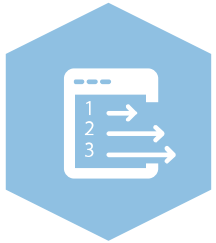
Active Learning for Open-set Annotation

Kun-Peng Ning[1*] Xun Zhao[2†] Yu Li[3*] Sheng-Jun Huang[1†]
[1]Nanjing University of Aeronautics and Astronautics
[2]Applied Research Center, Tencent PCG [3]International Digital Economy Academy (IDEA)
{ningkp,huangsj}@nuaa.edu.cn emmaxunzhao@tencent.com liyu@idea.edu.cn

汇报人：蒋明忠

时间：2025.11



CVPR

Background

深度学习的显著成功在很大程度上归功于具有人类注释标签的大型数据集的收集。然而,用高质量的注释标记大规模数据是非常昂贵和耗时的。因此,使用有限的标记数据进行学习是一个重大的挑战。

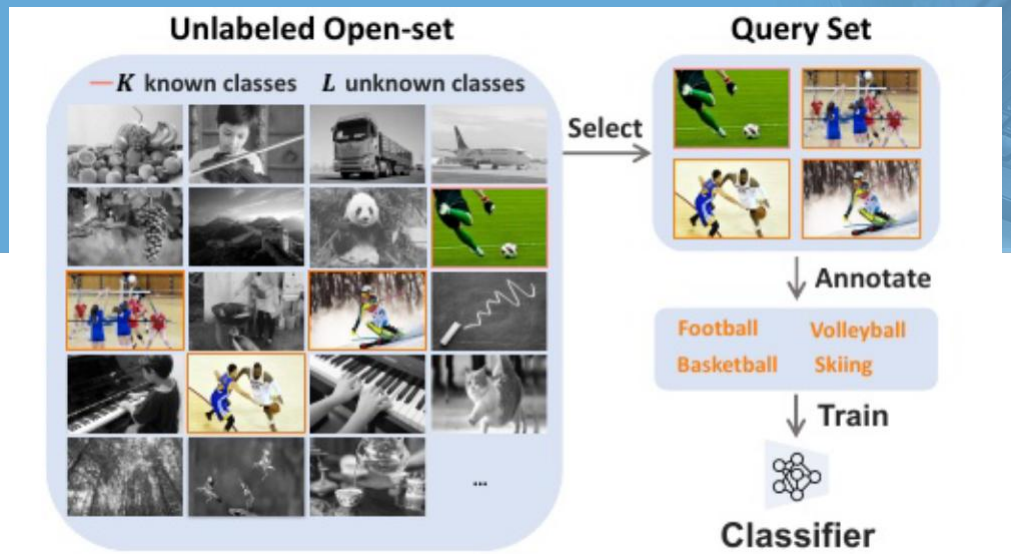
主动学习(AL)是解决这一问题的主要方法，它迭代地从未标记的数据中选择最有用的例子，在降低标注成本的同时实现有竞争力的性能。

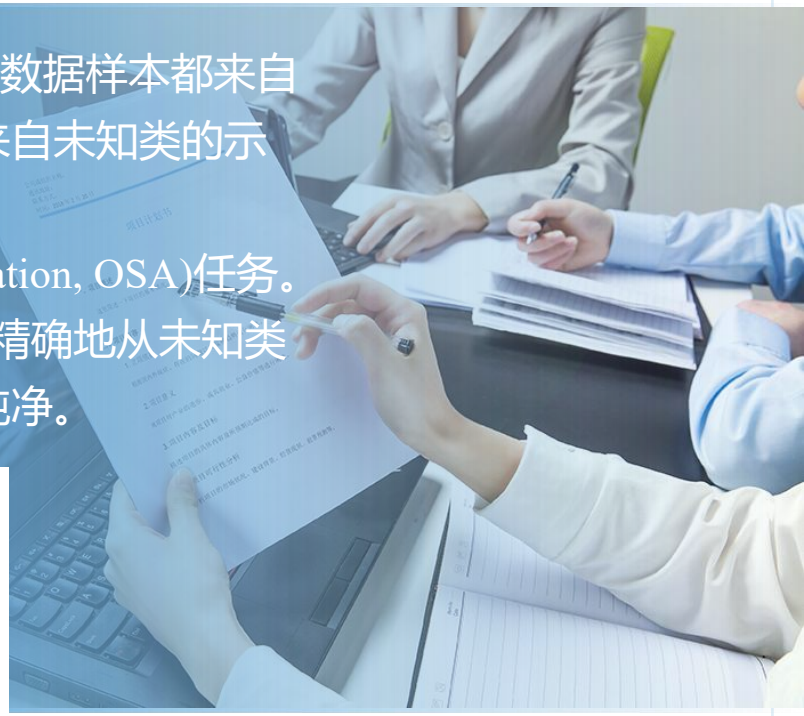
但是对于开集，现有的闭集AL系统无法准确地将这些不相关的图像与未知的类别区分开来,而是 倾向于选择它们进行标注,因为它们包含更多的不确定性或信息。

Background

现有的主动学习研究通常在闭集环境下工作假设所有要标记的数据样本都来自已知的类。然而,在真实的标注任务中,未标注的数据通常包含大量来自未知类的示例,导致大多数主动学习方法失败。

在本文中,我们将这个问题表述为一个开集标注 (open-set annotation, OSA)任务。如图所示,未标记集 包含K个已知类和L个未知类,其中L>K。目标是精确地从未知类中过滤出样本,同时主动选择一个包含已知类样本的查询集,尽可能纯净。





Method

LfOSA包括两个分别用于检测和分类的网络。具体来说,检测器使用高斯混合模型对每类最大激活值(MAV)分布建模,将未标记的开放集动态划分为已知和未知集,然后从已知集中选择具有较大确定性的样本构建用于标注的查询集。标记完成后, 分类模型将使用来自已知类的新样本进行更新。同时, 由于查询集不可避免地会包含一些未知类的无效样例, 这些无效样例将被用作负训练样例来更新检测器。

