

基于源行为偏离度的降水超分模型黑盒指纹验证方法

袁程胜^{1,2}, 孔令志¹, 李欣亭³, 宋甫元^{1,2}, 刘庆程¹, 付章杰^{1,2}

(1. 南京信息工程大学 网络空间安全学院, 江苏 南京 210044; 2. 南京信息工程大学 数字取证教育部工程研究中心, 江苏 南京 210044;
3. 国防科技大学 外国语学院, 江苏 南京 210039)

摘要: 深度学习驱动的降水数据超分辨率重建已取得显著进展, 但随着派生模型的大量出现, 降水超分模型的版权保护与来源验证面临新的挑战。准确判定候选模型与受保护源模型之间的派生关系, 对保护降水超分模型知识产权至关重要。然而, 超分模型输出为连续空间场, 缺乏分类任务中离散标签所提供的明确决策边界; 同时, 针对相同任务独立训练的不同模型, 受共同任务规律与相近数据分布影响, 也可能出现行为相似性。针对上述问题, 本文提出一种基于源行为偏离度 (source-behavior deviation, SBD) 的降水超分模型黑盒指纹验证方法。首先, 针对连续场输出缺乏分类决策边界的问题, 提出局部守恒反事实成对查询方法, 通过对局部降水量进行双向守恒重分配以激发模型响应, 并仅利用模型最终预测结果构建其局部行为表示。其次, 针对独立训练模型行为相似干扰派生判定的问题, 构建基于源行为参考子空间的 SBD 方法, 以源模型行为响应所定义的变化空间作为统一参照, 将候选模型相对于源模型的总体行为形变投影至源行为参考子空间的正交补, 利用归一化 Frobenius 范数量化无法由源模型行为结构解释的正交偏离程度, 在严格黑盒条件下实现派生关系的有效验证。在 28 个候选模型上的实验结果表明, 当查询预算 $K = 800$ 时, SBD 正确识别 15 个源派生模型并拒绝全部 10 个独立训练模型, 准确率、平衡准确率分别为 89.29%、91.7%。SBD 仅依赖模型最终输出, 能够在严格黑盒条件下有效区分真实派生关系与共同任务导致的行为相似性。

关键词: 黑盒指纹验证; 模型指纹; 降水超分辨率; 源行为偏离度; 反事实扰动

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.1000

Black-Box Fingerprint Verification of Precipitation Super-Resolution Models via Source-Behavior Deviation

Yuan Chengsheng^{1,2}, Kong Lingzhi¹, Li Xinting³, Song Fuyuan^{1,2}, Liu Qingcheng¹, Fu Zhangjie^{1,2}

1. School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
2. Ministry of Education Engineering Research Center for Digital Forensics, Nanjing University of Information Science and Technology, Nanjing 210044, China
3. School of Foreign Languages, National University of Defense Technology, Nanjing 210039, China

Abstract: Deep learning-driven super-resolution reconstruction of precipitation data has achieved significant progress. However, with the rapid proliferation of derivative models, copyright protection and provenance verification of precipitation super-resolution models have become increasingly challenging. Accurately determining the derivation relationship between a candidate model and a protected source model is therefore crucial for safeguarding the intellectual property of precipitation super-resolution models. Nevertheless, super-resolution models produce continuous spatial fields and thus lack the explicit decision boundaries provided by discrete labels in classification tasks. Meanwhile, independently trained

收稿日期: 2026-08-28; 修回日期: 2026-09-22

通信作者: 李欣亭, 邮箱: lixt@tju.edu.cn

基金项目: 国家自然科学基金资助项目 (No.U22B2062, No.U23B2023)

