arXiv:2303.01806v2 [cs.LG] 1 Jun 2023

特权信息何时能解释消除标签噪声?

Guillermo Ortiz-Jimenez [*1] **Mark Collier** [*2] **Anant Nawalgaria** [2] **Alexander D'Amour** [2] **Jesse Berent** [2] **Rodolphe Jenatton** [2] **Effrosyni Kokiopoulou** [2]

摘要

利用特权信息（PI）——即训练时可用、但测试时不可用的特征——近期已被证明是应对标签噪声的有效方法。然而，其有效性的原因尚未得到充分理解。本研究探讨 PI 的不同属性在解释消除标签噪声方面所发挥的作用。通过在多个包含真实 PI 的数据集（CIFAR-N/H）和一个新的大规模基准测试 ImageNet-PI 上开展实验，我们发现，当 PI 能让网络轻松区分干净数据与错误标记数据，同时提供一种学习捷径来记忆错误标注样本时，PI 最有帮助。有趣的是，当 PI 对目标标签的预测能力过强时，PI 方法的表现往往不如无特权信息基线。基于这些发现，我们提出了对最先进的特权信息方法的若干改进，并展示了 PI 在应对标签噪声方面的潜力。最后，我们展示了如何轻松地将由此得到的 PI 方法与现有的、旨在处理标签噪声的无 PI 技术相结合。

1. 引言

标签噪声，即训练数据中的错误标签，是机器学习中普遍存在的问题。随着我们使用更多弱标注数据训练更大的模型，这一问题也日益常见。人工标注者往往是这类噪声的来源，他们会为某些样本赋予错误标签（Snow 等人，2008；Sheng 等人，2008），例如类别划分过于细致时。错误标注也可能源于使用其他模型提供代理标签（Prabhu 等人，2022），或从网络抓取数据（Radford 等人，2021）。然而，监督学习中的标准做法是忽略这一问题，并将所有标签视为正确，这会导致模型性能显著下降，因为模型往往会记住噪声标签，并使学习到的表征退化（Zhang 等人，2017）。

最近，一些研究提出利用特权信息（PI）来减轻标签噪声的影响（Vapnik 和 Vashist，2009年；Collier 等人，2022），也就是在训练时可用、但在测试时不可用的特征。特权信息的例子包括描述提供特定标签的人类标注者的特征，例如标注者 ID、提供标签所需的时间、标注者的经验等。尽管多种 PI 方法在减轻标签噪声的影响方面取得了令人鼓舞的进展（Lopez-Paz 等人，2016年；Lambert 等，2018；Collier 等人，2022），但其成功背后的原因尚未得到充分理解。此外，现有结果是在异构设置下得出的，这使得比较和得出结论变得困难。

在本研究中，我们旨在规范特权信息的评估，并大规模研究特权信息在解释标签噪声方面的作用。我们考察了若干 PI 方法在不同噪声数据集上的表现，并根据可用特权信息的预测属性分析其行为。有趣的是，我们发现，当特权信息对目标标签的预测能力过强时，由于模型无法学习非特权输入特征与目标之间的关联，大多数 PI 方法的表现会显著低于无特权信息基线。相反，我们发现，最强形式的特权信息具有两个主要属性：（i）它使网络能够轻松区分干净样本和错误标注样本；（ii）它有助于形成一条更容易的"学习捷径"（其含义将在后文阐明），从而对错误标注样本进行过拟合（见图1）。当这两个属性同时具备时，PI 方法能够更好地抵御标签噪声，表现显著优于对应的无特权信息方法。

总体而言，我们观察到，在训练过程中使用特权信息可以使模型发现捷径，从而阻碍其学习特征与标签之间的关系（Geirhos 等人，2020年；D'Amour 等人，2020 年）。一方面，这些由特权信息促成的捷径可能产生积极作用：当被忽略的关系仅涉及错误标签时，它们

[*]贡献相同 [1]洛桑联邦理工学院（EPFL）。此项工作是在谷歌实习期间完成的。 [2]Google Research。通讯作者：Guillermo Ortiz-Jimenez <guillermo.ortizjimenez@epfl.ch>, MarkCollier<mark.collier@google.com>。

发表于 40th 国际机器学习会议，美国夏威夷檀香山。PMLR 202, 2023。版权所有 © 2023，作者所有。

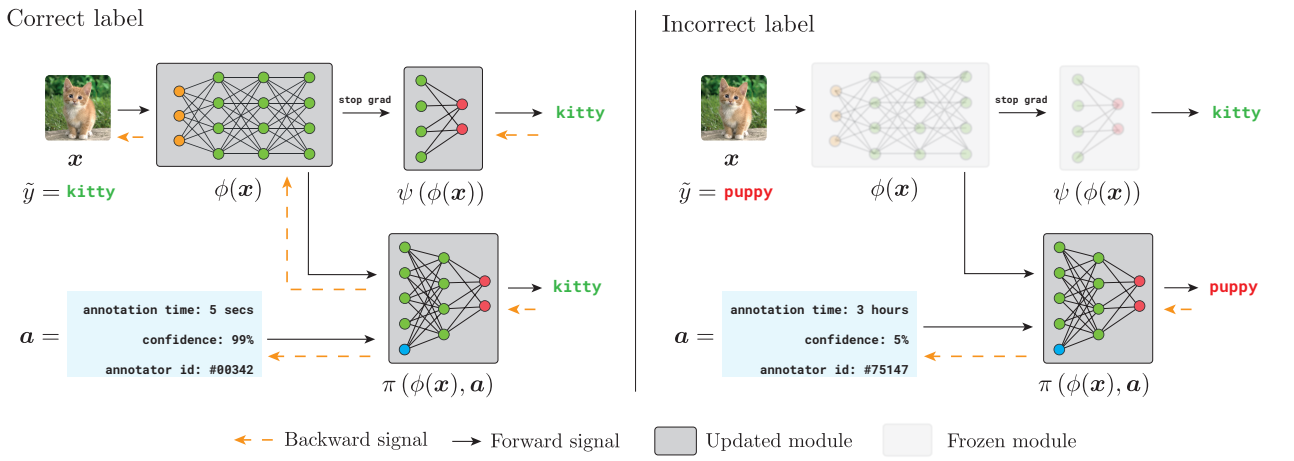

图1. 概念示意图：使用带噪声标签训练 TRAM（Collier 等人，2022）等特权信息方法时，理想的信号传播过程。访问特权信息后，网络可以利用学习捷径，仅凭特权信息记忆错误标注样本。这能保护从实际数据中提取特征的过程，而这些特征只会通过正确标记的样本得到优化。

可以防止学到的特征表示受到标签噪声的污染。另一方面，当干净标签也受到影响时，这些捷径可能产生负面作用，阻碍模型学习特征与目标之间的正确关联。捷径对于理解特权信息在缓解标签噪声中的作用至关重要，并且与深度学习动态直接相关（Zhang 等人，2017年）。

聚焦于最强 PI 方法的动态过程时，我们发现，使用 PI 可以在噪声水平更高的数据集上训练更大的模型。PI 能够抵消记忆特征与错误标签之间错误关联所带来的负面影响，因为它提供了一条主要影响错误标注样本的捷径。我们利用这些新见解改进了当前最先进的 PI 算法。

总体而言，我们的主要贡献如下：

- 我们首次大规模研究了 PI 在各种类型的带噪声监督数据集中的作用。
- 我们发布了 ImageNet-PI，这是目前规模最大的测试平台，可用于研究 PI 与标签噪声。
- 我们发现，有效的特权信息能够使模型仅在错误标记数据上学习捷径，从而显著提升性能。
- 我们通过简单的改进提升了多种 PI 方法的性能，并展示了这些方法与其他最先进的噪声标签方法相结合时能够带来的累积收益。

我们认为，我们的发现能够对未来关于标签噪声和特权信息的研究产生重大影响。它们不仅揭示了理想特权信息应具备的属性（这有助于设计和收集 PI 特征并将其应用于实践），还为改进现有方法提供了切实可行的见解。对我们的实证结果进行形式化刻画，是未来研究的另一个有前景的方向。

2. 方法

我们的大规模实验全面分析了存在标签噪声时特权信息的效用。以往的研究结果是在异质设置下得出的，测试所用的数据集和特权信息类型各不相同，因此难以进行公平比较。我们的目标是统一这些比较，并将统一实验中的所有设置作为核心贡献的一部分。我们的代码可在 https://github.com/google/uncertainty-baselines 获取。接下来，我们将简要介绍本研究中使用的数据集和基线方法。

2.1. 数据集

在本文中，我们研究 Collier 等人（2022）所描述的带有特权信息和标签噪声的监督学习设置。我们的训练数据由三元组（$\boldsymbol{x}$，$\tilde{y}$，$\boldsymbol{a}$）组成，其中，$\boldsymbol{x} \in \mathbb{R}^d$ 是一组输入特征，$\tilde{y} \in \{1, \ldots, K\}$ 是噪声目标标签（假设有 K 个类别），$\boldsymbol{a} \in \mathbb{R}^p$ 是特权信息特征向量。在本文中，我们主要关注这些特权信息特征与标注过程相关的情况，因为这是标签噪声的常见来源（Snow 等人，2008；Sheng 等人，2008）。这些特权信息可能包括有关标注者的信息，例如其 ID 或经验；也可能包括有关过程本身的信息，例如标注时长或置信度。在测试时，我们无法获取任何特权信息，并且仅基于来自数据分布的干净（$\boldsymbol{x}$，y）对来评估我们的模型。

在实验中，我们使用经过重新标注的标准图像识别数据集，这些数据集提供了有关标注过程的各种形式的特权信息。这些数据集使我们能够获取干净标签（y）和噪声标签（$\tilde{y}$），但在训练和超参数选择中，我们只使用噪声标签（有关此步骤中噪声标签影响的讨论，详见附录 D）。干净标签仅用于评估。我们所使用的数据集涵盖了多种训练条件，包括不同数量的样本和类别、不同程度的噪声、不同类型的特权信息、图像尺寸以及标注过程，因此我们的发现具有广泛的适用性。

其中一些数据集为每个样本提供多重标注。尽管如此，为了创建统一基准，我们对于提供多个标签的数据集，每个样本只抽取一个标签。为了控制噪声水平并考察其对 PI 方法性能的影响，我们尽可能为每个数据集创建高噪声和低噪声版本。按照 Wei 等人（2022年）的术语，我们将低噪声版本称为“均匀”版本，即从可用标签中均匀随机地选取一个；将高噪声版本称为“最差”版本，即只要有错误标签可选，就总是选取错误标签。按照设计，“最差”版本的噪声比“均匀”版本更大。

CIFAR-10/100N。 这是 CIFAR-10/100 数据集（Krizhevsky，2009年）的重新标注版本，每张图像都包含多重标注（Wei 等人，2022年）。原始数据包含有关标注过程的信息，例如标注时间和标注者 ID，但这些信息并未提供到样本级别，而是以样本批次的平均值形式提供，因此属于粗粒度特权信息。我们将展示 PI 基线方法在该数据集上的表现不佳。CIFAR-10N 的“均匀”版本有 82.6% 的时间与干净标签一致，而“最差”版本为 59.8%。CIFAR-100N 与干净标签的一致率为 59.8%。作为参考，在没有标签噪声且不使用 PI 的情况下训练我们的基础架构，在 CIFAR-10 和 CIFAR-100 上分别达到 93.5% 和 77.9% 的测试准确率。

CIFAR-10H。CIFAR-10 的另一种由人工重新标注的版本，新标签仅提供于测试集（Peterson 等人，2019年）。与 Collier 等人（2022）一样，在 CIFAR-10H 上训练时，我们在原始 CIFAR-10训练集上评估模型的性能（因为 CIFAR-10H 仅重新标注了验证集））与 CIFAR-N 数据集不同，CIFAR-10H 包含丰富的样本级特权信息，并带有关于标注过程的高质量元数据。“均匀”版本与干净标签的一致率为 95。1%，而“最差”版本为 35。4%。作为参考，我们在没有标签噪声和特权信息的情况下训练基础架构，其在 CIFAR-10H 上的测试准确率为 88。4%。