此外,通过降低交叉熵(CE)损失的温度T, 进一步增强检测器的可分辨性。

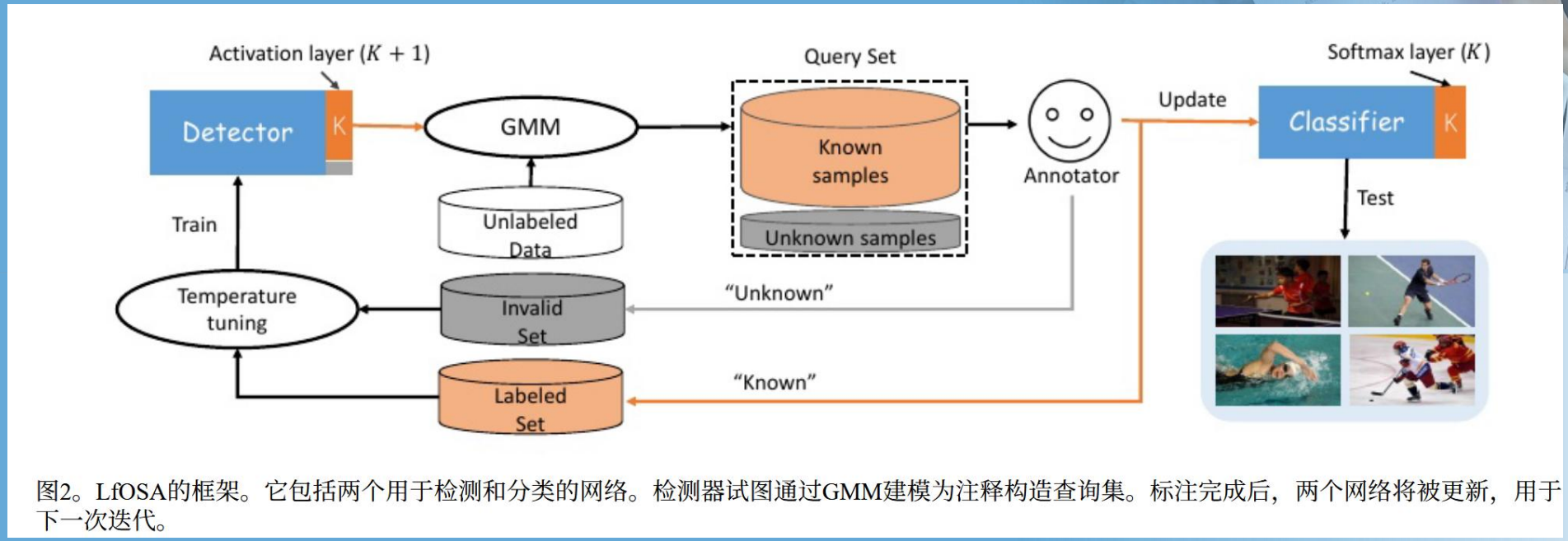


图2。LfOSA的框架。它包括两个用于检测和分类的网络。检测器试图通过GMM建模为注释构造查询集。标注完成后，两个网络将被更新，用于下一次迭代。

Method

检测器训练、主动采样和分类器训练三部分组成。具体来说,我们首先在使用低温机制的同时,通过利用已知和未知监督来训练一个用于检测未知示例的网络。然后,通过使用高斯混合模型(GMM)建模每个类的最大激活值(MAV)分布, 可以主动选择最确定的已知示例进行注释。最后,分类模型将使用来自已知类的新样本进行更新。

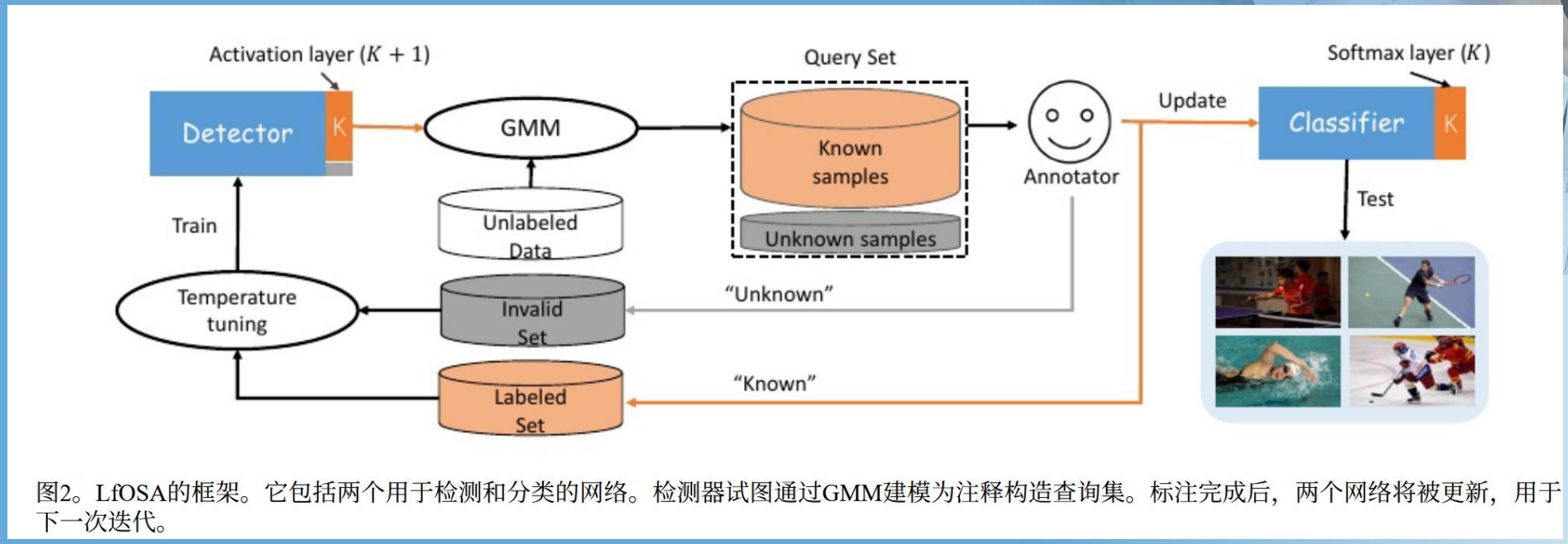


图2。LfOSA的框架。它包括两个用于检测和分类的网络。检测器试图通过GMM建模为注释构造查询集。标注完成后，两个网络将被更新，用于下一次迭代。

Method

检测器训练:

检测器是一个 $(K+1)$-way 分类网络，训练数据包括：

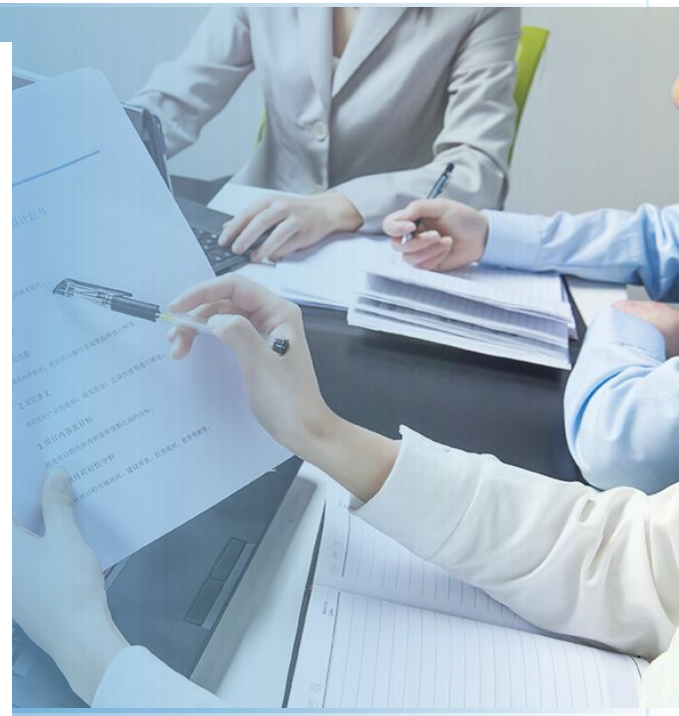
- **正样本**：初始已标注集 $\mathcal{D}_L$（仅含已知类）；
- **负样本**：无效集 $\mathcal{D}_I$（历史查询中被发现的未知类样本）。

损失函数（低温交叉熵）：

$$\mathcal{L}_D(x,c) = -\sum_{c=1}^{K+1} p_c \log(q_c^T)$$

其中：

- p_c 是 one-hot 标签（已知类标签为对应位置 1，未知类标签为第 $K+1$ 位为 1）；
- $q_c^T = \frac{\exp(a_c/T)}{\sum_j \exp(a_j/T)}$ 是温度 T 调节后的 softmax；
- **关键设置**：$T = 0.5 < 1$（低温）。



$$\frac{\partial \mathcal{L}_D}{\partial a_c} = \frac{1}{T}(q_c^T - p_c) = \frac{1}{T}(\frac{exp(a_c/T)}{\sum_j exp(a_j/T)} - p_c). \quad (4)$$

$$T \downarrow \Rightarrow \frac{1}{T} \uparrow, \frac{exp(a_c/T)}{\sum_j exp(a_j/T)} - p_c \uparrow \Rightarrow \frac{\partial \mathcal{L}_D}{\partial a_c} \uparrow .$$

Method

检测器训练:

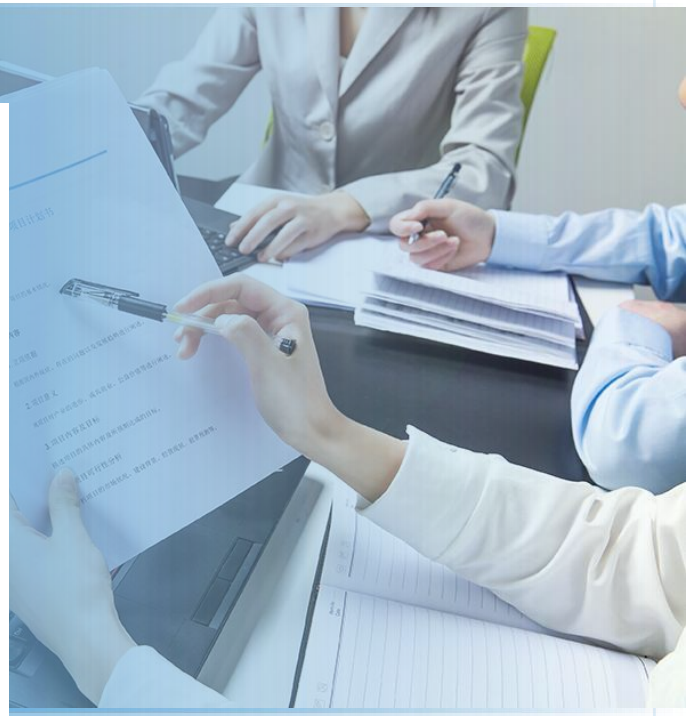
✅ **对于正确类（$c = y$）：**

- $\frac{\partial \mathcal{L}_D}{\partial a_y} = \frac{1}{T}(q_y^T - 1) < 0$
- 梯度绝对值变大 → 在 SGD 更新中，a_y 增加更多；
- 即：**模型更强烈地“拉高”正确类的激活值。**

✅ **对于错误类（$c \neq y$）：**

- $\frac{\partial \mathcal{L}_D}{\partial a_c} = \frac{1}{T}(q_c^T - 0) > 0$
- 梯度变大 → a_c 被压得更低；
- 即：**模型更强烈地“压制”错误类的激活值。**

👉 所以，**梯度变大 → 正确类 logit 上升更快，错误类下降更快 → 分布更尖锐。**



$$\frac{\partial \mathcal{L}_D}{\partial a_c} = \frac{1}{T}(q_c^T - p_c) = \frac{1}{T}(\frac{exp(a_c/T)}{\sum_j exp(a_j/T)} - p_c). \quad (4)$$

$$T \downarrow \Rightarrow \frac{1}{T} \uparrow, \frac{exp(a_c/T)}{\sum_j exp(a_j/T)} - p_c \uparrow \Rightarrow \frac{\partial \mathcal{L}_D}{\partial a_c} \uparrow .$$