Foundation Items: The National Natural Science Foundation of China (No.U22B2062, No.U23B2023)

models developed for the same task may also exhibit similar behaviors due to shared task-specific regularities and similar data distributions. To address these challenges, this paper proposes a black-box fingerprint verification method for precipitation super-resolution models based on source-behavior deviation (SBD). First, to address the absence of explicit decision boundaries in continuous-field outputs, a locally conservative counterfactual paired-query method is proposed. By bidirectionally redistributing local precipitation amounts under a conservation constraint, the method elicits model responses and constructs local behavioral representations using only the final predictions of the model. Second, to mitigate the interference of behavioral similarity among independently trained models in derivation identification, an SBD method based on a source-behavior reference subspace is developed. Specifically, the variation space defined by the behavioral responses of the source model is used as a unified reference, and the behavioral deformation of a candidate model relative to the source model is projected onto the orthogonal complement of the source-behavior reference subspace. The normalized Frobenius norm is then employed to quantify the degree of orthogonal deviation that cannot be explained by the behavioral structure of the source model, thereby enabling effective verification of derivation relationships under a strict black-box setting. Experimental results on 28 candidate models show that, with a query budget of $K = 800$, SBD correctly identifies 15 source-derived models and rejects all 10 independently trained models, achieving an accuracy of 89.29%, a balanced accuracy of 91.7%. Relying solely on the final model outputs, SBD can effectively distinguish genuine derivation relationships from behavioral similarities induced by shared task characteristics under strict black-box conditions.

Key words: black-box fingerprint verification, model fingerprint, precipitation super-resolution, source-behavior deviation, counterfactual perturbation

0 引言

近年来,深度学习技术已广泛应用于降水降尺度与高分辨率场重建任务^[1-5]。降水降尺度旨在由粗空间分辨率降水信息估计细尺度降水分布,而高分辨率场重建则结合相关气象变量,还原降水场的局部空间结构信息,从而为精细化降水分析与极端降水过程刻画提供更细致的数据支撑。随着相关模型从研究走向业务部署,经由微调、剪枝、量化、知识蒸馏等技术手段可衍生出大量派生版本,源模型的版权保护因此面临现实挑战:在仅能访问模型输入与输出的黑盒条件下,如何准确判定候选模型是否源于给定源模型,并有效区分真实的派生关系与独立训练模型因共享任务及数据分布而产生的输出相似性。尤其对于降水超分模型,其输出为连续空间场,缺乏分类任务中离散标签所提供的明确决策边界,增加了上述判定问题的难度。

模型指纹技术为黑盒条件下的模型来源验证提供了一种非侵入式手段。现有方法主要利用敏感查询样本、模型输出特征、分类决策边界、或对抗扰动行为关系构建模型指纹^[6-9]。Quan等^[10]将该方法扩展至图像恢复任务,但其所依赖的自然图像统计特征与降水场的数值分布特性及物理空间结构存在差异,限制了其向降水超分辨率场景的直接迁移。具体而言,降水超分辨率模型具有两方面特点。其

一,模型输出为具有明确物理意义的连续空间场,而分类任务中依赖离散标签决策边界的指纹方法在此条件下并不适用。其二,对于面向相同任务且在相近数据分布上独立训练的不同模型,受共享降水映射规律的影响,其输出行为可能呈现相似性;而派生模型则因参数或结构调整,其输出行为会与源模型产生偏移。

针对上述问题,本文提出一种基于源行为偏离度的降水超分辨率模型黑盒指纹验证方法。该方法以固定降水输入为对象,构造局部守恒的反事实成对扰动,以模型对正、负扰动的响应差异刻画其局部行为特性;进而利用源模型响应构建源行为参考子空间,通过量化候选模型行为变化中无法由源结构解释的正交偏离,判断其来源关系。整个验证过程仅依赖模型最终预测输出,无需访问模型参数、梯度和中间层特征。本文主要贡献如下:

(1) 针对降水超分辨率模型连续场输出缺乏天然决策边界、致使传统分类模型指纹难以直接适用的问题,本文提出一种局部守恒反事实成对查询方法。该方法通过对局部降水量进行双向守恒重分配以激发模型响应,并仅利用模型最终预测结果构建其局部行为表示。

(2) 针对同任务独立训练模型行为相似、容易干扰模型派生关系判定的问题,提出一种基于源行

为参考子空间的源行为偏离度方法,通过量化候选模型行为变化中无法由源模型行为结构解释的正交偏离,在黑盒条件下有效验证模型来源关系,提升固定判别条件下源派生模型识别的稳定性。

(3) 为验证所提方法的性能,本文在包含微调、剪枝、量化、知识蒸馏及独立训练模型在内的 28 个候选模型上进行了测试。在给定查询预算下, SBD 方法的准确率、平衡准确率和受试者工作特征曲线下面积 (area under the receiver operating characteristic curve, ROC-AUC) 分别达到 89.29%、91.7% 和 0.978, 并成功拒识全部 10 个独立训练模型。

