馏损失的影响。因此，我们在所有后续实验中将 $i = 3$ 作为默认设置。

CrossKD 与特征模仿的比较. 我们将 CrossKD 与先进的特征模仿方法 PKD [5] 进行比较。为了公平比较，我们在与我们的 CrossKD 相同的位置执行 PKD，包括 FPN 特征和检测头部的第三层。结果在 Tab. 2 中报告。可以看出，当 PKD 应用于教师-学生对的 FPN 特征时，可以达到 38.0 AP。在检测头部，PKD 甚至显示出性能下降。相比之下，我们的 CrossKD 达到了 38.7 AP，比应用于 FPN 特征的 PKD 高出 0.7 AP。

为了进一步调查 CrossKD 的优势，我们可视化了检测头部潜在特征上的梯度，如图 Fig. 5 所示。如图所示，由 PKD 生成的梯度对整个特征图有大范围且广泛的影响，这是低效且不具针对性的。相反，由 CrossKD 生成的梯度可以专注于具有潜在语义的区域，这些区域包含感兴趣的物体。

表 4. CrossKD 和预测模仿的集体效应。教师-学生对是具有 ResNet-50 和 ResNet-18 骨架的 GFL。

CrossKD	LD	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
-	-	35.8	53.1	38.2	18.9	38.9	47.9
✓		38.7	56.3	41.6	21.1	42.2	51.1
	✓	37.8	55.5	40.5	20.0	41.4	49.5
✓	✓	38.1	55.6	40.9	20.4	41.6	51.1

CrossKD v.s. 预测模仿。我们首先分别在分类和边界框回归分支上单独执行预测模仿和 CrossKD。结果在 Tab. 3中报告。可以看出，无论分类还是回归分支，用 CrossKD 替换预测模仿都能带来稳定的性能提升。具体来说，预测模仿在分类和回归分支上分别产生了 37.3 AP 和 36.8 AP，而 CrossKD 产生了 37.7 AP 和 37.2 AP，这代表了对预测模仿相应结果的一致改进。如果在两个分支上执行知识蒸馏（KD），我们的方法仍然可以比预测模仿高出 +0.9 AP。此外，我们进一步评估了预测模仿和 CrossKD 的集体效应，如 Tab. 4所示。有趣的是，我们观察到同时使用预测模仿和 CrossKD 的最终结果是 38.1 AP，这甚至低于单独使用 CrossKD 的结果。我们认为这是因为预测模仿再次引入了目标冲突问题，这使得学生模型在学习上感到困难。