ImageNet-PI。受 Collier 等人（2022）启发，这是 ImageNet（Deng 等人，2009年）的重新标注版本，标签由一组架构各异的预训练深度神经网络提供。在重新标注过程中，我们从每个模型针对每个样本生成的温度缩放预测分布中抽取一个随机标签。这会产生非对称且依赖特征的标签噪声。重新标注过程和温度缩放的技术细节见附录 A。该数据集的特权信息来自模型对抽样标签的置信度、模型的参数数量，以及模型在干净测试分布上的测试准确率。这些特权信息特征可作为各模型预期可靠性的良好代理。我们使用的 ImageNet-PI 高噪声版本与干净标签的一致率为 16。2%，低噪声版本则为 51。9%。作为参考，我们在没有标签噪声和特权信息的情况下训练基础架构，其在 ImageNet 上的测试准确率为 76。2%。本研究的一项贡献是开放源代码发布 ImageNet-PI（包含不同程度的标签噪声），以推动对特权信息和标签噪声的进一步研究，其规模超越了当前 CIFAR-N/H 所能实现的范围。数据可在 https://github.com/google-research-datasets/imagenet_pi 公开获取。

2.2. PI 算法

我们研究了四种利用特权信息的代表性方法的性能。这些方法都已被证明能够有效减轻标签噪声的影响（Collier 等人，2022年）：

无 PI。 一种标准的监督学习基线，在无法访问特权信息的情况下，最小化噪声标签上的交叉熵损失，以近似 $p(\tilde{y}|\boldsymbol{x})$。

蒸馏（Lopez-Paz 等，2016年）。 一种知识蒸馏方法：首先利用特权信息，通过标准的最大似然估计训练教师模型，以近似 $p(\tilde{y}|\boldsymbol{x},\ \boldsymbol{a})$。随后，在无法访问特权信息的情况下，使用相同架构训练学生模型，使其匹配教师模型的输出。我们还提供了标准自蒸馏基线的结果，其中教师模型无法访问特权信息（Hinton 等人，2015年）。

TRAM（Collier 等人，2022年）。 一种基于双头模型的方法，其中一个输出头可以访问特权信息，而另一个不能。在训练时，将共同的特征表示 $\phi(\boldsymbol{x})$ 输入两个分类头 $\pi(\phi(\boldsymbol{x})\ \boldsymbol{a})$（“PI 头”）和 $\psi(\phi(\boldsymbol{x}))$（“无 PI 头”），以联合解决

$$\min_{\phi,\pi,\psi} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{a},\tilde{y})} \left[\ell(\pi(\phi(\boldsymbol{x}),\boldsymbol{a}),\tilde{y}) + \ell(\psi(\phi(\boldsymbol{x})),\tilde{y})\right]. \quad (1)$$

重要的是，在训练期间，无 PI 特征提取器 ϕ仅使用来自 PI 头的梯度进行更新。在测试时，仅使用无 PI 头进行预测。

近似完全边缘化（Collier 等人，2022年）。首先使用最大似然估计训练一个神经网络，并利用 PI 来近似 $p(\tilde{y}|\boldsymbol{x},\boldsymbol{a})$。推理期间，使用蒙特卡洛估计来近似边缘

$$p(\tilde{y}|\boldsymbol{x}) = \int p(\tilde{y}|\boldsymbol{x},\boldsymbol{a})p(\boldsymbol{a}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{a}$$

通常还会进一步假设 $p(\boldsymbol{a}|\boldsymbol{x}) \approx p(\boldsymbol{a})$ 相互独立。请注意，这一过程会增加推理时的内存和计算开销，因为它需要针对 $\boldsymbol{a}$ 的每个不同采样值计算网络的输出（实践中，通常需要额外采样约 1，000 个值）。

所有方法都使用相同的底层网络架构，只做少量修改以满足各自的特定需求，这与 Collier 等人（2022）的方法类似。具体而言，在推理阶段，除近似完全边缘化（AFM）基线方法外，所有方法都使用完全相同的网络；AFM 需要额外的参数来处理采样得到的特权信息。

在所有实验中，我们都采用相同的流程，使用噪声标签进行训练，并在干净测试集上评估。由于早停是应对标签噪声的强基线（Bai 等人，2021），我们始终报告在留出的噪声训练集验证比例上取得最佳性能的轮次结束时的测试准确率。我们在附录 D 中复现了不使用早停的结果。除非另有说明，我们都会通过网格搜索为每种方法在每项实验中调整最重要的超参数，并报告五次运行的平均测试准确率和标准差。实验设置的更多细节见附录 B。

3. PI 何时有帮助?

表1（原始）展示了不同 PI 算法在我们收集的噪声数据集上的表现[1]，，可以看出，利用 PI 并不总能大幅提升性能。事实上，TRAM 和 AFM 虽然在 CIFAR-10H 和 ImageNet-PI 上显著优于无特权信息基线，但在 CIFAR-10N 和 CIFAR-100N 上的表现并没有好很多。此外，与普通的自蒸馏基线相比，蒸馏（PI）的收益也很小。

同一算法在数据集上的表现差异——这些数据集的主要差异来源是可用的特权信息（PI），即 CIFAR-10N与 CIFAR-10H——表明利用 PI 并非总是有益。事实上，结果因 PI 的预测属性和噪声分布而大不相同。这引出了两个问题：i）“是什么让 *PI* 对这些算法有效？”以及 ii）“它们如何利用 *PI* 来解释消除标签噪声？”。

[1]正文中我们展示了高噪声条件下的结果。附录 C 中给出了表1在低噪声条件下的复现结果。

为回答这些问题，我们开展了一系列受控实验，使用不同的 PI 特征（包括真实特征和合成特征）训练 PI 方法。通过这些实验，我们旨在找出促使这些算法取得最佳表现的主要机制。

3.1. 完全预测性 PI

> **假设**：PI $\boldsymbol{a}$总是补充关于标签 $\tilde{y}$ 的信息，而这些信息包含在 $\boldsymbol{x}$.中

自然会认为，在 $\boldsymbol{x}$ 的基础上获知 $\boldsymbol{a}$ 有助于预测 $\tilde{y}$，从而改善监督学习的效果。然而，这种推理存在缺陷，因为它忽略了模型在推理时无法利用 $\boldsymbol{a}$。相反，正如我们将看到的，如果 $\boldsymbol{a}$ 对目标 $\tilde{y}$具有很强的预测能力，测试性能可能会严重下降。

我们通过在噪声数据集上重新训练算法来检验这一假设，但使用 $\boldsymbol{a} = \tilde{y}$，而非原始 PI 特征。也就是说，让算法能够访问完全预测性 PI。

> **发现**：当 $\boldsymbol{a}$ 能够完全预测 $\tilde{y}$时，大多数 PI 方法的表现都不如无特权信息基线。

如表1（标签）所示，在这种情况下，所有 PI 基线方法的表现都大幅下降。原因很简单：当 PI 对目标标签的信息过于充分时，模型会严重依赖 PI 来解释标签，因此不再积极学习 $\boldsymbol{x}$ 与$\tilde{y}$ 之间的关联，也无法学到有意义的特征表示。在这方面，我们可以看到，蒸馏（PI）的表现与蒸馏（无 PI）大致相同，而 TRAM 和 AFM 的测试准确率则非常低。

极具预测性的 PI 会损害这些算法的表现，这一事实凸显了其动态机制的一个关键要素：PI 可能促使模型学习捷径（D'Amour 等人，2020年；Geirhos 等人，2020年），从而阻碍模型学习 $\boldsymbol{x}$ 与 $\tilde{y}$之间的某些关联；这可能是因为更新 ϕ（$\boldsymbol{x}$）的梯度信号受到削弱（Pezeshki 等人，2021年）。这具有实际意义：它提醒我们，在训练 $\boldsymbol{a}$ 时，不应盲目附加任意复杂的元数据，因为这些元数据可能对目标标签具有很强的预测性。

3.2. 噪声指示器

> **假设**：特权信息之所以有帮助，是因为它能够将错误标记样本与正确样本区分开来。

我们发现，当 $\boldsymbol{a}$ 对 $\tilde{y}$的预测性过强时，PI 方法的表现会很差。现在，我们转而探讨 PI 有助于解释消除标签噪声的另一种假设：PI 特征可以帮助网络区分干净样本与错误标注样本。事实上，Collier 等人（2022）最初提出使用 PI 来应对标签噪声，其动机是标注者特征（例如置信度）可以充当

表*1.* 使用不同特征作为特权信息训练的几种方法的测试准确率（灰色 和 斜体 所示的基线方法不使用特权信息）。此处，原始表示数据集的标准特权信息，指示器 表示区分干净样本和噪声样本的二元信号，标签 表示独热编码标签，近似最优 表示一种合成特征：对于标注错误的样本，它提供标注者标签；否则提供零向量。**粗体数字** 表示所有特权信息特征中的显著最大值，其中显著性指 p 值 < 0.05。

		原始	指示器	标签	近似最优
CIFAR-10H（最差）	*无 PI*	55.0±1.5	55.0±1.5	55.0±1.5	55.0±1.5
	蒸馏（无 PI）	47.9±0.0	47.9±0.0	47.9±0.0	47.9±0.0
	TRAM	64.9±0.8	63.3±1.1	38.3±0.2	**67.8±0.2**
	近似 FM	64.0±0.6	66.7±2.1	29.5±0.5	**74.4±0.1**
	蒸馏（PI）	45.4±0.8	**49.9±0.7**	44.5±0.1	48.2±0.9
CIFAR-10N（最差）	*无 PI*	80.6±0.2	80.6±0.2	80.6±0.2	80.6±0.2
	蒸馏（无 PI）	80.4±0.0	80.4±0.0	80.4±0.0	80.4±0.0
	TRAM	80.5±0.5	87.9±0.4	48.9±0.2	**89.3±0.3**
	近似 FM	82.0±0.3	91.2±0.3	22.6±0.2	**92.0±0.1**
	蒸馏（PI）	80.2±0.3	80.1±0.3	**80.7±0.2**	80.2±0.3
CIFAR-100N	*无 PI*	60.4±0.5	60.4±0.5	60.4±0.5	60.4±0.5
	蒸馏（无 PI）	60.6±0.2	60.6±0.2	60.6±0.2	60.6±0.2
	TRAM	59.7±0.3	62.4±0.3	34.9±0.2	**67.4±0.3**
	近似 FM	60.0±0.2	66.4±0.2	20.1±0.3	**70.2±0.1**
	蒸馏（PI）	61.1±0.2	**61.8±0.3**	60.5±0.2	**61.5±0.3**
ImageNet-PI（高噪声）	*无 PI*	47.7±0.8	47.7±0.8	47.7±0.8	47.7±0.8
	蒸馏（无 PI）	50.2±0.8	50.2±0.8	50.2±0.8	50.2±0.8
	TRAM	53.3±0.5	53.6±0.5	41.0±0.7	**56.5±0.3**
	近似 FM	55.6±0.3	55.3±0.6	0.8±0.2	**58.3±0.1**
	蒸馏（PI）	**51.0±0.4**	**50.6±0.2**	39.0±4.6	27.5±22.7

识别错误标注样本的代理。直观地说，主要假设是：如果 PI 能够准确识别错误标注样本，那么它就应当发挥专家知识的作用，避免模型专注于难以处理的错误标注样本，转而只在正确且容易处理的样本上学习（Vapnik 与 Vashist，2009）。

虽然这个假设很自然，但此前尚未经过检验。不过，我们可以利用本研究中的数据集进行检验。请回想一下，我们可以获取所有训练样本的干净标签和噪声标签，因此能够合成一个指示信号$\mathbb{1}(\tilde{y} \neq y)$，当干净标签与噪声标签一致时取值为1，当二者不一致时取值为0。表1（指示器）展示了使用 $\boldsymbol{a} = \mathbb{1}(\tilde{y} \neq y)$ 进行训练的结果。

发现：一些 PI 方法使用原始 PI 时的表现优于使用预言机噪声指示器时的表现。

有趣的是，尽管我们发现指示器列上的性能通常高于原始列，但情况并非总是如此，有时指示器的表现不如原始特权信息（PI），或未能显著优于它（参见 CIFAR10H 和 ImageNet-PI 上的 AFM 和 TRAM）。这表明，PI 方法所做的不仅仅是利用噪声指示能力，PI。显然，如果即使将理想的噪声指示器信号$\mathbb{1}(\tilde{y} \neq y)$用作 PI，原始 PI 有时仍能胜过它，那么 PI 中必然还包含算法可加以利用、从而提升性能的其他信息。

3.3. 记忆动态发挥着重要作用

检查算法的训练动态有助于理解前述结果。例如，图2展示了在 CIFAR-10H 上使用不同 PI 特征的 TRAM 模型的测试准确率和训练准确率变化过程[2]。这原始 PI 带来的最终测试准确率高于噪声指示器。与此同时，使用标注者标签作为 PI 训练的模型似乎没有学到任何有用的东西。这些差异可以用模型通过各个 TRAM 输出头拟合错误标记样本和正确样本的速率来解释。