Method

主动采样:

如上所示 对检测器进行训练后,我们发现网络的激活层（就是取最大的logits）具有区分未知样例的能力,即未知类样例的最大激活值(MAV)往往与已知类样例的平均MAV存在显著差异。

$$mav_i^c = \max_c a_c^i.$$

第一句："根据当前检测器的预测，将所有未标记的样例分为 $K+1$ 个类。"

- 检测器输出 $K+1$ 维 logits;
- 使用 argmax 得到预测类别 $\hat{y}_i \in \{1, 2, \ldots, K+1\}$;
- 所以每个样本被分配到一个"预测类";
- 其中:
  - $\hat{y}_i \leq K \rightarrow$ 预测为某个已知类;
  - $\hat{y}_i = K+1 \rightarrow$ 预测为"未知"。

目标：**从这些"预测为已知类"的样本中，选出真正属于已知类的高质量样本进行标注。**

- **过滤**：只保留那些预测为 $\hat{y}_i \in \{1, \ldots, K\}$ 的样本;
- 这些样本被认为是"可能属于已知类"的候选;
- 不考虑预测为"未知"的样本（因为它们可能是新类，不参与本轮标注）。

筛选后得到一个集合：$\mathcal{X}_{\text{known-predicted}}$

$$\mathcal{W}^c = GMM(mav^c, \theta_D), \quad (6)$$

$$\mathcal{W} = sort(\mathcal{W}^1 \cup \mathcal{W}^2 \cup \ldots \cup \mathcal{W}^K). \quad (7)$$

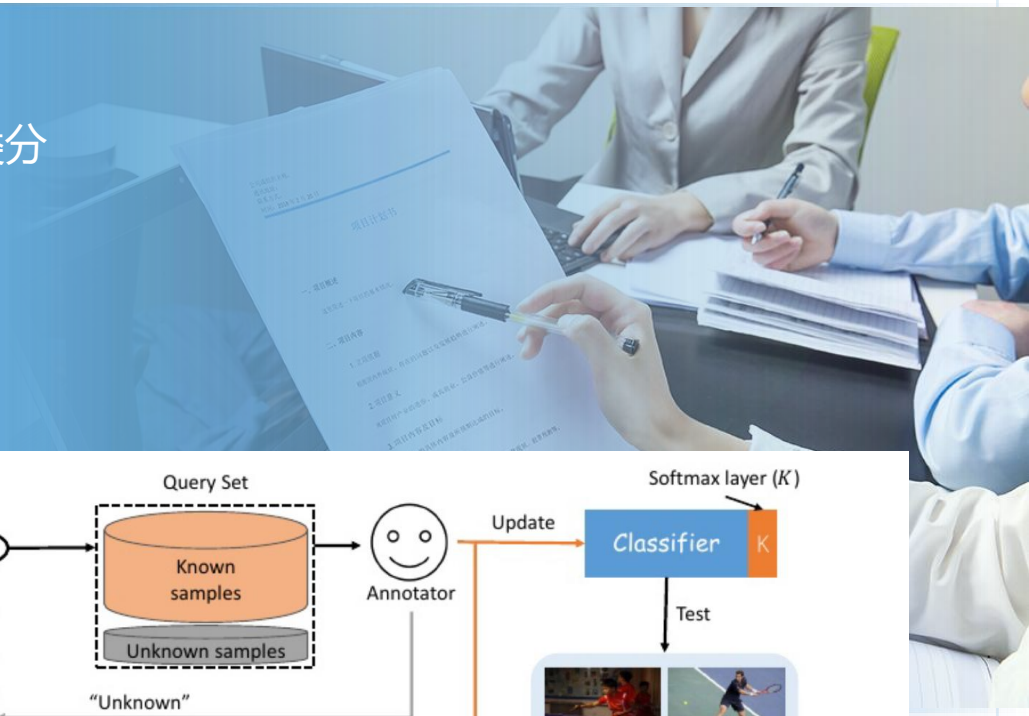
$$X^{query} = \{(x_i, w_i) | w_i \geq \tau, \forall (x_i, w_i) \in (D_U, \mathcal{W})\}. \quad (8)$$

Method

分类器训练:

基于当前标注的数据DL,我们通过最 小化标准交叉熵损失来训练k类分类器

$$\mathcal{L}_C(x_i, y_i) = -\sum_{i=1}^{n^L} y_i * log(f(x_i; \theta_C)), \qquad (9)$$



步骤	目标	方法
检测器训练	学会区分已知/未知	使用 D_L 和 D_I 联合训练
MAV 计算	量化模型置信度	取前 K 类 logit 的最大值
GMM 建模	区分“真已知”和“伪已知”	双分量 GMM 拟合 MAV 分布
概率排序	优先标注高质量样本	按 w_i 排序，选 top-b
分类器训练	提升已知类精度	仅用 D_L 训练标准分类器