此外，我们可视化训练过程中的统计变化，以对

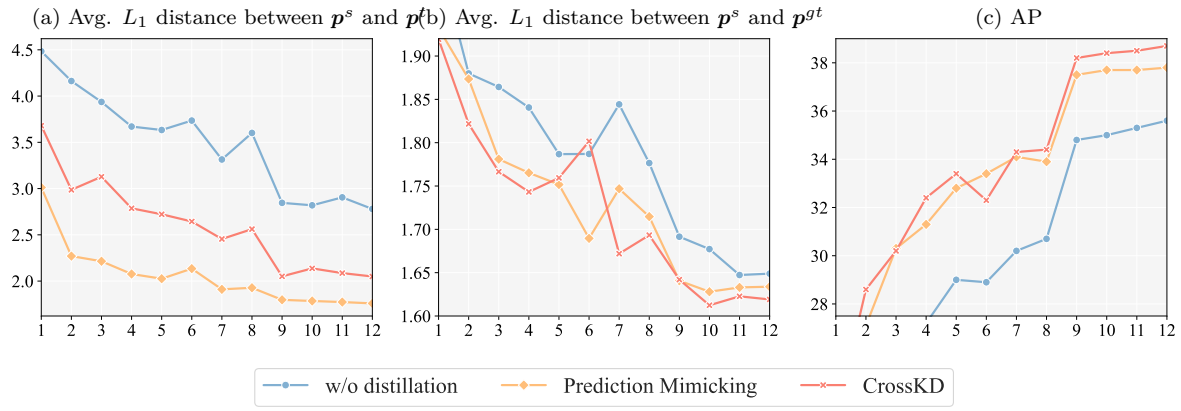

图 6. 训练过程中统计数据变化的可视化。(a) 学生预测 p^s 与教师预测 p^t 之间平均 L_1 距离的曲线。(b) 学生预测 p^s 与正真实标注目标 p^{gt} 之间平均 L_1 距离的曲线。(c) 平均精度 (AP) 的曲线。所有曲线都是在 COCO *minival* 集上评估的。X 轴指的是周期编号。(a) 和 (b) 中的 Y 轴表示平均 L_1 距离，而 (c) 中表示 AP 的值。

CrossKD 和预测模仿进行进一步分析。我们首先计算每个时期学生预测 p^s 与教师预测 p^t 以及真实目标 p^{gt} 之间的 L_1 距离。如 Fig. 6(a) 所示，通过我们的 CrossKD，距离 $L_1(p^s, p^t)$ 可以显著减少，而预测模仿实现最低距离是合理的，因为蒸馏是直接施加在 p^s 上的。然而，由于存在优化目标冲突，预测模仿涉及一个矛盾的优化过程，因此通常产生比我们的 CrossKD 更大的距离 $L_1(p^s, p^{gt})$，如 Fig. 6(b) 所示。在 Fig. 6(c) 中，我们的方法显示出更快的训练过程，并实现了 37.8 AP 的最佳性能。

4.3. 与 SOTA 知识蒸馏方法的对比

在本节中，我们在 GFL [35] 框架上评估各种 state-of-the-art 目标检测知识蒸馏 (KD) 方法，并与我们提出的 CrossKD 进行公平比较。我们使用 ResNet-101 作为教师检测器的主干网络，它使用 2× 训练计划和多尺度增强进行训练。对于学生检测器，我们采用 ResNet-50 作为主干网络。我们使用 1× 训练计划来训练学生。教师检测器基于 ResNet101-FPN 主干网络，使用 2× 训练计划和多尺度增强进行训练，而学生仅采用 ResNet50-FPN 主干网络，并使用 1× 训练计划进行训练。教师的预训练检查点直接从 MMDetection[9] 模型库中借用。

我们在 Tab. 5中报告了所有结果。正如我们所见，在相同条件下，CrossKD 能够在没有额外装饰的情况下达到 43.7 AP，将学生的准确性提高了 3.5 AP，超过了所有其他 state-of-the-art 方法。值得注意的是，CrossKD 比先进的特征模仿方法 PKD 高出 0.4 AP，比先进的预测模仿方法 LD 高出 0.7 AP，这证明了 CrossKD 的有效性。此外，我们还观察到 CrossKD 也与特征模仿方法正交。在 PKD 的帮助下，CrossKD 达到了最高的 43.9 AP，与基线相比提高了 3.7 AP。

表 5. 与 COCO 上的 state-of-the-art 检测知识蒸馏方法的比较。* 表示结果参考自 LD [73] 和 PKD [5]。所有结果都在 COCO 的 *minival* 集上进行评估。

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
GFL-R101 (T)	44.9	63.1	49.0	28.0	49.1	57.2
GFL-R50 (S)	40.2	58.4	43.3	23.3	44.0	52.2
FitNets* [53]	40.7 (0.5↑)	58.6	44.0	23.7	44.4	53.2
Inside GT Box*	40.7 (0.5↑)	58.6	44.2	23.1	44.5	53.5
Defeat* [19]	40.8 (0.6↑)	58.6	44.2	24.3	44.6	53.7
Main Region* [73]	41.1 (0.9↑)	58.7	44.4	24.1	44.6	53.6
Fine-Grained* [61]	41.1 (0.9↑)	58.8	44.8	23.3	45.4	53.1
FGD [64]	41.3 (1.1↑)	58.8	44.8	24.5	45.6	53.0
GID* [13]	41.5 (1.3↑)	59.6	45.2	24.3	45.7	53.6
SKD [14]	42.3 (2.1↑)	60.2	45.9	24.4	46.7	55.6
LD [73]	43.0 (2.8↑)	61.6	46.6	25.5	47.0	55.8
PKD* [5]	43.3 (3.1↑)	61.3	46.9	25.2	47.9	56.2
CrossKD	43.7 (3.5↑)	**62.1**	47.4	**26.9**	48.0	56.2
CrossKD+PKD	**43.9** (**3.7↑**)	62.0	**47.7**	26.4	**48.5**	**57.0**