1 相关工作

1.1 基于水印的方法

模型水印技术通过将所有权信息嵌入模型参数、网络结构或特定输入-输出映射中,在发生版权争议时,可通过提取水印或验证触发响应实现归属判定。袁等^[11]针对伪造指纹检测任务,提出了一种融合主动保护与被动验证的版权保护方法,利用对抗样本和后门映射实现模型水印验证。Kim 等^[12]提出基于边界的神经网络水印方法,通过增大水印样本与决策边界之间的间隔,提升水印在面临模型提取和蒸馏攻击时的保持能力。Lv 等^[13]提出 MEA-Defender, 将水印样本与主任务的输入、输出空间相结合,使水印行为能够随模型功能一同迁移至提取模型中。Zhou 等^[14]从频域构造水印样本,以增强水印在模型提取、剪枝及压缩等操作下的鲁棒性。随着生成模型的发展,水印技术已扩展

至扩散模型, Yuan 等^[15-16]面向 Stable Diffusion 模型设计了相应的模型水印方案。

1.2 基于模型指纹的方法

与主动嵌入所有权信息的水印方法不同,模型指纹通过提取模型训练后自然形成的参数特征或输入-输出行为特征来建立模型身份表示,无需对原模型施加额外修改。早期研究主要面向图像分类模型, He 等^[3]通过构造对模型参数变化敏感的查询样本检测模型修改情况; ModelDiff^[4]利用不同模型在相同测试输入上的输出差异构建行为距离; Peng 等^[5]借助针对源模型生成的通用对抗扰动提取模型全局行为特征; DeepJudge^[6]从模型测试行为角度进行版权验证; SAC^[7]进一步利用多个查询样本输出之间的相关关系建立模型指纹。

近年来,模型指纹研究逐步扩展至严格黑盒、连续输出以及生成模型等复杂场景。Quan 等^[10]利用关键图像的颜色统计及局部空间响应识别图像去噪和超分辨率模型; Bai 等^[17]和 He 等^[18]分别从边界敏感样本和严格黑盒完整性验证角度提高模型修改检测能力, Bai 等^[9]进一步提出 SDBF, 通过陡峭决策边界构造高敏感指纹样本。针对不同类型模型, Sinitsa 和 Fried^[19]、ManiFPT^[20] 以及 Teng 等^[21]分别从生成图像固有特征、生成模型响应以及扩散模型去噪行为中提取模型指纹, Merge-Print^[22]则进一步研究模型合并条件下大语言模型指纹的保持能力^[23]。上述研究说明,现有方法仍主要围绕类别标签、分类概率、决策边界或自然图像统计特征设计,对于以连续物理空间场为输出的降水超分辨率模型,其行为表示与来源判定依据仍