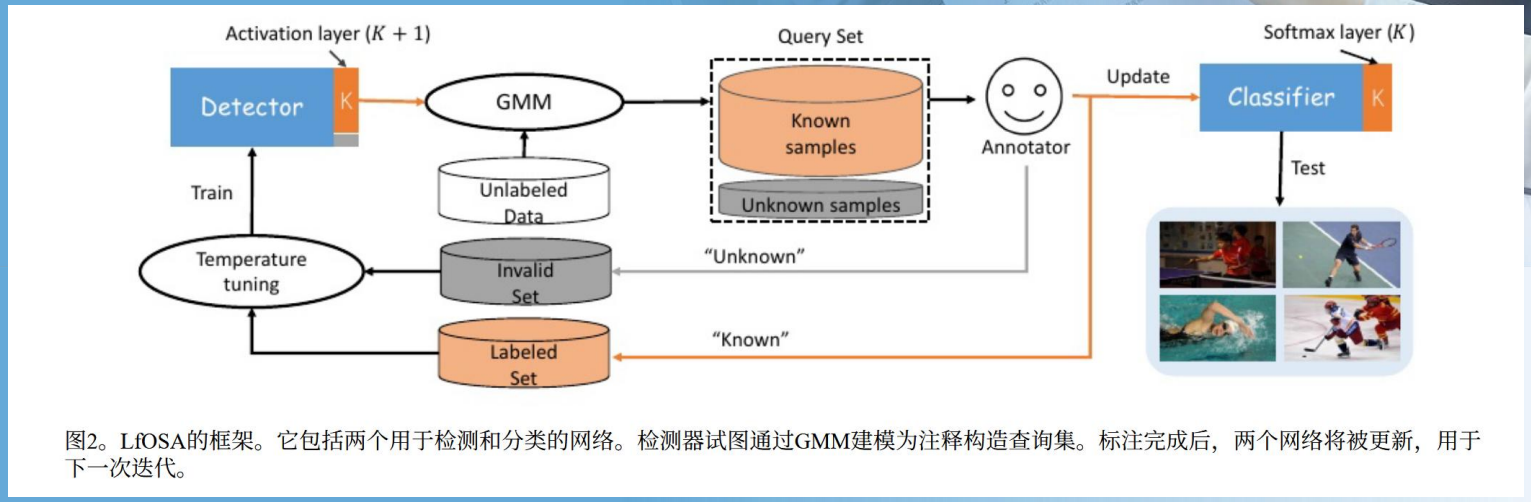


图2。LfOSA的框架。它包括两个用于检测和分类的网络。检测器试图通过GMM建模为注释构造查询集。标注完成后，两个网络将被更新，用于下一次迭代。

Experiments

为了验证该方法的有效性,我们在CIFAR10、 CIFAR100和Tiny-Imagenet数据集上进行了实验,这些数据集分别包含10个、100个和200个类别。为了 构建开集数据集,我们将所有实验的错配率分别设置 为20%、30%和40%,其中错配率表示已知类别数量占 总类别数量的比例。例如,当不匹配率设置为20%时, 在CIFAR10、CIFAR100和Tiny-Imagenet上,前2、20、 40个类分别被视为分类器训练的已知类,后8、80、 160个类分别被视为未知类。

维度	CIFAR-10	CIFAR-100	Tiny-ImageNet
语义粒度	粗粒度（如“猫”、“车”）	细粒度（如“枫树” vs “橡树”，“海狸” vs “松鼠”）	更细粒度（200 个 ImageNet 子类）
类间相似性	较低（类别差异大）	较高（易混淆，如不同鸟类）	高（许多动物/物体外观接近）
图像分辨率	极低（32×32）	极低（32×32）	较低（64×64），但仍比 CIFAR 清晰
数据多样性	有限（简单背景、小目标）	更丰富（但受限于分辨率）	更真实（来自 ImageNet，含复杂背景）
任务难度	简单（闭集准确率 >95%）	中等（~75–80%）	较难（~60–70%）

Experiments

第i次选择中已知类的查全率和查准率可以定义为:

$$recall_i = \frac{\sum_{j=0}^{i} k^i}{n_{kno}}, \tag{1}$$

$$precision_i = \frac{k_i}{k_i + l_i}, \tag{2}$$

方法	是否考虑未知类?	主要策略	开放集适用性	缺点
Random	❌	随机	差	效率低
Uncertainty	❌	选最不确定	差	易选未知类
Certainty	⚠️（隐式）	选最确定	中	可能 overconfident，缺乏探索
Coreset	❌	选多样本	差	会选代表性未知类
BALD	❌	贝叶斯不确定性	中	计算贵，仍可能选未知类
OpenMax	✅	开集识别 + 过滤	好	未结合主动学习信息量
LiFOSA	✅✅	GMM 建模 MAV + 概率排序	**优秀**	实现稍复杂