表 6. CrossKD 用于具有同质主干网络的检测器。教师检测器使用 ResNet-101 作为主干网络，而学生检测器使用 ResNet-50 作为主干网络。所有结果都在 COCO 的 *minival* 集上进行评估。

Student	Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [39]	R101	38.9	58.0	41.5	21.0	32.8	52.4
	R50	37.4	56.7	39.6	20.0	40.7	49.7
	CrossKD	39.7	58.9	42.5	22.4	43.6	52.8
FCOS [56]	R101	40.8	60.0	44.0	24.2	44.3	52.4
	R50	38.5	57.7	41.0	21.9	42.8	48.6
	CrossKD	41.3	60.6	44.2	25.1	45.5	52.4
ATSS [69]	R101	41.5	59.9	45.2	24.2	45.9	53.3
	R50	39.4	57.6	42.8	23.6	42.9	50.3
	CrossKD	41.8	60.1	45.4	24.9	45.9	54.2

表 7. CrossKD 用于具有不同标签分配器的教师-学生对。所有结果都在 COCO 的 *minival* 集上进行评估。

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
GFL-R50 (S)	40.2	58.4	43.3	23.3	44.0	52.2
ATSS[69]-R101 (T)	41.5	59.9	45.2	24.2	45.9	53.3
KD	39.7	57.9	42.8	21.8	44.2	51.5
CrossKD	42.1	60.5	45.7	24.5	46.3	54.5
GFL-R50 (S)	40.2	58.4	43.3	23.3	44.0	52.2
Retinanet[39]-R101 (T)	38.9	58.0	41.5	21.0	32.8	52.4
KD	30.3	49.2	31.2	20.0	38.1	34.4
CrossKD	41.2	59.4	44.8	24.0	45.1	53.5

4.4. CrossKD 用于不同的检测器

除了在 GFL 上执行 CrossKD 外，我们选择了三种常用的检测器，即 RetinaNet[39]、FCOS [56] 和 ATSS [69]，来研究 CrossKD 的有效性。我们严格遵循学生设置进行训练，并参考 MMDetection 模型库中的教师和学生结果。结果在 Tab. 6中呈现。如 Tab. 6所示，CrossKD 显著提升了这三种类型检测器的性能。具体来说，采用我们的 CrossKD 的 RetinaNet、FCOS 和 ATSS 分别达到了 39.7 AP、41.3 AP 和 41.8 AP，分别比它们相应的基线高出 2.3 AP、2.8 AP 和 2.4 AP。蒸馏后的所有结果甚至超过了原始教师的性能，表明 CrossKD 可以在不同的密集检测器上很好地工作。

4.5. 在严重目标冲突下的蒸馏

在这一小节中，我们在具有不同分配器的检测器之间执行预测模仿和我们的 CrossKD，以探索 CrossKD 对抗目标冲突问题的有效性。如 Tab. 7所示，目标冲突问题对学生的优化有很大的影响，导致性能较差。具体来说，当教师为 RetinaNet（与 GFL 具有不同的分配器）时，预测模仿将 AP 降低到 30.3。此外，即使 ATSS 与 GFL 具有相同的分配器，学生的 AP 也只能蒸馏到 39.7，低于没有知识蒸馏的性能。相比之下，即使在真实目标和蒸馏目标之间存在很大的差异，CrossKD 仍然可以显著提高学生的准确性。当将 ATSS 作为教师应用时，CrossKD 将 GFL-R50 的准确性提高到 42.1（+1.9 AP）。即使在弱教师 RetinaNet 的指导下，CrossKD 仍然将 GFL-R50 的性能提高到 41.2 AP，比基线高出 1.0 AP。这展示了我们的 CrossKD 在面对严重目标冲突时的鲁棒性。