聚焦于特权信息头的训练准确率，图2（右列）解释了为什么将标签作为特权信息会损害测试性能[3]。标签模型很快就能在所有样本（错误标记的100% 和正确的）上获得训练准确率

[2]其他设置的结果见附录 E.1。[3]图2（右上）中，蓝线位于黄线后方。

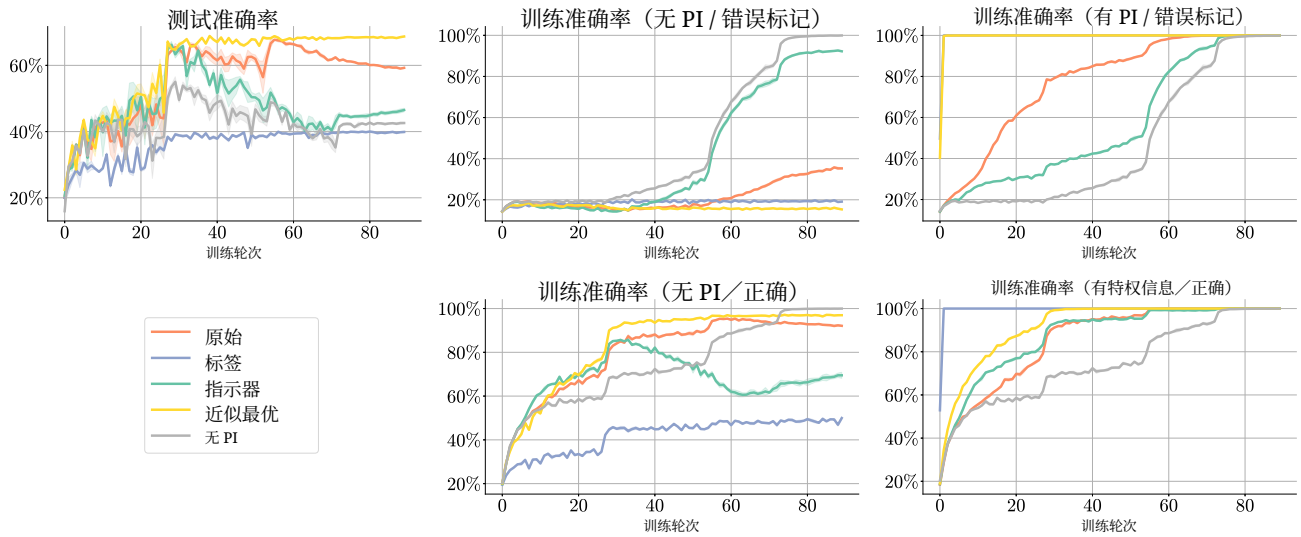

图*2*。不同特权信息下 TRAM 在 CIFAR-10H 上的动态表现。**左上**：测试准确率。**中上**：在无 PI 头上评估的错误标注样本训练准确率；**右上**：在 PI 头上评估的错误标注样本训练准确率。**中下**：在无 PI 头上评估的干净样本训练准确率。**右下**：在 PI 头上评估的干净样本训练准确率。

使用特权信息头，从而减慢无 PI 头的训练速度（中间列）。这是因为特征提取器只由来自特权信息头的梯度更新；如果模型仅使用 PI 特征来拟合所有样本，就会导致 $p(\tilde{y}|\boldsymbol{x})$ 缺乏有意义的表示。

聚焦于图2中无 PI 头的训练准确率（中间一列），最佳模型是在正确样本上达到最高训练准确率、同时不过拟合错误标注样本的模型。指示器与原始在测试表现上的差异，可以用原始模型更难对错误标注样本过拟合来解释。有趣的是，原始模型借助 PI 头记忆错误标注样本的速度，比指示器更快。看起来，借助 PI 头快速拟合训练样本，反而阻碍了模型通过无 PI 头拟合相同样本；换言之，PI 使模型能够利用原始PI 学习捷径来记忆错误标注样本，而无需使用 $\boldsymbol{x}$。这可能是因为指示器信号对所有样本都只取 $\{0,1\}$中的值，而这些值不足以区分噪声训练集。事实上，正如我们将看到的，能够获取可在错误标注样本上轻易记忆的 PI，对于最大化性能至关重要。

3.4. 近最优 PI 特征

假设：最优PI使学习捷径能够记住仅错误标记的标签。

使用标注者标签作为特权信息的实验，清楚地展示了一种由特权信息促成的学习捷径，而这种捷径对模型性能极为有害。另一方面，原始模型的动态特性表明，当同一种捷径机制仅应用于错误标注样本时，也可能产生积极效果。为了检验这一假设，我们设计了一种新的特权信息特征形式，在表格和图表中记作近似最优。顾名思义，这种特权信息应能使模型的性能非常接近其最佳水平。设计近似最优特征的目的，是仅在错误标注样本上利用特权信息捷径，让模型能够在正确标注的样本上自由学习。为此，近似最优特权信息特征由两个拼接的值组成：（i）一个指示信号，用于说明给定样本是否被错误标注；（ii）仅当该样本被错误标注时才提供的标注者标签；否则，则将一个与独热编码标签维度相同的全零向量与该指示信号拼接。

发现：仅当学习捷径仅提供给错误标注样本时，PI 方法才能取得最佳性能。

表1中的结果（近似最优）表明，在使用 TRAM 或 AFM 时，这些 PI 特征在所有数据集上的表现都大幅优于其他所有 PI 特征[4]。同样，在图2中，我们观察到近最优模型的动态表现完全符合预期。近似最优模型在 PI 头上的错误标注样本上训练得最快，因此在

[4]近似最优 并不总是优于 原始 在使用蒸馏（PI）时的表现，但请注意，总体而言，蒸馏（PI）相较于无 PI 的性能提升远小于 TRAM 和 AFM。鉴于此，我们将寻找蒸馏（PI）的近似最优策略这一目标留作未来工作的开放问题。

无 PI 头上的错误标注样本上，训练速度非常慢。此外，由于错误标注样本不再影响特征提取的更新（因为它们在 PI 头上的准确率已经达到最大值），我们观察到模型在正确样本上的表现要好得多。

同样的解释也适用于 AFM，其动态过程如附录 E.1 所示。在这种情况下，仅使用 PI 对错误标注样本进行记忆，也能保护无 PI 特征。这样，在推理过程中，从错误标注样本中采样的 PI 只会给错误标签的预测概率增加一个恒定的噪声底。这个平均噪声底通常远小于使用无 PI 的干净特征预测出的概率，因此不会导致频繁的误分类错误。

4. 改进 PI 算法

本节中，我们利用前述分析的见解来改进 PI 方法的设计。我们针对主要 PI 方法的不同设计参数进行了消融研究，并找出了能够显著提升其性能的简单改动。当 PI 有帮助时（参见表1），TRAM 和 AFM 的表现远优于蒸馏（PI），因此我们主要关注这两种方法。此处给出说明性结果，完整结果见附录 E。

4.1. 模型大小

我们探究模型大小如何影响性能。特别需要注意的是，所有 PI算法的参数数量可以拆分为两部分：用于标准特征的特征提取器 ϕ，以及处理 PI 的塔 $\boldsymbol{x}\ \pi$；见公式（1）和第 2.2 节。因此，我们进行了消融研究，分别对模型的这两部分进行缩放。

特征提取器。 图 3 展示了随着 PI方法中特征提取器大小增加，测试准确率的变化情况。性能呈 U 形，即模型规模超过某个点后，最终性能会受到损害。事实上，更大的容量会抑制模型使用 PI 特征，并导致其对标准特征过拟合，这一点表现为错误标注样本上的训练准确率同时上升，而测试准确率下降。

发现：增加特征提取器大小会抑制将 PI 用作捷径。

PI 头部大小。 图 4 展示了在固定特征提取器大小的同时，缩放 PI 处理塔大小的结果。我们观察到，更大的 PI 头部能够提升性能，因为它们促使模型仅利用 PI 进行更多记忆，并保护无 PI 特征的提取。这一点可以从以下现象看出：对于更大的 PI 头，

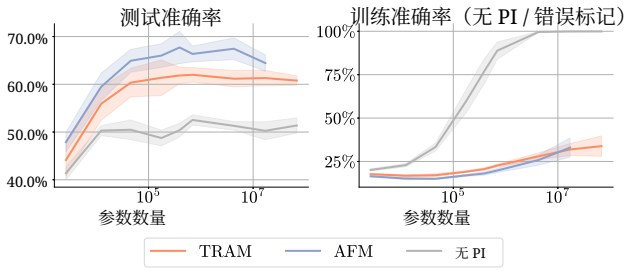

图 3。在固定 PI 塔的情况下，增加特征提取器的参数数量时，不同 PI基线在 CIFAR-10H 上的性能。更大的模型会出现过拟合，因为它们倾向于利用更大的容量对错误标注样本进行过拟合，从而不利于模型利用 PI。

错误标注样本在无 PI 头上的训练准确率下降。

发现：增加 PI 塔的容量会阻-
鼓励使用 PI 作为捷径。

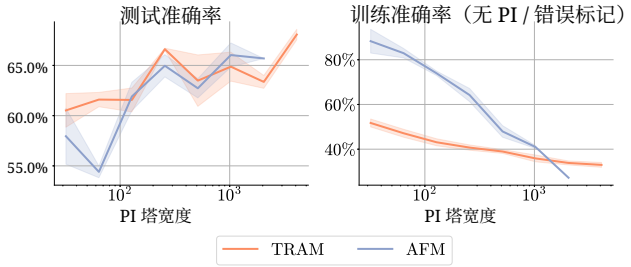

图 4。 增大 PI 头部大小时，不同 PI方法在 CIFAR-10H 上的性能。更大的 PI 头部大小会促使模型利用 PI 记忆错误标注样本，从而进一步将 PI 用作捷径。

4.2. 随机 PI 可以启用正向捷径

假设：能够唯一标识每个样本的随机 PI 可以启用一种 PI捷径，防止模型利用 $\boldsymbol{x}$记忆错误标签。

这些近似最优、标签和指示器信号均为表1中的合成 PI特征，在实际中无法使用，因为它们依赖于对哪些样本被错误标记、哪些样本正确的了解。然而，这些信号表明，若能获取一种比标准特征更容易被记住的信号 $\boldsymbol{x}$，并让它作用于错误标注样本，就能有效提升性能。话虽如此，错误标签的一个关键属性是，顾名思义，它们本身就是错误的。从这个意义上说，拟合一个错误的训练标签，不过是记住某个特定训练样本；该样本的特征无法预测目标标签，也就是说，这些特征仅仅充当样本标识。事实上，任何一组能够充分区分各个样本的特征，都可以充当这样的标识。

表 2。性能比较 比较无 PI、TRAM、TRAM++、SOP、HET、TRAM+SOP 和 TRAM+HET 在不同 PI 数据集上的表现 .