Experiments

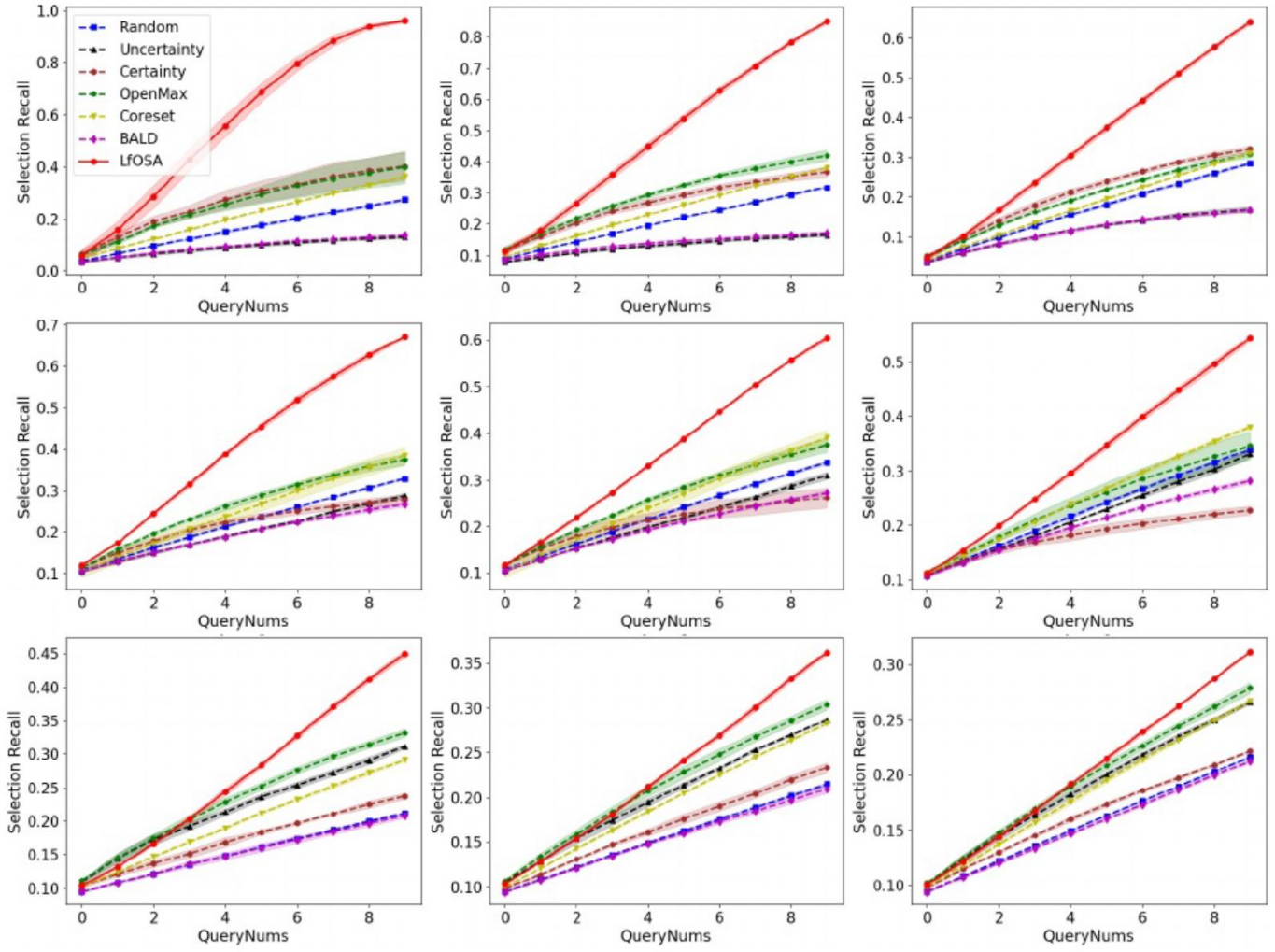


图3。CIFAR10(第一行)、CIFAR100(第二行)和Tiny-Imagenet(第三行)上20%(第一列)、30%(第二列)和40%(第三列)错配率的选择召回比较。

Experiments

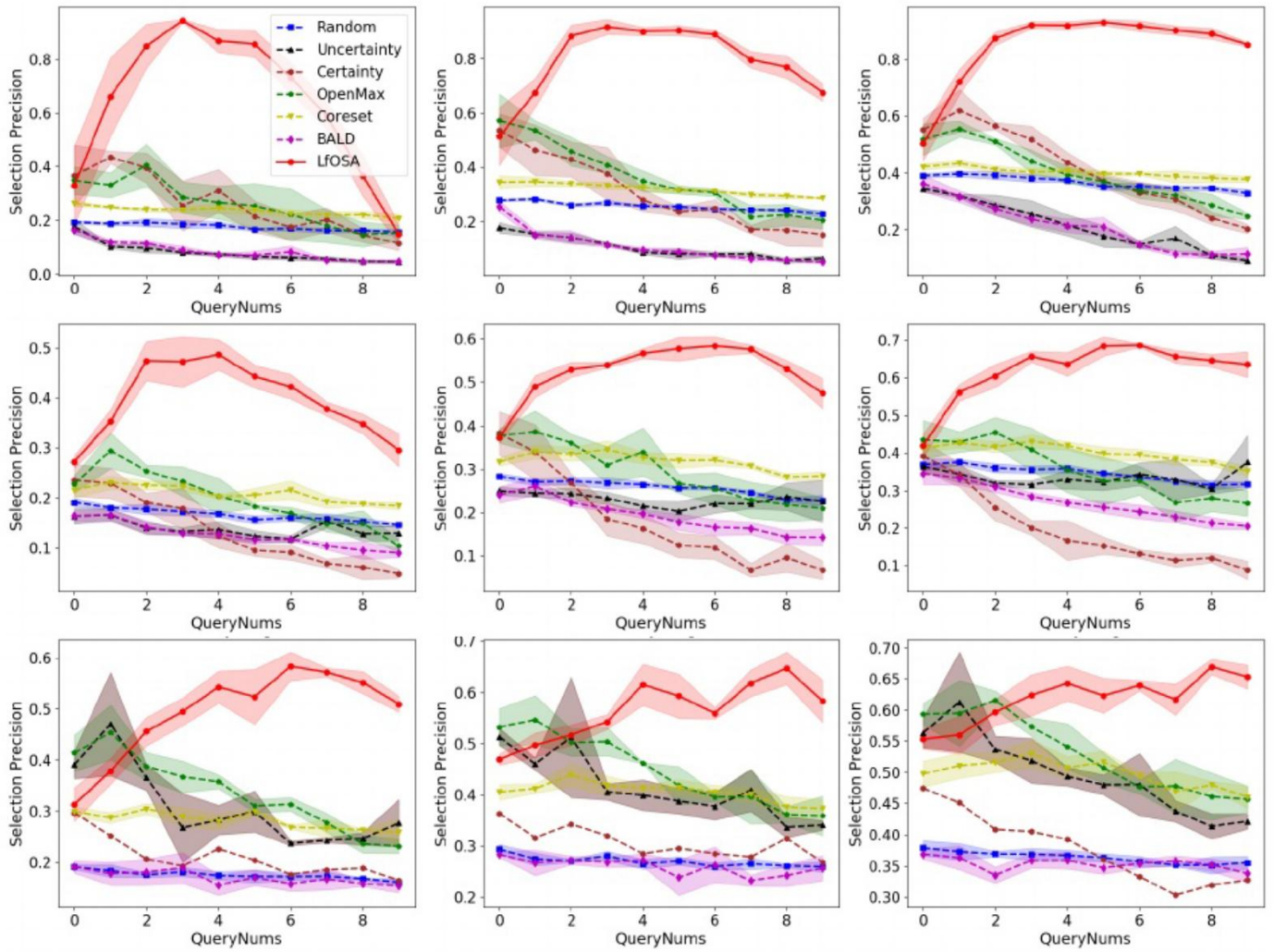
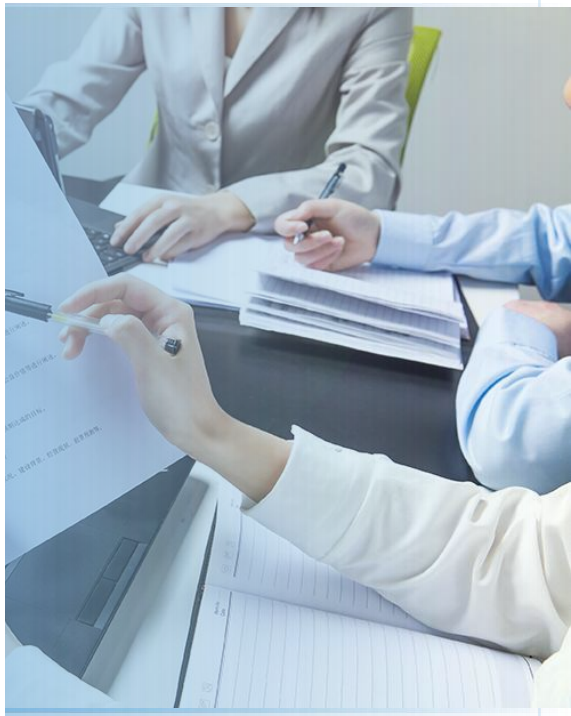