4.6. 异构主干网络间的蒸馏

在这一小节中，我们探索了我们的 CrossKD 在蒸馏异构学生方面的能力。我们在具有不同主干网络的 RetinaNet [39] 上进行知识蒸馏，并与最新的先进方法 PKD [5] 进行比较。具体来说，我们选择了两种典型的异构主干网络，即变换器主干 Swin-T [42] 和轻量级主干 MobileNetv2 [54]。所有检测器都以单尺度策略训练 12 个周期。结果在 Tab. 8中呈现。我们可以看到，当从 Swin-T 蒸馏知识时，CrossKD 达到了 38.0 AP（+1.5 AP），比 PKD 高出 0.8 AP。CrossKD 还将具有 MobileNetv2 主干的 RetinaNet 的结果提高到 34.1 AP，比基线高出 3.2 AP，超过了 PKD 的 0.9 AP。

5. 结论与讨论

在本文中，我们介绍了 CrossKD，这是一种新颖的知识蒸馏 (KD) 方法，旨在提升密集目标检测器的性能。CrossKD 将学生头部的中间特征转移到教师头部，以产生用于蒸馏的跨头部预测，这是一种有效缓解监督和蒸馏目标之间冲突的方法。我们的结果显示，CrossKD 可以提高蒸馏效率并实现 state-of-the-art 性能。将来，我们将进一步将我们的方法扩展到其他相关领域，例如 3D 目标检测。

致谢。本研究得到了国家自然科学基金 (62225604 号，

表 8. CrossKD 用于具有异构主干网络的其他检测器对。为方便起见，仅列出以下主干列表，其中 'SwinT' 指的是带有小型 Swin Transformer 版本的 RetinaNet [42]。所有结果都在 COCO 的 *minival* 集上进行评估。

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SwinT (T) [42]	37.3	57.5	39.9	22.7	41.0	49.6
ResNet-50 (S)	36.5	55.4	39.1	20.4	40.3	48.1
PKD	37.2	56.7	39.5	21.2	41.2	49.7
CrossKD	38.0	58.1	40.5	23.1	41.8	49.7
ResNet-50 (T)	36.5	55.4	39.1	20.4	40.3	48.1
MobileNetv2 (S) [54]	30.9	48.7	32.5	16.3	33.5	41.9
PKD	33.2	51.3	35.0	16.5	36.6	46.5
CrossKD	34.1	52.7	36.5	18.8	37.1	45.4

62276145 号)、中央高校基本科研业务费 (南开大学, 070-63223049)、中国科学技术协会通过青年精英科学家资助计划 (编号 YESS20210377) 的支持。计算得到了南开大学超算中心 (NKSC) 的支持。

References

[1] Haoli Bai, Jiaxiang Wu, Irwin King, and Michael Lyu. Few shot network compression via cross distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 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] Qinglong Cao, Zhengqin Xu, Yuantian Chen, Chao Ma, and Xiaokang Yang. Domain-controlled prompt learning. *arXiv preprint arXiv:2310.07730*, 2023.

[4] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. Domain prompt learning with quaternion networks. *arXiv preprint arXiv:2312.08878*, 2023.

[5] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. In *Advances in Neural Information Processing Systems*, 2022.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[7] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

[8] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] 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.

[10] Shaoyu Chen, Tianheng Cheng, Jiemin Fang, Qian Zhang, Yuan Li, Wenyu Liu, and Xinggang Wang. Tinydet: accurately detecting small objects within 1 gflops. *Science China Information Sciences*, 66(1):119102, 2023.

[11] Yuming Chen, Xinbin Yuan, Ruiqi Wu, Jiabao Wang, Qibin Hou, and Ming-Ming Cheng. Yolo-ms: Rethinking multi-scale representation learning for real-time object detection, 2023.

[12] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[13] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7842–7851, June 2021.