	无 PI	TRAM	TRAM++	SOP	TRAM+SOP	HET	TRAM+HET
CIFAR-10H （最差）	55.0±1.5	64.9±0.8	66.8±0.3	59.2±0.8	**70.9±0.5**	50.8±1.4	67.7±0.7
CIFAR-10N （最差）	80.6±0.2	80.5±0.5	83.9±0.2	87.9±0.2	**88.5±0.3**	81.9±0.4	82.0±0.3
CIFAR-100N	60.4±0.5	59.7±0.3	61.1±0.2	65.3±0.3	**66.1±0.2**	60.8±0.4	62.1±0.1
ImageNet-PI（高噪声）	47.7±0.8	53.3±0.5	53.9±0.4	–	–	51.5±0.6	**55.8±0.3**

发现：使用 $\boldsymbol{x}$，随机PI能有效减少对错误标签的过拟合。

我们在表 2 中评估这一假设，并在其中引入 TRAM++: 这是 TRAM 的一个版本，其中原始特权信息特征都加入了一个为每个样本单独生成的随机向量（实验细节见附录 F，AFM 的结果见++ 附录 G）。正如我们所见，TRAM++ 总体上优于单独使用 TRAM；在过拟合问题更严重的数据集（即 CIFAR）上，提升幅度更大。

5. 与其他无特权信息技术的结合

在本节中，我们通过实验表明，PI 方法在噪声数据集上取得的性能提升，可以与噪声标签文献中的其他最先进技术产生协同作用。具体而言，我们展示了 TRAM++ 可以轻松地与稀疏过参数化（SOP）（Liu 等，2022）和异方差输出层（Collier 等，2021）相结合，同时相较于这些基线方法[5]带来累积增益。

5.1. 稀疏过参数化（SOP）

稀疏过参数化（SOP）（Liu 等，2022）是一种最先进的方法，它利用随机梯度下降（SGD）的隐式偏置和过参数化来估计并校正噪声标签信号；这一概念已被证明具有良好效果（Zhao 等，2022）。该方法通过添加两组新的 K维参数$\{\boldsymbol{u}_i\}_{i=1}^N$ 和 $\{\boldsymbol{v}_i\}_{i=1}^N$来实现，其中 N 表示训练样本的数量，并求解

$$\min_{\boldsymbol{\theta},\{\boldsymbol{u}_i,\boldsymbol{v}_i\}_{i=1}^N} \frac{1}{N}\sum_{i=1}^{N} \ell\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{u}_i \odot \boldsymbol{u}_i - \boldsymbol{v}_i \odot \boldsymbol{v}_i, \tilde{y}\right)$$

使用 SGD。这种特定的参数化会使 SGD 的解偏向于恢复噪声信号$\boldsymbol{\epsilon}_i = \boldsymbol{u}_i \odot \boldsymbol{u}_i - \boldsymbol{v}_i \odot \boldsymbol{v}_i$，该信号会污染 y，即 $\tilde{y}_i \approx y_i + \boldsymbol{\epsilon}_i$，这隐含地假设 $\boldsymbol{\epsilon}_i$ 在整个数据集中是稀疏的。

在这项工作中，我们探讨 TRAM++ 与 SOP 结合能否在应对标签噪声方面带来累积收益。具体而言，我们提出一种简单的两步训练流程来结合二者：(i) 首先使用 TRAM++ 预训练神经网络；(ii) 然后使用 SOP 损失微调网络的无 PI 一侧，且不进行停止梯度操作。表 2 展示了该方法的结果[6]，我们发现，TRAM+SOP 确实在所有数据集上的表现都显著优于单独使用 TRAM++ 或 SOP。更多实验细节见附录 I。

5.2. 异方差输出层

最后，我们进一步分析了 TRAM 与 HET 的结合。HET 是噪声标签文献中另一种最先进的无特权信息基线，可扩展至 ImageNet 规模（Collier 等，2021）。这里的 HET 指的是使用异方差输出层，在无特权信息的情况下对预测的偶然不确定性进行建模。具体而言，我们将 HET 层应用于 TRAM 的两个输出头++，并沿用相同的训练设置。我们将由此得到的方法称为 TRAM+HET。

我们在表 2 中展示的实验表明，TRAM+HET 模型优于单独应用的 TRAM++和 HET。关于这一模型组合的更多实验细节，请参见附录 J。总而言之，这些结果印证了我们的主要发现：

发现：特权信息方法与噪声标签文献中的其他无特权信息算法协同发挥作用。

6. 相关工作

特权信息学习的一般框架（Vapnik 和 Vashist，2009年）已在深度学习领域得到广泛研究，许多研究探索了不同的基线方法，包括损失操纵（Yang 等，2017）、蒸馏（Lopez-Paz 等，2016）或高斯 dropout（Lambert 等，2018）。这条研究路线主要关注无噪声场景，将特权信息视为一种指导信号，用于识别容易或困难的样本（Vapnik 与 Izmailov，2015年）。与我们的工作类似，Yang 等（2022）也研究了特权信息在提升深度学习方法性能方面的作用，但他们关注的是在无噪声设置下使用蒸馏方法进行学习排序的任务。

[5]我们在附录 H 中提供了将 TRAM 与标签平滑（Szegedy 等，2016）结合的更多实验。

[6]我们没有提供 ImageNet-PI 的结果，因为 SOP 难以扩展到如此大规模的数据集。

最近，Collier 等人（2022年）提出了关于特权信息的一种新视角，认为它可以提高模型对噪声的鲁棒性。他们提出的特权信息方法称为 TRAM，在各种实验设置中均取得了提升，涵盖合成噪声和真实世界噪声。然而，他们的结果缺乏对特权信息的不同来源如何影响性能的详细分析。

我们的工作受到深度学习理论的启发。该领域对神经网络的记忆动态进行了丰富研究（Zhang 等人，2017年；Rolnick 等人，2017年；Toneva 等人，2019年；Maennel 等人，2020年；Baldock 等人，2021年）。在无 PI 的设定中，神经网络的动态表现为错误标签往往会在训练后期被记忆，这一现象已被噪声标签研究群体广泛利用，并发展出早停和正则化等技术（Liu 等人，2020年；Bai 等人，2021年）。其他研究则利用干净样本与错误标注样本在学习过程中的内在差异，通过自监督（Veit 等人，2017年；Li 等人，2020年）、协同教学（Han 等人，2018年）或正则化（Cheng 等人，2021年）来检测并纠正分类错误。最后，许多研究尝试通过估计标签转移矩阵（Patrini 等人，2017年），或直接在预测空间中估计噪声信号（Liu 等，2022年），对标签污染过程进行建模。总的来说，我们认为这条关于噪声标签的研究路线（Song 等人，2020年）与特权信息的使用相互独立；实验结果也表明，我们的特权信息方法实际上具有互补性，能够与这些技术自然结合。

这项工作的某些方面让人联想到因果推理。特别是，当某个变量所取的值存在多种解释时，解释消除 是一种众所周知的现象，例如，该值究竟是真实标签被正确标注的结果，还是标注者出错所致（Pearl，2009）。尽管我们在这项工作中没有明确使用因果形式化，但我们在结果中观察到了类似的学习动态。特权信息（通常称为辅助标签）也用于以因果为动机的鲁棒机器学习工作中，不过这类工作通常关注的是处理伪相关这一不同问题，而不是克服标签噪声（Kallus 等人，2018；Veitch 等人，2021；Makar 等人，2022）。在自监督学习中，消除捷径同样是一个受到关注的研究主题（Minderer 等人，2020）。

7. 结论

在这项工作中，我们进行了系统研究，探讨哪些形式的特权信息更能解释消除标签噪声。我们发现，最有帮助的特权信息能够让网络在特征空间中区分正确标注样本与错误标注样本，同时也能提供更容易的学习捷径，以记忆错误标注样本。我们还表明，使用恰当特权信息来解释消除标签噪声的方法，可以与其他最先进的方法相结合，以去除噪声并取得累积增益。进一步探索这一方向是未来研究的一个有前景的途径。我们的见解表明，使用特权信息是对抗标签噪声的一条有前景的研究路径。我们的见解还强调，为数据集收集恰当的特权信息需要仔细考量，才能实现有效捷径的学习。

致谢

感谢 Jannik Kossen 对这项工作提出的宝贵意见。我们也感谢 Josip Djolonga 和 Joan Puigcerver 就基础设施和数据处理相关问题展开的有益讨论。

参考文献

Bai, Y.、Yang, E.、Han, B.、Yang, Y.、Li, J.、Mao, Y.、Niu, G. 和 Liu, T. 理解并改进噪声标签学习中的早停。在 神经信息处理系统进展（*NeurIPS*），2021年。

Baldock, R. J. N.、Maennel, H. 和 Neyshabur, B. 透过样本难度审视深度学习。在神经信息处理系统进展（*NeurIPS*），2021年。

Cheng, H.、Zhu, Z.、Li, X.、Gong, Y.、Sun, X. 和 Liu, Y. 实例依赖的标签噪声学习：一种样本筛选方法。在 国际学习表征会议（*ICLR*），2021年。

Collier, M.、Mustafa, B.、Kokiopoulou, E.、Jenatton, R. 和 Berent, J. 大规模图像分类中的相关的输入依赖型标签噪声。在*IEEE*计算机视觉与模式识别会议（*CVPR*），2021年。

Collier, M.、Jenatton, R.、Kokiopoulou, E. 和 Berent, J. 迁移并边缘化：利用特权信息解释标签噪声。在 国际机器学习会议（*ICML*），2022年。

D'Amour, A.、Heller, K. A.、Moldovan, D.、Adlam, B.、Alipanahi, B.、Beutel, A.、Chen, C.、Deaton, J.、Eisen- stein, J.、Hoffman, M. D.、Hormozdiari, F.、Houlsby, N.、Hou, S.、Jerfel, G.、Karthikesalingam, A.、Lucic, M.、Ma, Y.、McLean, C. Y.、Mincu, D.、Mitani, A.、Monta- nari, A.、Nado, Z.、Natarajan, V.、Nielson, C.、Osborne, T. F.、Raman, R.、Ramasamy, K.、Sayres, R.、Schrouff, J.、Seneviratne, M.、Sequeira, S.、Suresh, H.、Veitch, V.、Vlady myrov, M.、Wang, X.、Webster, K.、Yadlowsky, S.、Yun, T.、Zhai, X. 和 Sculley, D. 欠具体化给现代机器学习的可信度带来了挑战。*CoRR*，abs/2011.03395，2020年。URL https://arxiv.org/abs/2011.03395。

Deng, J., Dong, W., Socher, R., Li, L. J., Kai, L. 和 Li, F. F. ImageNet：大规模分层图像

数据库。在*IEEE*计算机视觉与模式识别会议 *(CVPR)*, 第 248–255 页，2009年。

Geirhos, R., Jacobsen, J., Michaelis, C.,Zemel, R. S.,Brendel, W., Bethge, M., and Wichmann, F. A. 深度神经网络中的捷径学习。自然机器智能, 2 (11):665–673, 2020年。

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. 协同教学：使用极度噪声标签对深度神经网络进行鲁棒训练。发表于神经信息处理系统进展 *(NeurIPS)*, 2018。

Hinton, G. E., Vinyals, O., and Dean, J. 神经网络中的知识蒸馏。*CoRR*,abs/1503.02531, 2015年。

Kallus, N., Puli, A. M., 和 Shalit, U. 通过实验基础移除隐藏混杂。载于神经信息处理系统进展（*NeurIPS*），2018年。Krizhevsky, A. 从微型图像中学习多层特征。技术报告，多伦多大学，2009年。Lambert, J., Sener, O., 和 Savarese, S. 使用异方差丢弃，在特权信息下进行深度学习。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），2018年。Li, J., Socher, R., 和 Hoi, S. C. H. Dividemix：将噪声标签学习视为半监督学习。载于国际学习表征会议（*ICLR*），2020年。Liu, S., Niles-Weed, J., Razavian, N., 和 Fernandez-Granda, C. 早期学习正则化可防止对噪声标签的记忆。载于神经信息处理系统进展（*NeurIPS*），2020年。Liu, S., Zhu, Z., Qu, Q., 和 You, C. 通过过参数化实现标签噪声下的鲁棒训练。载于国际机器学习会议（*ICML*），2022年。Lopez-Paz, D., Bottou, L., Sch¨olkopf, B., 和 Vapnik, V. 统一蒸馏与特权信息。载于国际学习表征会议（*ICLR*），2016年。Maennel, H., Alabdulmohsin, I. M., Tolstikhin, I. O., Baldock, R. J. N., Bousquet, O., Gelly, S., 和 Keysers, D. 使用随机标签训练时，神经网络学到了什么？载于神经信息处理系统进展（*NeurIPS*），2020年。Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., 和 D'Amour, A. 受因果动机驱动的捷径移除

使用辅助标签。载于 人工智能与统计国际会议（*AISTATS*），2022年。

Minderer, M., Bachem, O., Houlsby, N., and Tschannen, M. 用于自监督表征学习的自动捷径消除。发表于国际机器学习会议 *(ICML)*, 2020年。

Nado, Z., Band, N., Collier, M., Djolonga, J., Dusenberry, M. W., Farquhar, S., Feng, Q., Filos, A., Havasi, M., Jenatton, R., 等。《不确定性基线方法：深度学习中不确定性与鲁棒性的基准测试》。*arXiv*预印本 *arXiv:2106.04015*, 2021年。