图4。CIFAR10(第一行)、CIFAR100(第二行)和Tiny-Imagenet(第三行)在20%(第一列)、30%(第二列)和40%(第三列)错配率下的选择精度比较。



Experiments

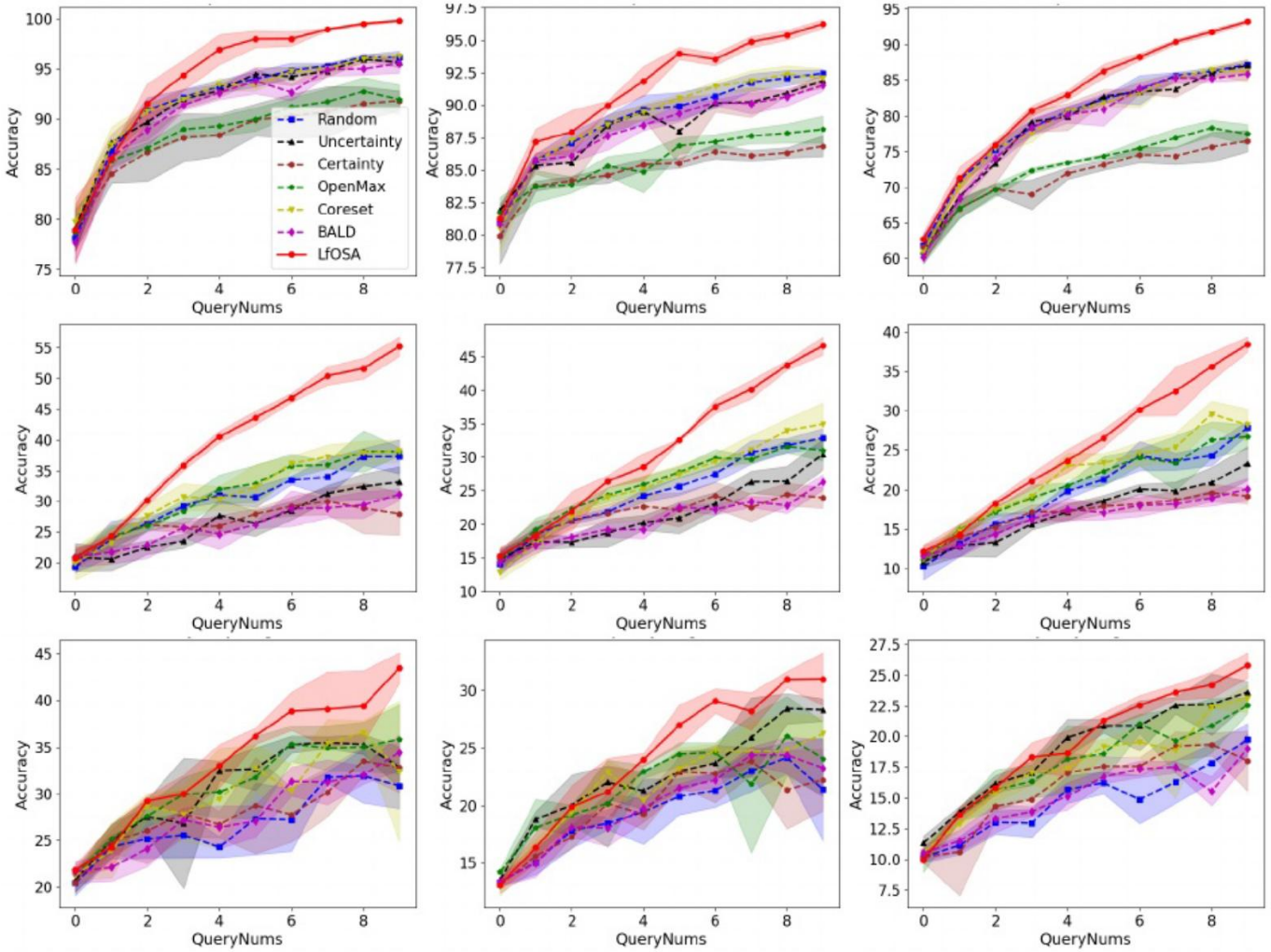
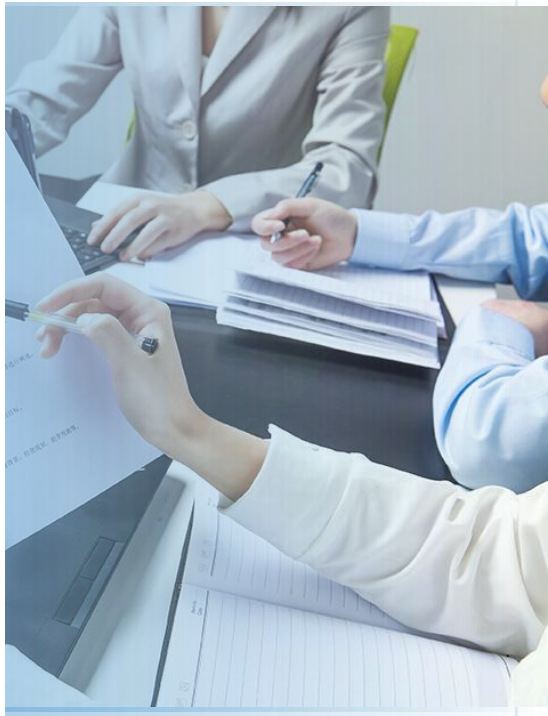