[14] Philip De Rijk, Lukas Schneider, Marius Cordts, and

Dariu Gavrila. Structural knowledge distillation for object detection. In *Advances in Neural Information Processing Systems*, 2022.

[15] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3510–3519, October 2021.

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

[17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[18] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[19] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2154–2164, June 2021.

[20] Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Xinghao Chen, Chunjing Xu, Chang Xu, and Yunhe Wang. Positive-unlabeled data purification in the wild for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2653–2662, 2021.

[21] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[22] 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 (CVPR)*, June 2016.

[23] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network (2015). *arXiv preprint arXiv:1503.02531*, 2, 2015.

[25] Zihao Jia, Shengkun Sun, Guangcan Liu, and Bo Liu. Mssd: multi-scale self-distillation for object detection. *Visual Intelligence*, 2(1):8, 2024.

[26] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

[27] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020.

[28] Animesh Koratana, Daniel Kang, Peter Bailis, and Matei Zaharia. LIT: Learned intermediate representation training for model compression. In *International Conference on Learning Representations*, 2019.

[29] Yuqing Lan, Yao Duan, Chenyi Liu, Chenyang Zhu, Yueshan Xiong, Hui Huang, and Kai Xu. Arm3d: Attention-based relation module for indoor 3d object detection. *Computational Visual Media*, 8(3):395–414, 2022.

[30] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13619–13627, June 2022.

[31] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1306–1313, 2022.

[32] Guilin Li, Junlei Zhang, Yunhe Wang, Chuanjian Liu, Matthias Tan, Yunfeng Lin, Wei Zhang, Jiashi Feng, and Tong Zhang. Residual distillation: Towards portable deep neural networks without shortcuts. *Advances in Neural Information Processing Systems*, 33:8935–8946, 2020.

[33] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11632–11641, June 2021.

[35] 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21002–21012. Curran Associates, Inc., 2020.

[36] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 2, pages 1504–1512, 2023.

[37] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11740–11750, 2021.

[38] Tsung-Yi Lin, Piotr Dollar, 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 (CVPR)*, July 2017.

[39] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[41] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022.

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021.

[43] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022.

[44] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3651–3660, October 2021.

[45] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5191–5198, 2020.

[46] Chuong H. Nguyen, Thuy C. Nguyen, Tuan N. Tang, and Nam L.H. Phan. Improving object detection by label assignment distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1005–1014, January 2022.

[47] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[48] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[49] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detec-

tion with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[53] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[54] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[55] Lin Song, Jin-Fu Yang, Qing-Zhen Shang, and Ming-Ai Li. Dense face network: A dense face detector based on global context and visual attention mechanism. *Machine Intelligence Research*, 19(3):247–256, 2022.

[56] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[57] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[58] Luting Wang, Xiaojie Li, Yue Liao, Zeren Jiang, Jianlong Wu, Fei Wang, Chen Qian, and Si Liu. Head: Hetero-assists distillation for heterogeneous object detectors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 314–331. Springer, 2022.

[59] Luequan Wang, Hongbin Xu, and Wenxiong Kang. Mvcontrast: Unsupervised pretraining for multi-view 3d object recognition. *Machine Intelligence Research*, 20(6):872–883, 2023.

[60] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[61] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[62] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L. Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[63] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2021.

[64] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4643–4652, June 2022.

[65] Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, and Tong Zhang. G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3591–3600, October 2021.

[66] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[67] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023.

[68] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2021.

[69] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[70] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[71] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11953–11962, June 2022.

[72] Wangbo Zhao, Jiasheng Tang, Yizeng Han, Yibing Song, Kai Wang, Gao Huang, Fan Wang, and Yang You. Dynamic tuning towards parameter and inference efficiency for vit adaptation, 2024.

[73] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9407–9416, June 2022.

[74] Du Zhixing, Rui Zhang, Ming Chang, Shaoli Liu, Tianshi Chen, Yunji Chen, et al. Distilling object detectors with feature richness. *Advances in Neural Information Processing Systems*, 34:5213–5224, 2021.

[75] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[76] 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*, 2021.