Patrini, G., Rozza, A., Menon, A. K., Nock, R., 和 Qu, L。《使深度神经网络对标签噪声具有鲁棒性：一种损失校正方法》。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），2017年。Pearl, J.《因果关系：模型、推理与推断》。剑桥大学出版社，2009年。Peterson, J. C., Battleday, R. M., Griffiths, T. L., 和 Russakovsky, O.《人类的不确定性使分类更加鲁棒》。载于*2019*年*IEEE/CVF*国际计算机视觉会议（*ICCV 2019*），韩国首尔，*2019*年*10*月*27*日至*11*月*2*日，2019年。Pezeshki, M., Kaba, S., Bengio, Y., Courville, A. C., Precup, D., 和 Lajoie, G.《梯度饥饿：神经网络的一种学习倾向》。载于神经信息处理系统进展（*NeurIPS*），2021年。Prabhu, V., Yenamandra, S., Singh, A., 和 Hoffman, J.《通过探测注意力条件掩码一致性来适配自监督视觉 Transformer》。载于神经信息处理系统进展（*NeurIPS*），2022年。Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等。《从自然语言监督中学习可迁移视觉模型》。载于国际机器学习会议，第8748–8763页。PMLR，2021年。Rolnick, D., Veit, A., Belongie, S., 和 Shavit, N.《深度学习能够抵御大规模标签噪声》。*arXiv*预印本 *arXiv:1705.10694*，2017年。Sheng, V. S., Provost, F., 和 Ipeirotis, P. G.《再获取一个标签？使用多个带噪声的标注者提升数据质量和数据挖掘》。载于第*14*届*ACM SIGKDD*国际知识发现与数据挖掘会议论文集（*KDD*），2008年。

Snow, R., O'Connor, B., Jurafsky, D. 和 Ng, A. 便宜又快速——但效果好吗？评估自然语言任务中的非专家标注

。见自然语言处理经验方法会议（*EMNLP*），2008年10月。Song, H., Kim, M., Park, D. 和 Lee, J. 使用深度神经网络从噪声标签中学习：综述。*CoRR*, abs/2007.08199，2020年。Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. 和 Wojna, Z. 重新思考计算机视觉的Inception架构。见*IEEE*计算机视觉与模式识别会议（*CVPR*），2016年。Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y. 和 Gordon, G. J. 深度神经网络学习过程中样本遗忘的实证研究。见国际学习表征会议（*ICLR*），2019年。Vapnik, V. 和 Izmailov, R. 利用特权信息进行学习：相似性控制与知识迁移。《机器学习研究杂志》，16:2023–2049，2015年。Vapnik, V. 和 Vashist, A. 一种新的学习范式：利用特权信息进行学习。神经网络，22 (5-6)，2009年。Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A. 和 Belongie, S. J. 以最少监督从大规模噪声数据集中学习。见*IEEE*计算机视觉与模式识别会议（*CVPR*），2017年。Veitch, V., D'Amour, A., Yadlowsky, S. 和 Eisenstein, J. 文本分类中对伪相关的反事实不变性。见神经信息处理系统进展（*NeurIPS*），2021年。Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G. 和 Liu, Y. 重新审视噪声标签学习：一项使用真实世界人工标注的研究。见国际学习表征会议（*ICLR*），2022年。Yang, H., Zhou, J. T., Cai, J. 和 Ong, Y. MIML-FCN+: 利用特权信息的全卷积网络实现多示例多标签学习。见*2017*年 *IEEE*计算机视觉与模式识别会议（*CVPR*），2017年。Yang, S., Sanghavi, S., Rahmanian, H., Bakus, J. 和 Vishwanathan, S. 探索学习排序中特权特征蒸馏的原理。见神经信息处理系统进展（*NeurIPS*），2022年。Zhang, C., Bengio, S., Hardt, M., Recht, B. 和 Vinyals, O. 理解深度学习需要重新思考泛化。见国际学习表征会议（*ICLR*），2017年。

Zhao, P.、Yang, Y. 和 He, Q.-C. 通过隐式正则化进行高维线性回归。*Biometrika*，109（4）：1033–1046，2022年2月。doi：10.1093/biomet/asac010。URL https://doi.org/10.1093%2Fbiomet% 2Fasac010。

附录

本附录的组织如下：附录 A 描述了为生成 ImageNet-PI 而进行的重新标注过程。附录 B 详细介绍了我们的实验细节和超参数调优策略。附录 C 在 PI 数据集的低噪声版本中复现了我们的主要发现。附录 D 讨论并通过消融实验分析了早停在我们实验中的影响。附录 E 提供了我们在其他数据集上进行消融研究的额外结果。附录 F 和附录 G 分别进一步介绍了 TRAM++和AFM++。最后，附录 I 和附录 J 分别详细描述了将 SOP 和 HET 与 TRAM 结合时所采用的实验设置。

A. ImageNet-PI

ImageNet-PI 是标准 ILSVRC2012 ImageNet 数据集的重新标注版本，其标签由 16 个采用不同架构、在标准 ILSVRC2012 上预训练的深度神经网络提供。具体而言，这些预训练模型从 tf.keras.applications[7] 下载，包括：ResNet50V2、ResNet101V2、ResNet152V2、DenseNet121、DenseNet169、DenseNet201、InceptionResNetV2、InceptionV3、MobileNet、MobileNetV2、MobileNetV3Large、MobileNetV3Small、NASNetMobile、VGG16、VGG19 和 Xception。

在重新标注过程中，我们不会直接采用各模型置信度最高的预测，而是针对每个样本，从各模型对该样本的预测分布中随机抽取一个标签。此外，为了控制重新标注数据集时引入的标签噪声量，ImageNet-PI 提供了使用随机温度缩放来增加预测分布熵的选项。此过程的随机性由参数 β 控制，该参数控制伽马分布（形状参数为 $\alpha = 1.0$）的逆尺度；温度值从该分布中采样，代码片段如下：

```
# 获取模型标注者的预测分布。预测_分布  =model.predict(...)

# 对温度进行采样。温度 = tf.random.gamma([



# 计算新的预测分布。log_probs  =tf.math.log(pred_


# 从新的预测分布中采样。class_预测  = tf.random.categorical(tf.math.log(new_pred_dist),
```

直观地说，β 的值越小，温度值就越高，如图 5 所示；这会导致更高程度的标签噪声，因为在高温下，softmax 会趋近于均匀分布。

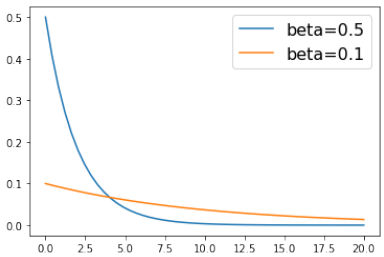

图 5。 参数 β 对温度采样的影响。

[7]https://www.tensorflow.org/api docs/python/tf/keras/applications

这一重新标注过程可能产生任意程度的噪声标签，其分布与对称分布相去甚远，也就是说，并非所有误分类发生的可能性都相同。例如，相似的犬种更有可能彼此混淆，而将“狗”重新标注为“椅子”的可能性则较低。

此数据集中的特权信息来自模型对所采样标签的置信度、模型的参数数量，以及它们在干净测试分布上的测试准确率。这些特权信息特征能够很好地反映各个模型的预期可靠性。在数据集发布中，我们将提供以下文件：

• **labels-train.csv, labels-validation.csv** 这些文件分别包含训练集和验证集的新（带噪声）标签。这些新标签由预训练标注模型提供。每个文件均以 CSV 格式提供标签：

```
<图像 编号>,<标签 1>,<标签 2>,...,<标签 16>
```

• **confidences-train.csv, confidences-validation.csv** 这些文件分别包含各个标注模型对其标注结果的置信度，涵盖训练集和验证集。每个文件均以 CSV 格式提供置信度：

```
<图像 编号>,<置信度 1>,<置信度 2>,...,<置信度 16>
```

• **annotator-features.csv** 此文件以 CSV 格式包含标注者特征（即有关模型标注者本身的元数据）（共 16 行，每位模型标注者一行）：

```
<模型 准确率>, <模型 参数 数量>
```

具体而言，我们将提供通过应用上文讨论的温度抽样过程获得的两组标准化抽样标注：一组对应于 $\beta = 0.1$ 高标签噪声，另一组对应于 $\beta = 0.5$ 低标签噪声。

B. 实验细节

我们基于开源 Uncertainty Baselines 代码库（Nado 等人，2021年）中的实现和超参数开展工作。本文中的所有结果均基于 5 个随机种子报告。

B.1. 特定数据集的训练设置

B.1.1. CIFAR

所有 CIFAR 模型均使用 SGD 优化器进行训练，采用 0.9 的 Nesterov 动量，训练 90 轮，每批次大小为 256。我们对初始学习率 $\{0.01, 0.1\}$ 进行搜索，并在训练 27、54 和 72 轮后将学习率衰减至原来的 0.2 倍。我们对 L2 正则化参数 $\{0.00001, 0.0001, 0.001\}$ 进行搜索。遵循 Nado 等人（2021）的做法，我们采用宽型残差网络架构，模型宽度倍数为 10，模型深度为 28。

除非另有说明，对于 TRAM 和 AFM 模型，我们将 PI 塔宽度设为 1024，因为这是 Collier 等人（2022）采用的默认参数。我们对蒸馏（PI）的教师模型使用相同的架构。这控制了用于整合 PI 的子网络的大小。该子网络的参数化方式为：将预处理后的 PI（随后通过一个 Dense + ReLU 层）与基础宽型残差网络模型对非 PI 输入的表征进行拼接，然后依次通过一个 Dense + ReLU 层、一个残差连接，最后将联合特征空间与非 PI 表征拼接。Dense 层中的单元数由“PI 塔宽度”控制。

对于蒸馏模型，我们在 $[0.5, 10]$的温度区间内进行均匀采样。对于CIFAR-10N和CIFAR-100N，我们将原始训练集划分为训练集和验证集；其中98%的样本用于训练，剩余2%的样本用作验证集。由于CIFAR-10H训练集规模较小（按照Collier 等人（2022）的做法，它实际上是原始CIFAR测试集），因此将原始训练集的96%用作训练集，剩余4%用作验证集。

对于 TRAM++ ，以及在适用时对于 AFM，我们搜索无 PI 损失权重 $\{0.1, 0.5\}$、PI 塔宽度$\{512, 1024, 2048, 4096\}$ 和随机 PI 长度 $\{8, 14, 28\}$。对于异方差CIFAR模型，我们将

异方差协方差矩阵低秩分量的因子数量（Collier 等，2021）设为：CIFAR-10H 和 CIFAR-10N 为 3，CIFAR-100N 为 6；并搜索异方差温度的取值 $\{0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 2.0, 3.0, 5.0\}$ 。

B.1.2. IMAGENET-PI

ImageNet 模型使用 SGD 优化器进行训练，采用 0.9 的 Nesterov 动量，训练 90 轮次，批量大小为 128。初始学习率设为 0.05，并在训练 30、60 和 80 轮次后将学习率衰减为原来的 0.1。我们对 L2 正则化参数 $\{0.00001, 0.0001\}$进行网格搜索。模型采用 ResNet-50 架构。

对于 TRAM 和 AFM 模型，默认将 PI 塔宽度设为 2048，PI 塔采用与 CIFAR 模型相同的参数化方式。对于蒸馏模型，我们将蒸馏温度设为 0.5。我们将原始 ImageNet 训练集的 1% 用作验证集。

对于 TRAM++ ，以及适用时的 AFM，我们将无 PI 损失权重设为 0.5，并将随机 PI 长度设为 30。对于异方差模型，我们将异方差协方差矩阵低秩分量的因子数设为 15，并对异方差温度 $\{0.75, 1.0, 1.5, 2.0, 3.0\}$ 进行搜索。

B.2. 超参数调优策略

除非另有说明，我们报告的是在超参数取最佳验证集准确率的 arg max 所确定的取值时的测试集准确率（其中，轮次数也被视为待最大化集合的一部分）。所用验证集的噪声标签由与训练集相同的过程生成。这种做法实现了带早停的现实且存在噪声的超参数搜索；我们认为，这最能复现真实场景中可能采用的做法，因为在这些场景中，干净验证集可能并不可用。然而，其他论文报告的测试集指标，是通过假设有干净验证集可用和/或不采用早停的超参数搜索得出的，这可能会对所报告的测试集指标产生重大影响（以这种方式计算的结果见附录 D）。

C. 低噪声设置下的结果

在正文中，我们始终　　各数据集高噪声设置下报告的结果。然而，我们现在展示了　　t
表中的所有发现　　1 也适用于低噪声设置。

表 3. 在数据集的低噪声版本上，使用不同特征作为特权信息训练的若干方法的测试准确率 (基线方法以灰色和斜体表示，不使用特权信息)。此处，原始表示数据集的标准特权信息，指示器是一种将干净样本与噪声样本区分开的二元信号，标签是独热编码标签，而近似最优是一种合成特征：对于标注错误的样本，提供标注者标签；对于其他样本，则提供零向量。**粗体数字**表示不同 PI 特征中的显著最大值，其中显著性表示 p值 < 0.05。