图5。CIFAR10(第一行)、CIFAR100(第二行)和Tiny-Imagenet(第三行)在20%(第一列)、30%(第二列)和40%(第三列)错配率下的分类性能比较。



Experiments

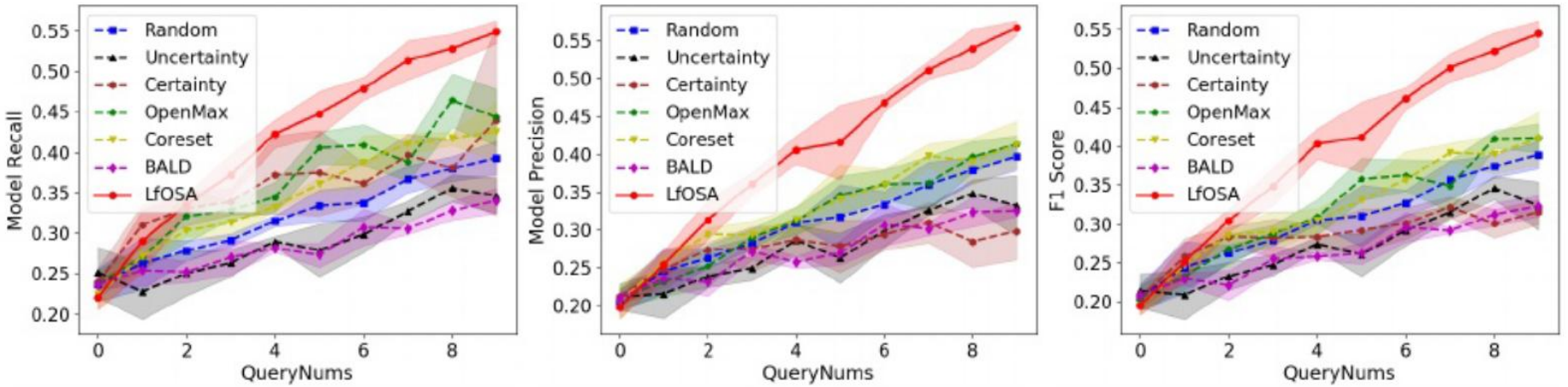


图6.。分类查全率(第一列)、查准率(第二列)、F1(第三列)在20%错配率CIFAR100上的性能比较。

表1。时间复杂度的比较。

Random	Certainty	Openmax	Coreset	BALD	LfOSA
$\sim 0s$	$23s$	$29s$	$165s$	$182s$	$26s$

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

变体	含义	性能影响
✅ **LfOSA（红色实线）**	完整方法：检测器 + 分类器 + GMM + 低温 softmax + 无效集训练	✅ 最优性能，全程领先
❌ **w/o Detector（绿色虚线）**	不使用检测器，直接用分类器做开放集识别	⬇️ 显著下降 → 检测器不可替代
❌ **w/o Classifier（蓝色虚线）**	不使用分类器，只用检测器输出做分类	⬇️ 大幅下降 → 分类器是必需的
❌ **w/o invalid set（紫色菱形）**	检测器训练时不使用无效集 D_I	⬇️ 中等下降 → 无效集提供关键监督信号
❌ **high temperature（黄色虚线）**	温度 $T = 2$（softmax 更平滑）	⬇️ 下降 → 高温削弱判别力
❌ **w/o temperature（黑色三角）**	温度 $T = 1$（标准 softmax）	⬇️ 下降 → 低温增强判别性
❌ **Dichotomies（棕色方块）**	用二分类（known/unknown）代替多类检测器	⬇️ 下降 → 多类建模更优

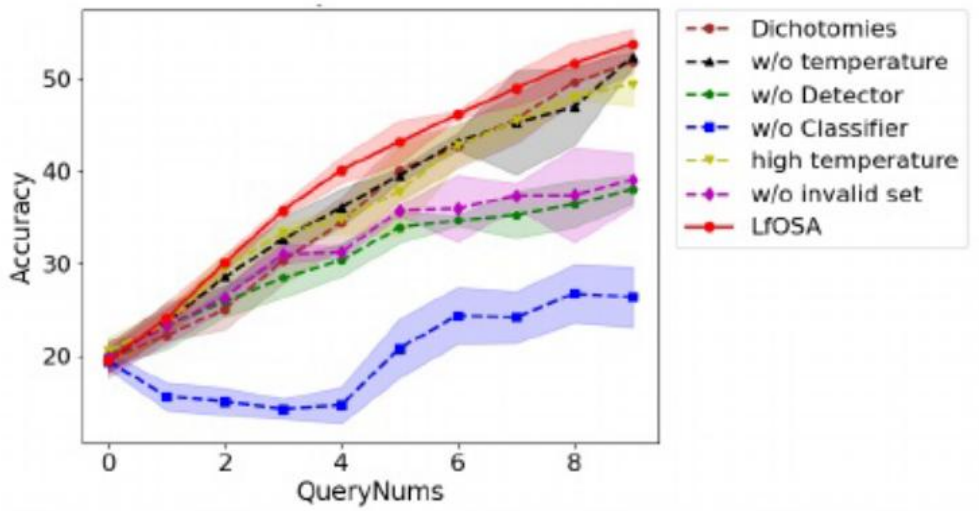


图9。LfOSA各组分对20%失配率下CIFAR100的影响。

Thanks