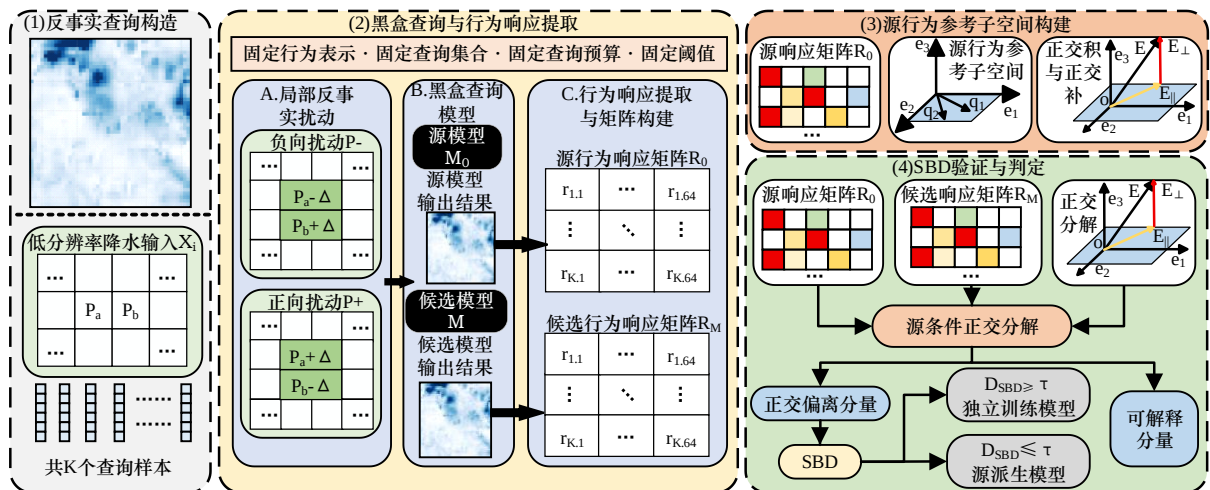


图 1 源行为偏离度方案(SBD)黑盒指纹验证总体架构图

需探索。

1.3 降水数据超分模型的研究

近期,相关视觉超分工作^[24-25]表明,多域特征融合能够有效改善细节恢复与重建质量。在降水数据处理领域, PreciDBPN^[26]将降水分布特点及物理因素融入深度反投影网络,实现小时尺度降水降尺度和偏差订正; Srivastava 等^[27]和 Glawion 等^[28]分别利用视频扩散和生成模型恢复高频降水时空结构。Yi 等^[29]于 2026 年提出条件小波扩散模型,在小波域学习多尺度降水结构,在提高公里级降尺度的同时改善生成效率。

2 基于源行为偏离度的黑盒指纹验证方法

针对同任务模型输出相似性对派生关系判定造成的干扰,本文提出一种 SBD 黑盒指纹验证方法,其整体框架如图 1 所示,主要包括反事实查询构造、黑盒查询与行为响应提取、源行为参考子空间构建和 SBD 验证与判定四个阶段。

2.1 问题定义

给定源模型为 M_0 ,待验证的候选模型为 M 。对于输入样本 x ,记模型的输出为 $M(x)$ 。本文目标是在有限黑盒查询条件下判断 M 与 M_0 是否存在源派生关系。本文将直接继承源模型参数并通过微调、剪枝或量化得到的模型,以及在训练过程中利用源模型输出进行知识传递的蒸馏模型,统一定义为源派生模型。已有研究表明,知识蒸馏能够在异构模型之间实现有效的知识传递^[30],因此即使学生模型与源模型具有不同网络结构,其训练过程仍直接继承了源模型所蕴含的知识。基于此,本文将此类蒸馏模型纳入源派生模型范畴;未继承源模型参数且未利用其输出进行知识传递的模型,则定义为独立训练模型。

根据上述定义,将候选模型 M 相对于给定源模型 M_0 的派生关系标签表示为:

$$y(M, M_0) = \begin{cases} 1, M \text{ 为源派生模型} \\ 0, M \text{ 为独立训练模型} \end{cases} \quad (1)$$

在此基础上,进一步构造验证函数,在有限的查询次数内对上述派生关系标签进行估计。验证阶段不得访问模型权重、梯度、中间层特征及训练过程信息。所有候选模型均使用相同查询集合、行为表示和判别规则。针对黑盒条件下的非自适应模型来源验证场景,验证者拥有待保护源模型 M_0 ,仅

可向源模型与候选模型 M 提交相同查询输入并获取模型返回的最终高分辨率降水预测结果,无法获取候选模型的权重、梯度、中间层特征、网络结构标识等信息。

2.2 局部守恒反事实查询与行为响应构造

为降低正常输入中共同任务规律的影响,本节通过局部降水量进行重新分配,构造方向相反且总量守恒的反事实输入,并利用模型对两端点的输出差异刻画其局部行为响应。具体地,设低分辨率降水场为 P ,枚举所有不跨越边界的水平相邻格点对,以两格点降水量中的较小值作为共享可转移容量 c ,并选择 c 最大格点对作为扰动施加位置,

$$c = \min(p_l, p_r) \quad (2)$$

若存在并列,则按固定行优先顺序选择。令扰动系数 $\varepsilon = 0.2$,实际转移量 Δ 定义为:

$$\Delta = \varepsilon c \quad (3)$$

为明确局部守恒约束,设第 i 个查询中被选中的两个格点记为 a 和 b , e_a 、 e_b 为对应单位基向量,则反事实端点可等价写为:

$$P^+ = P + \Delta(e_a - e_b) \quad (4)$$

$$P^- = P + \Delta(e_b - e_a) \quad (5)$$

由于 $\Delta = \varepsilon c \leq c = \min(p_a, p_b)$,两个端点的格点降水量均保持非负;同时 $\mathbf{1}^T(P^+ - P) = \mathbf{1}^T(P^- - P) = 0$,因此所选局部区域以及全场总降水量严格守恒。由此, P^+ 与 P^- 构成关于原输入 P 的对称、局部守恒反事实扰动对。 P^+ 与 P^- 之间的差异仅源于该格点对上的反向降水重分配,确保不同模型面对同一局部物理变化时具备一致的可比输入条件。对于模型 M ,两个反事实输入得到的最终预测结果分别记为 Y_M^+ 和 Y_M^- 。考虑到降水量具有非负性,且其数值分布呈现显著的长尾特征,本文首先对预测结果施加统一的非线性变换,以压缩量级差异并抑制极端值对行为表征的主导影响:

$$g(Y) = \log[1 + \max(Y, 0)] \quad (6)$$

其中,非负处理保证输出降水量为正, $\log$ 变换用于压缩长尾分布造成的数值尺度差异。上述变换仅作用于模型已返回的最终预测结果,不涉及模型内部参数或计算过程。对于第 i 个查询样本,定义模型 M 的有符号对称行为响应场为:

$$S_i(M) = \frac{1}{2\varepsilon} [g(Y_M^+) - g(Y_M^-)] \quad (7)$$

公式(7)可视为模型输入沿样本特定局部守恒方向 $c(\mathbf{e}_a - \mathbf{e}_b)$ 关于扰动系数 ε 的中心有限差分。该响应描述了在同一局部降水重新分配情形下,模型沿两个相反扰动方向所产生的有限差分变化。与直接使用单一预测场相比,该响应更侧重于模型对局部扰动方向及空间重分配方式的敏感性,而非原始预测结果中绝对数值的影响。

为保留降尺度场中粗粒度空间结构信息,并统一不同模型的响应维度,将每个行为响应场通过 8×8 自适应平均池化压缩,再按固定顺序展开为 64 维行为响应向量,记为 $\mathbf{r}_i(\mathbf{M})$ 。设固定查询集合包含 K 个样本,将所有响应向量按查询顺序堆叠,构成模型 M 的行为响应矩阵:

$$\mathbf{R}_M = [\mathbf{r}_1(M), \mathbf{r}_2(M), \mathbf{r}_3(M), \dots, \mathbf{r}_K(M)]^T \quad (8)$$

其中,上式 $\mathbf{R}_M$ 的尺寸为 $K \times 64$ 。源模型与所有候选模型均采用完全相同的查询样本、扰动位置选择规则、输出变换和空间池化方式,使得不同模型的行为响应矩阵在查询样本和响应维度上均可实现一一对应比较。

2.3 源条件下的行为形变建模

在获得对齐的源模型与候选模型行为响应矩阵后,进入源行为参考子空间构建阶段。由于各模型面向相同的降水超分辨率任务,其响应中既包含共同任务规律所形成的行为共性,也包含由模型训练及派生过程所致的个体差异。因此,来源验证问题可进一步表述为:在候选模型相对于源模型的行为变化中,哪些部分仍可由源模型已有的跨查询响应模式解释,哪些部分已超出该解释范围。设源模型和候选模型的行为响应矩阵分别为 $\mathbf{R}_0$ 和 $\mathbf{R}_M$, 定义候选模型相对于源模型的总体行为形变:

$$\mathbf{E} = \mathbf{R}_M - \mathbf{R}_0 \quad (9)$$

建立如下源条件仿射形变模型:

$$\mathbf{R}_M = \mathbf{R}_0 \mathbf{A} + \mathbf{1} \mathbf{b}^T + \mathbf{L} \quad (10)$$

其中,行为响应维数 $D = 64$, 则 $\mathbf{R}_0, \mathbf{R}_M, \mathbf{L} \in \mathbb{R}^{K \times D}$, $\mathbf{A} \in \mathbb{R}^{D \times D}$ 为响应维度间的线性变换矩阵,用以刻画源模型不同响应维度之间可能发生的线性重组、尺度变化和耦合, $\mathbf{1} \in \mathbb{R}^{K \times 1}$ 为全 1 向量, $\mathbf{b} \in \mathbb{R}^D$ 表示与查询样本无关的响应偏移, $\mathbf{L}$ 表示无法由上述仿射变换解释的剩余行为变化。由此,候选模型响应中能够延续源模型已有行为结构的部分记为 $\mathbf{R}_0 \mathbf{A} + \mathbf{1} \mathbf{b}^T$ 。模型经过微调、压缩或知识蒸馏后,