		原始	指示器	标签	近似最优
CIFAR10-H（均匀）	*无 PI*	85.8±0.3	85.8±0.3	85.8±0.3	85.8±0.3
	蒸馏（无 PI）	82.7±0.0	82.7±0.0	82.7±0.0	82.7±0.0
	TRAM	85.1±0.2	86.3±0.3	47.3±1.7	**87.6±0.1**
	近似 FM	85.9±0.1	86.8±0.2	38.1±0.5	**87.4±0.1**
	蒸馏（PI）	81.0±0.0	**83.2±0.3**	81.1±0.3	**83.4±0.3**
CIFAR10-N（均匀）	*无 PI*	88.3±0.4	88.3±0.4	88.3±0.4	88.3±0.4
	蒸馏（无 PI）	85.4±0.0	85.4±0.0	85.4±0.0	85.4±0.0
	TRAM	88.7±0.6	92.2±0.3	54.2±0.1	**93.7±0.1**
	近似 FM	88.6±0.6	93.0±0.2	37.4±1.7	**94.0±0.1**
	蒸馏（PI）	88.1±0.0	**88.8±0.6**	**88.8±0.7**	**88.6±0.4**
ImageNet-PI（低噪声）	*无 PI*	68.3±0.3	68.3±0.3	68.3±0.3	68.3±0.3
	蒸馏（无 PI）	67.9±0.3	67.9±0.3	67.9±0.3	67.9±0.3
	TRAM	**69.7±0.3**	**69.8±0.2**	55.9±0.5	66.8±0.1
	近似 FM	**70.6±0.2**	**70.7±0.3**	24.9±13.6	**70.8±0.2**
	蒸馏（PI）	**67.2±0.2**	**67.4±0.2**	**52.5±26.3**	**67.2±0.7**

D. 早停的影响

由于早停是应对标签噪声最有效的基线方法之一，在我们的所有实验中，我们都从噪声训练集中留出一小部分，并报告验证准确率最高的轮次对应的干净测试准确率。不过，为了确保我们的发现不取决于是否使用早停，也不取决于验证集中的标签噪声量，我们现重新呈现表1中的结果：分别采用禁用早停，或使用干净验证集进行早停和超参数调优的设置。

D.1. 不进行早停

表4展示了未使用早停时的基准测试结果。总体而言，我们观察到，在不使用早停的情况下，大多数基线方法的表现明显更差，因为它们对噪声标签的过拟合更加严重。在这方面，由于特权信息（PI）的主要优势之一是能够防止对噪声标签的记忆，我们发现，不使用早停时，特权信息技术相对于其无 PI 基线的相对改进幅度要大得多。

表4. 使用不同特征作为特权信息训练的多种方法的测试准确率（不使用早停）（灰色和斜体表示基线方法不使用特权信息）。此处，原始表示数据集的标准特权信息，指示器表示区分干净样本和噪声样本的二元信号，标签表示独热编码的标签，近似最优表示一种合成特征：对于标注错误的样本，赋予其标注者标签；否则赋予零向量。**粗体数字**表示在各项特权信息特征中显著的最大值，其中显著性指 p值 < 0.05。

		原始	指示器	标签	近似最优
CIFAR-10H（最差）	*无 PI*	42.6±0.3	42.6±0.3	42.6±0.3	42.6±0.3
	蒸馏（无 PI）	45.2±0.1	45.2±0.1	45.2±0.1	45.2±0.1
	TRAM	59.2±0.2	46.5±0.6	39.9±0.3	**77.4±0.1**
	近似 FM	61.2±0.7	39.3±0.6	10.0±0.0	**79.3±0.3**
	蒸馏（PI）	45.2±0.0	**46.3±0.1**	45.7±0.1	**46.4±0.1**
CIFAR-10N（最差）	*无 PI*	67.7±0.6	67.7±0.6	67.7±0.6	67.7±0.6
	蒸馏（无 PI）	71.4±0.2	71.4±0.2	71.4±0.2	71.4±0.2
	TRAM	67.0±0.4	79.2±0.9	51.6±0.2	**91.9±0.2**
	近似 FM	69.8±0.5	89.6±0.1	10.0±0.0	**92.3±0.2**
	蒸馏（PI）	71.1±0.2	**73.1±0.3**	70.9±0.3	70.9±0.2
CIFAR-100N	*无 PI*	55.8±0.2	55.8±0.2	55.8±0.2	55.8±0.2
	蒸馏（无 PI）	58.6±0.1	58.6±0.1	58.6±0.1	58.6±0.1
	TRAM	56.4±0.3	56.0±0.4	34.9±0.2	**67.5±0.3**
	近似 FM	58.9±0.3	65.4±0.3	4.1±0.1	**70.8±0.4**
	蒸馏（PI）	58.9±0.2	58.9±0.3	56.8±0.2	**60.4±0.4**
ImageNet-PI（高噪声）	*无 PI*	47.7±0.8	47.7±0.8	47.7±0.8	47.7±0.8
	蒸馏（无 PI）	50.4±0.8	50.4±0.8	50.4±0.8	50.4±0.8
	TRAM	53.3±0.4	53.5±0.4	41.0±0.8	**56.5±0.3**
	近似 FM	55.0±0.4	55.5±0.3	0.4±0.1	**58.3±0.1**
	蒸馏（PI）	**50.9±0.4**	**50.6±0.2**	39.1±5.1	18.0±24.7

D.2. 干净验证集

我们研究的大多数数据集的训练集中都含有大量标签噪声。在这方面，我们从训练集中留出的少量验证集同样具有很大的噪声，这会影响早停和超参数调优的效果。因此，我们还在表5中提供了相关结果：在进行超参数调优和早停时，我们使用验证集中的干净标签。正如我们所看到的，在这种设置下，大多数方法的表现更好，尽管我们关于特权信息属性如何影响性能的主要发现仍然成立。

表*5*。 使用不同特征训练的多种方法的测试准确率（这些特征作为PI）使用干净验证集选择最佳超参数（基线方法以灰色和斜体表示，不使用PI）。此处，原始表示数据集的标准PI，指示器表示区分干净样本和噪声样本的二元信号，标签表示独热编码标签，而近似最优表示一种合成特征：对于标注错误的样本，赋予其标注者标签；否则赋予零向量。**粗体数字**表示PI特征中的显著最大值，其中显著性表示p值 < 0.05。

		原始	指示器	标签	近似最优
CIFAR-10H（最差）	*无 PI*	53.2±1.0	53.2±1.0	53.2±1.0	53.2±1.0
	蒸馏（无 PI）	53.4±1.0	53.4±1.0	53.4±1.0	53.4±1.0
	TRAM	67.7±0.1	64.9±0.6	39.7±0.3	**77.4±0.1**
	近似 FM	70.6±0.4	66.7±2.1	29.5±0.5	**79.1±0.2**
	蒸馏（PI）	**53.9±0.4**	**53.3±0.6**	**53.4±0.4**	51.6±0.0
CIFAR-10N（最差）	*无 PI*	81.4±0.5	81.4±0.5	81.4±0.5	81.4±0.5
	蒸馏（无 PI）	82.9±0.4	82.9±0.4	82.9±0.4	82.9±0.4
	TRAM	81.9±0.3	89.1±0.3	51.6±0.1	**91.1±0.1**
	近似 FM	82.0±0.3	91.2±0.3	22.6±0.2	**92.3±0.2**
	蒸馏（PI）	**80.8±0.3**	**80.7±0.5**	**81.1±0.4**	**80.8±0.2**
CIFAR-100N	*无 PI*	60.8±0.2	60.8±0.2	60.8±0.2	60.8±0.2
	蒸馏（无 PI）	60.8±0.1	60.8±0.1	60.8±0.1	60.8±0.1
	TRAM	60.6±0.3	63.3±0.2	34.8±0.4	**67.3±0.3**
	近似 FM	60.2±0.1	67.8±0.3	20.1±0.3	**70.9±0.2**
	蒸馏（PI）	61.1±0.2	**61.9±0.2**	60.5±0.2	61.5±0.3
ImageNet-PI（高噪声）	*无 PI*	48.3±0.1	48.3±0.1	48.3±0.1	48.3±0.1
	蒸馏（无 PI）	50.5±0.7	50.5±0.7	50.5±0.7	50.5±0.7
	TRAM	53.3±0.3	53.8±0.7	40.7±0.8	**56.5±0.2**
	近似 FM	55.6±0.3	55.5±0.4	0.8±0.2	**58.2±0.1**
	蒸馏（PI）	**51.0±0.4**	**50.7±0.3**	39.1±4.4	**27.6±22.7**

E. 更多结果

本节将展示使用其他数据集和算法进行实验的完整结果，以及论文中的主要发现。

E.1. 训练动态

在图2中，我们详细分析了采用不同 PI 特征时 TRAM 在 CIFAR-10H 上的动态变化。现在我们展示 TRAM 在 CIFAR-10N 和 CIFAR-100N 上的结果（分别见图 6 和图 7）。我们还展示 AFM 在 CIFAR-10H、CIFAR-10N 和 CIFAR-100N 上的结果（分别见图 8、图 9 和图 10）。

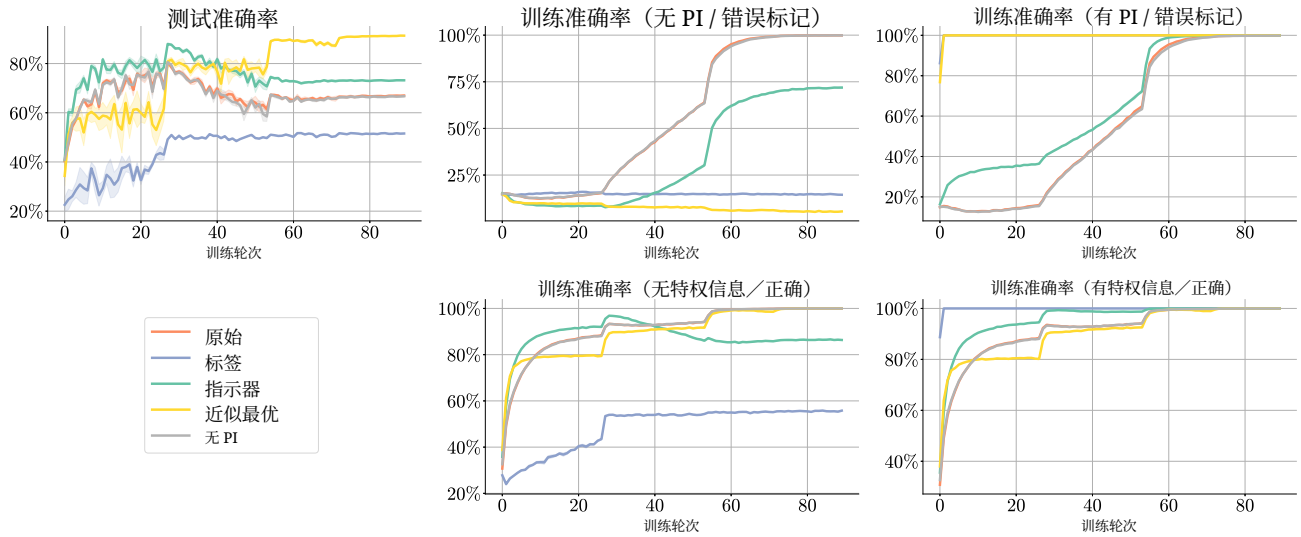

图 6. 不同 PI 特征下，TRAM 在 CIFAR-10N 上的动态。**左上**：测试准确率。**上中**：在无 PI 头上评估的噪声样本训练准确率，**右上**：在 PI 头上评估的噪声样本训练准确率。**下中**：在无 PI 头上评估的干净样本训练准确率。**右下**：在 PI 头上评估的干净样本训练准确率。