即使响应数值发生变化,只要这些变化在固定查询集合上的组织方式仍主要延续源模型的行为结构,仍可由 $\mathbf{R}_0 \mathbf{A} + \mathbf{1} \mathbf{b}^T$ 表示。相比之下,独立训练模型虽然可能学习到相似的降水映射规律,但其面对局部守恒反事实扰动时形成的响应组织方式,可能与源模型存在较大差异。源条件仿射形变模型仅是对固定查询集合上可解释行为变化的一种参数化描述,并非新的仿射几何理论,实际 SBD 计算不需要显式估计 $\mathbf{A}$ 和 $\mathbf{b}$, 而是利用源行为响应构建参考子空间并计算正交残差。将公式(10)可改写为:

$$\mathbf{E} = \mathbf{R}_0(\mathbf{A} - \mathbf{I}_D) + \mathbf{1} \mathbf{b}^T + \mathbf{L} \quad (11)$$

其中, $\mathbf{I}_D$ 为 64 维单位矩阵,总体形变 $\mathbf{E}$ 包含两部分: $\mathbf{R}_0(\mathbf{A} - \mathbf{I}_D)$ 和 $\mathbf{1} \mathbf{b}^T$ 由源模型原有的响应结构及统一偏移产生,属于源模型能够解释的行为变化, $\mathbf{L}$ 包含超出这一变化范围的剩余部分。因此,仅根据 $\mathbf{E}$ 的整体大小无法直接判断候选模型来源。

2.4 源行为参考子空间与正交分解

根据上述行为形变模型,源模型所能解释的行为形变均可由源模型行为响应矩阵 $\mathbf{R}_0$ 的列向量以及统一偏移向量 $\mathbf{1}$ 的线性组合构成。为描述其变化范围,构造源模型基矩阵:

$$\mathbf{B}_0 = [\mathbf{R}_0, \mathbf{1}] \in \mathbb{R}^{K \times (D+1)} \quad (12)$$

其中, $\mathbf{R}_0$ 的 64 列分别描述不同空间响应维度在全部固定查询样本上的变化模式,附加的全 1 列则表示与查询样本无关的统一偏移。根据公式(11),源模型能够解释的行为形变可进一步表示为

$$\mathbf{R}_0(\mathbf{A} - \mathbf{I}_D) + \mathbf{1} \mathbf{b}^T = \mathbf{B}_0 \begin{bmatrix} \mathbf{A} - \mathbf{I}_D \\ \mathbf{b}^T \end{bmatrix} \quad (13)$$

因此,

$$\mathbf{R}_0(\mathbf{A} - \mathbf{I}_D) + \mathbf{1} \mathbf{b}^T \in \text{Col}(\mathbf{B}_0) \quad (14)$$

其中,源可解释形变严格位于 $\mathbf{B}_0$ 的列空间中,本文将 $\text{Col}(\mathbf{B}_0)$ 定义为源行为参考子空间,该子空间刻画源条件仿射形变模型下可通过源模型固有行为模式与统一偏移量解释的变化范围,并作为对候选模型行为变化进行正交分解的参考空间。接着,对 $\mathbf{B}_0$ 进行奇异值分解,

$$\mathbf{B}_0 = \mathbf{U} \Sigma \mathbf{V}^T \quad (15)$$

根据预设数值秩判定准则得 $\text{rank}(\mathbf{B}_0) = r$, 并取对应的前 r 个左奇异向量构成 $\mathbf{Q}_0 \in \mathbb{R}^{K \times r}$,

$$\text{Col}(\mathbf{Q}_0) = \text{Col}(\mathbf{B}_0), \mathbf{Q}_0^T \mathbf{Q}_0 = \mathbf{I}_r \quad (16)$$