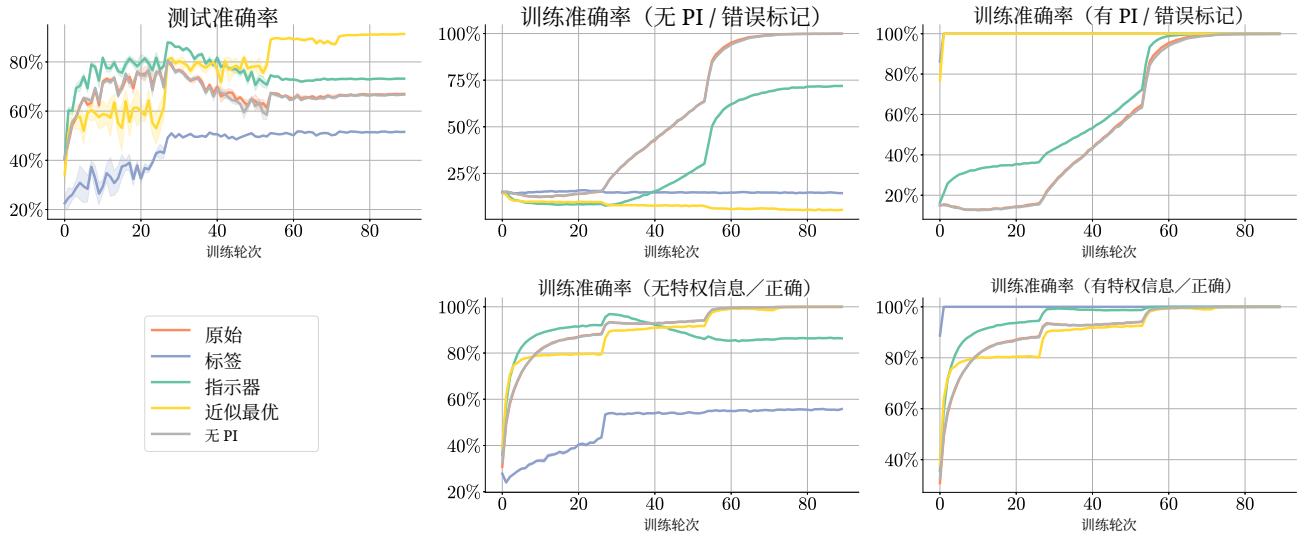

图 7. 不同 PI 特征下，TRAM 在 CIFAR-100N 上的动态。**左上**：测试准确率。**上中**：在无 PI 头上评估的噪声样本训练准确率，**右上**：在 PI 头上评估的噪声样本训练准确率。**下中**：在无 PI 头上评估的干净样本训练准确率。**右下**：在 PI 头上评估的干净样本训练准确率。

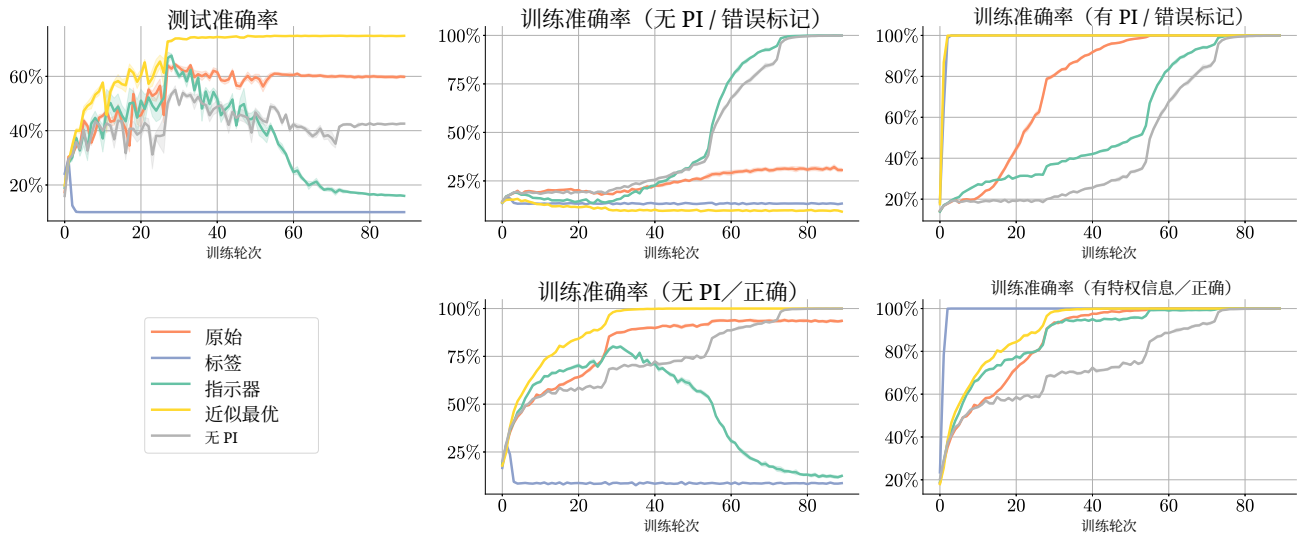

图 *8.* 使用不同 PI 特征时，AFM 在 CIFAR-10H 上的动态变化。**左上**：测试准确率。**上中**：通过边缘化评估的噪声样本训练准确率，**右上**：在特权信息头处评估的噪声样本训练准确率。**下中**：通过边缘化评估的干净样本训练准确率。**右下**：在 PI 头处评估的干净样本训练准确率。

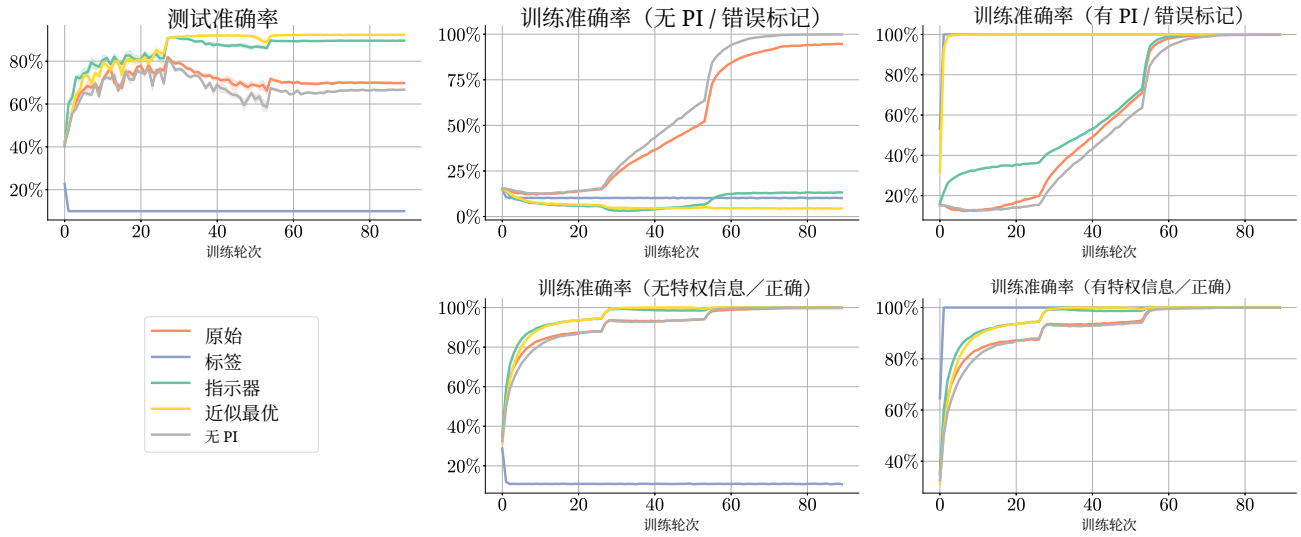

图 *9.* 使用不同 PI 特征时，AFM 在 CIFAR-10N 上的动态变化。**左上**：测试准确率。**上中**：通过边缘化评估的噪声样本训练准确率，**右上**：在特权信息头处评估的噪声样本训练准确率。**下中**：通过边缘化评估的干净样本训练准确率。**右下**：在 PI 头处评估的干净样本训练准确率。

E.2. 特征提取器大小

我们在其他设置下复现了图 3 中的结果，并得到了相同的发现。具体而言，我们展示了 CIFAR-10N 和 CIFAR-100N 上的结果（分别见图 11 和图 12）。

E.3. PI 头部大小

我们在 CIFAR-10N 和 CIFAR-100N 上复现了图 4 中的结果（分别见图 13 和图 14）。然而，在这种情况下，我们没有观察到明确的趋势，这可能是因为这些数据集上的原始PI 不够好，无法让 TRAM 和 AFM 发挥优势（参见表1）。在这方面，增大 PI 头部大小并不会带来更好的

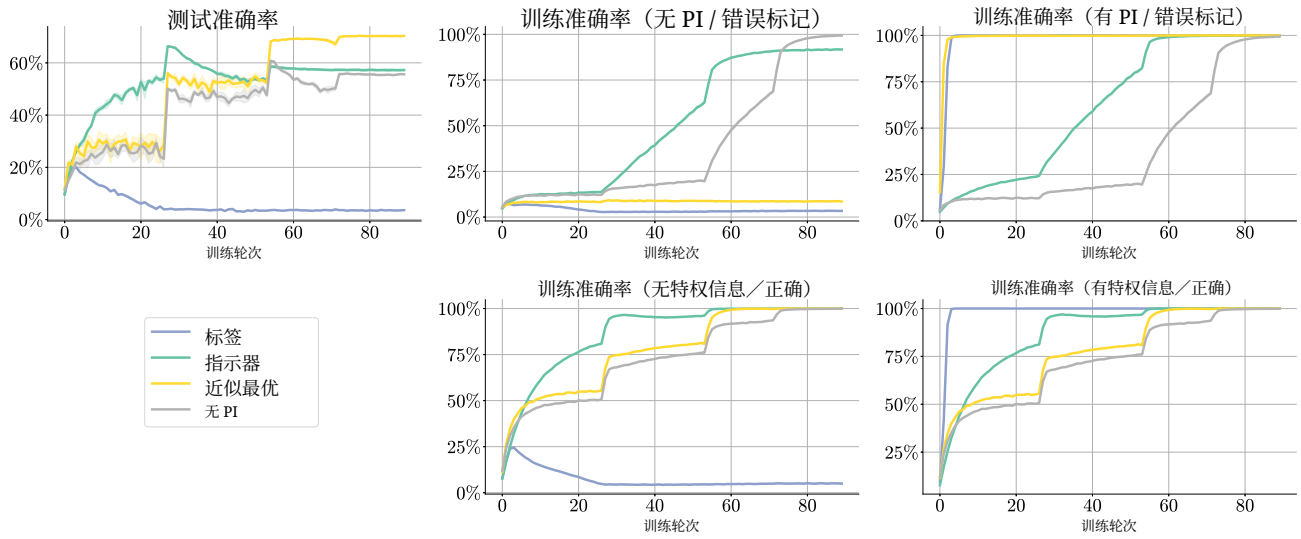

图 *10.* 不同 PI 特征下 AFM 在 CIFAR-100N 上的动态变化。**左上**：测试准确率。**中上**：通过边缘化评估的噪声样本训练准确率，**右上**：在 PI 头处评估的噪声样本训练准确率。**中下**：通过边缘化评估的干净样本训练准确率。**右下**：在 PI 头处评估的干净样本训练准确率。

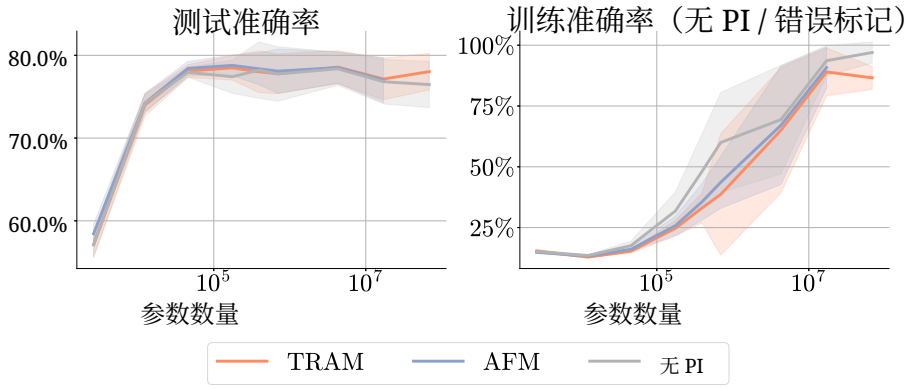

图 *11.* PI 塔保持不变、增加特征提取器参数数量时，不同 PI 基线方法在 CIFAR-10N 上的性能。较大的模型容易过拟合，因为它们倾向于利用更大的容量对噪声样本过拟合，从而阻碍模型利用 PI。

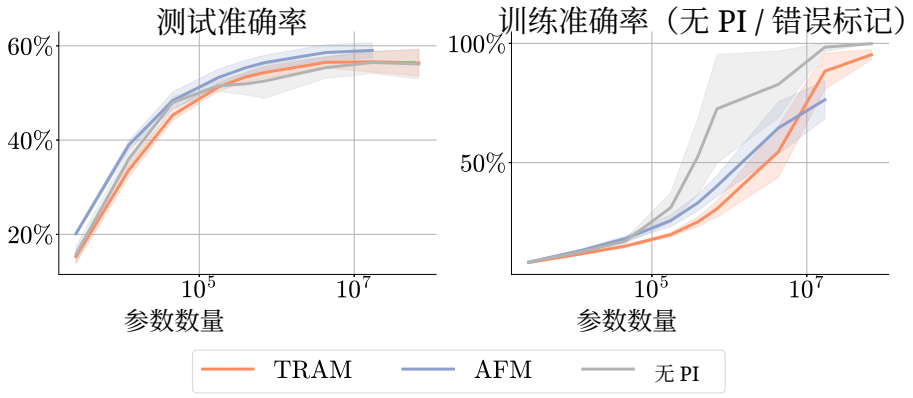

图 *12.* 在保持 PI 塔不变的情况下，增加特征提取器的参数数量时，不同 PI基线在 CIFAR-100N 上的性能。较大的模型更容易过拟合，因为它们倾向于利用更大的容量来拟合噪声样本，从而阻碍模型利用特权信息。

因为从特权信息中没有任何内容可供提取，性能也随之如此。

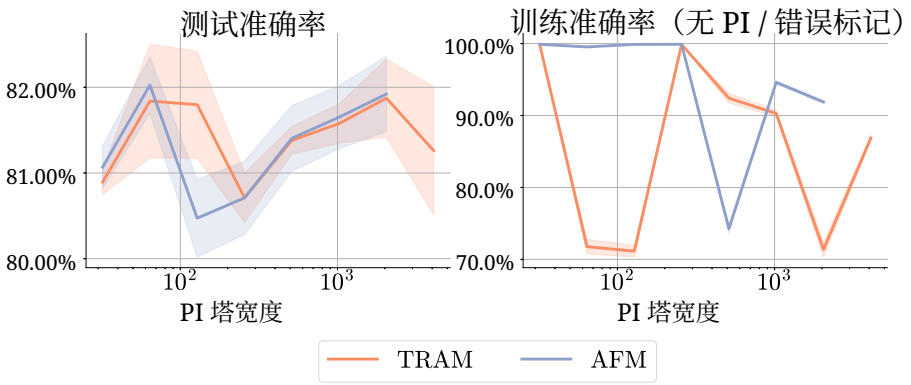

图 *13*。在增大 PI 头部大小时，CIFAR-10N 上不同 PI基线的性能。更大的 PI 头部大小会促使模型利用 PI 记忆噪声样本，从而更多地将 PI 用作捷径。

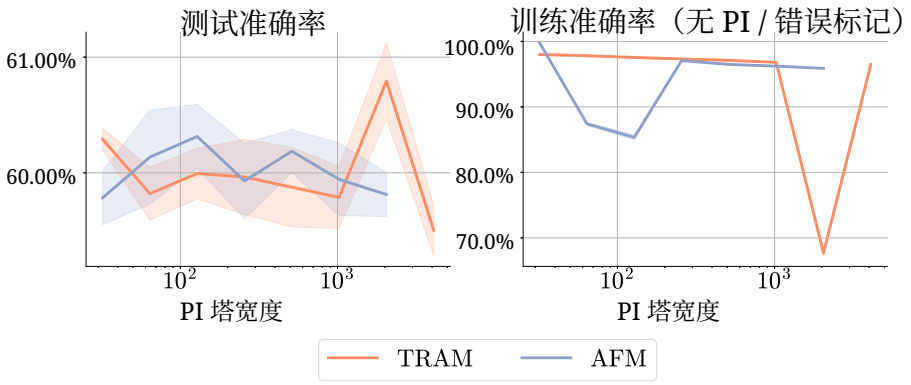

图 *14*。增大 PI 头部大小时，不同 PI 基线在 CIFAR-100N 上的表现。较大的 PI 头部大小会促使模型利用 PI 记住噪声样本，从而更多地将 PI 用作捷径。

F. TRAM++ 的设计细节

现在我们给出 TRAM 的设计细节++，这是 TRAM 的改进版本，它会向 原始 PI 追加一个独特的随机 PI 向量。具体而言，我们遵循了论文中其余 TRAM 实验采用的相同调参策略，并且还调整了 λ 这个用于对两个输出头的损失进行加权的参数，即

$$\min_{\phi,\pi,\psi} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{a},\tilde{y})} \left[\ell(\pi(\phi(\boldsymbol{x}),\boldsymbol{a}),\tilde{y}) + \lambda\,\ell(\psi(\phi(\boldsymbol{x})),\tilde{y})\right]. \tag{2}$$

Collier 等人（2022）提出，无 PI 头的梯度不会影响特征提取的更新，因此， λ 可以直接纳入 TRAM 全局学习率的调节中。然而，在我们的实验中，我们发现，在固定训练轮次的情况下调节 λ 可以显著提升性能，因为这能够减缓无 PI 头的训练速度。如图 15 所示，增大 λ 的效果与增大无 PI 头的学习率相同； $\lambda < 1$ 存在一个最佳取值范围，在该范围内，无 PI 头的训练速度足够快，能够拟合干净样本，同时又能避免学习所有噪声样本。

总体而言，为了尽可能贴近原始 TRAM 实现，在其他实验中均未调节 λ。然而，对于 TRAM++ 实验，为了使 TRAM 达到尽可能高的性能，我们调节了 λ。

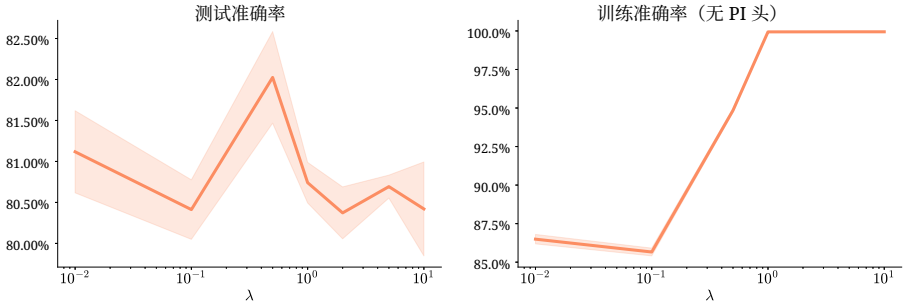

图 *15.* CIFAR-10N 中 TRAM 在不同损失权重 λ 下的性能。最优的 λ 应能在训练干净样本的同时，显著减缓对噪声样本的过拟合，从而取得良好平衡。

G. AFM 的设计++

在第 4.2 节中，我们已经看到，将能唯一标识每个样本的随机 PI 附加到 原始PI 上，有时可以在 TRAM 中诱导出有益的捷径++。现在，我们测试将同一策略应用于 AFM，并设计了 AFM++，这是 AFM 的增强版本，加入了额外的随机 PI。表 6 展示了实验结果，我们可以看到 AFM++ 也明显优于“原版”AFM。同样，对于过拟合问题本就更严重的数据集，性能提升也更大。

表 *6.* 无 PI、AFM 和 AFM++在不同 PI 数据集上的性能比较。

	无 PI	AFM	AFM++
CIFAR-10H（最差）	55.0±1.5	64.0±0.6	**68.2±0.6**
CIFAR-10N（最差）	80.6±0.2	82.0±0.3	**84.6±0.2**
CIFAR-100N	60.4±0.5	60.0±0.2	**61.9±0.2**
ImageNet-PI（高噪声）	47.7±0.8	**55.6±0.3**	**55.0±0.6**

H. TRAM 与标签平滑的结合

我们还评估了 TRAM 与标签平滑（LS）的组合。具体而言，我们遵循标准的标签平滑流程，并将标签平滑超参数加入表1中进行搜索的超参数范围。更具体地说，我们搜索标签平滑系数 0.2、0.4、0.6 和 0.8，并按照论文中所有实验采用的相同流程选择最优超参数设置。结果见表7。

我们观察到，在所有数据集上，将标签平滑添加到 TRAM 方法中都能提升性能，这表明 TRAM 可以与标签平滑成功结合。更普遍地说，这一观察进一步表明，TRAM 与 TRAM++ 彼此兼容；将它们与针对噪声标签开发的常用方法结合时，性能提升可以叠加。

表7。 不同 PI 数据集上无 PI、标签平滑（LS）、TRAM 和 TRAM + LS 的性能比较。

	无 PI	TRAM	LS	TRAM+LS
CIFAR-10H（最差）	55.0±1.5	64.9±0.8	59.9±1.5	**65.4±0.9**
CIFAR-10N（最差）	80.6±0.2	80.5±0.5	80.5±0.4	**82.4±0.2**
CIFAR-100N	60.4±0.5	59.7±0.3	60.0±0.46	**61.9±0.3**

I. SOP 和 TRAM+SOP 的实验细节

正如我们在第5节中所确立的，TRAM 和 SOP 的结合有望在对标签噪声的鲁棒性方面实现累积增益。研究表明，与不使用 PI 的模型相比，TRAM 结合其原始PI能够提升在具有密集噪声的数据集（如 CIFAR-10H（最差））上的性能。然而，PI 并不总能解释噪声；即使能够解释，也未必能完全解释消除所有噪声。此外，即使 PI 能够解释消除噪声，模型的特征提取器及其后续层仍可能容易受到噪声的影响。

另一方面，研究表明，SOP 对稀疏分布的噪声效果良好，其运作原理是对噪声进行建模消除，这与 TRAM 所采用的方法不同。由于这两种原理彼此互补，我们提出结合两种方法的优势，以实现累积增益。

如第5节所述，TRAM+SOP 结合了两个主要步骤：使用 TRAM 进行预训练，以及使用 SOP 进行微调。我们实现 TRAM 时采用了常规 TRAM，并加入了 TRAM++ 的一些改进，例如随机特权信息和更大的 PI 头部大小。需要注意的是，我们的实验使用了自行实现的 SOP；尽管该实现采用了 SOP 方法，并经过论文原作者的合理性检查，但我们的实验基线环境和搜索空间与他们的不同。因此，在 CIFAR-N 数据集上的测试准确率可能低于原始 SOP 论文报告的结果。不过，这些实验的主要目标是探究 TRAM + SOP 能否在 TRAM 和 SOP 各自的独立实现之上实现累积增益，而我们的结果支持这一假设。

在我们的实验中，SOP 和 TRAM+SOP 模型均训练了 120 轮次，学习率分别在第 40、80 和 110 轮次衰减。按照 Liu 等人（2022）的做法，我们在 TRAM 中采用带 Nesterov 动量的 SGD，在 SOP 中则采用常规动量。关于 SOP 参数的详细说明，请参阅原始 SOP 论文。需要注意的是，此处展示的 TRAM+SOP 方法结果并未包含 Liu 等人（2022）提出的所有改进。通过纳入这些改进，并联合优化 TRAM 和 SOP 预训练及微调阶段的超参数空间，性能或许还能进一步提升。

J. TRAM+HET 的实验细节

TRAM+HET 由一个简单的双输出头 TRAM 模型构成，其中两个输出头各自的最后一个线性层都替换为异方差线性层（Collier 等，2021）。因此，在这些实验中，我们也对异方差层的温度进行扫描。Collier 等人（2022）已经提出了类似的方法，称为 Het-TRAM；不过，我们在这里也运用自己的见解，允许模型在原始PI 特征之上使用随机 PI。有趣的是，与 TRAM+SOP 中的情况相反，添加随机 PI（即 TRAM++）并不总能提升 TRAM+HET 的性能。相反，视数据集而定（见表8），我们发现使用随机 PI 有时会损害模型的最终性能（例如 CIFAR-10H）。我们推测

这可能是因为 TRAM+HET 模型也利用随机 PI 记忆干净标签。理解为什么这种情况只会在使用异方差层时发生，是一个值得未来研究探索的方向。

表*8*。TRAM、TRAM++、HET、TRAM+HET（不含额外随机 PI）和 TRAM+HET（含额外随机 PI）在不同 PI 数据集上的性能比较。

	TRAM	TRAM++	HET	TRAM+HET（不含随机）	TRAM+HET（+随机）
CIFAR-10H（最差）	64.9±0.8	66.8±0.3	50.8±1.4	**67.7±0.7**	56.5±0.7
CIFAR-10N（最差）	80.5±0.5	83.9±0.2	81.9±0.4	82.0±0.3	**83.5±0.1**
CIFAR-100N	59.7±0.3	61.1±0.2	60.8±0.4	**62.1±0.1**	61.2±0.3
ImageNet-PI（高噪声）	53.3±0.5	53.9±0.4	51.5±0.6	**55.8±0.3**	**55.4±0.4**