将 $\mathbf{I}_r$ 记为 $r \times r$ 的单位矩阵。由于 $\mathbf{Q}_0$ 的列向量构成 $\text{Col}(\mathbf{B}_0)$ 的标准正交基, 因此定义源行为参考子空间上的正交投影算子及其正交补投影算子分别为

$$P_0 = \mathbf{Q}_0 \mathbf{Q}_0^T, P_0^\perp = \mathbf{I}_K - \mathbf{Q}_0 \mathbf{Q}_0^T \quad (17)$$

通过 $\mathbf{Q}_0$, 可将总体行为形变 $\mathbf{E}$ 分解为源行为参考子空间内的分量和与该空间正交的分量, 其在源行为参考子空间上的投影为:

$$\mathbf{E}_\parallel = P_0 \mathbf{E} = \mathbf{Q}_0 \mathbf{Q}_0^T \mathbf{E} \quad (18)$$

该分量对应能够由源模型行为结构解释的变化。从总体行为形变中去除 $\mathbf{E}_\parallel$, 得到源正交行为偏离:

$$\mathbf{E}_\perp = \mathbf{E} - \mathbf{E}_\parallel \quad (19)$$

进一步可写为:

$$(\mathbf{I}_K - \mathbf{Q}_0 \mathbf{Q}_0^T) \mathbf{E} = P_0^\perp \mathbf{E} \quad (20)$$

其中, $\mathbf{I}_K$ 为 K 维单位矩阵。由于 $\mathbf{Q}_0$ 为正交基, 因此可记为:

$$\mathbf{Q}_0^T \mathbf{E}_\perp = 0 \quad (21)$$

又由 $\text{Col}(\mathbf{Q}_0) = \text{Col}(\mathbf{B}_0)$ 可得, 正交补投影算子对 $\mathbf{B}_0$ 列空间内的任意分量均为 0, 即

$$P_0^\perp \mathbf{B}_0 = (\mathbf{I}_K - \mathbf{Q}_0 \mathbf{Q}_0^T) \mathbf{B}_0 = 0 \quad (22)$$

结合行为形变模型可知, $\mathbf{R}_0(\mathbf{A} - \mathbf{I}_D)$ 和 $\mathbf{1}\mathbf{b}^T$ 均位于 $\mathbf{B}_0$ 的列空间中, 因此经过正交投影后, 这两部分不会进入 $\mathbf{E}_\perp$ 。由此得到:

$$\mathbf{E}_\perp = (\mathbf{I}_K - \mathbf{Q}_0 \mathbf{Q}_0^T) \mathbf{L} \quad (23)$$

由公式(23)可知, 源条件正交分解消除了源行为参考子空间内的可解释分量, 最终 $\mathbf{E}_\perp$ 仅保留剩余行为变化 $\mathbf{L}$ 在该子空间正交补上的投影。因此, SBD 度量的是候选模型相对于源模型的结构性行为偏离, 而非全部的输出差异。本文据此提出来源判别假设: 若候选模型由源模型经过微调、剪枝、量化或知识蒸馏等方式得到, 其反事实响应结构通常会在一定程度上延续源模型已有的行为组织方式, 从而具有较小的源正交偏离; 而独立训练的模型尽管可能共享相同的任务规律, 但其局部反事实响应结构并不必然落入同一源行为空间。因此, $\mathbf{E}_\perp$ 刻画的是候选模型相对于指定源模型的结构性行为偏离, 为后续模型来源判定提供依据。此外, 源条件正交分解要求源行为参考子空间存在非平凡的正交

补空间, 即:

$$K - r > 0 \quad (24)$$

其中, 若 $K - r = 0$, 则源行为参考子空间占满整个查询样本空间, 任意行为形变均可被投影至该空间, 无法形成有效的正交偏离。因此, 实际计算时需要检查 $\mathbf{B}_0$ 的数值秩及其正交补空间维数, 以保证后续偏离度量具有可辨识性。从工程实现角度, 源基矩阵 $\mathbf{B}_0 = [\mathbf{R}_0, \mathbf{1}_K] \in \mathbb{R}^{K \times (D+1)}$, 因此其数值秩满足 $r \leq D + 1$ 。由此, $K > D + 1$ 是保证源参考子空间存在非平凡正交补的一个充分条件。当 $K \leq D + 1$ 时并不必然发生退化, 其是否可辨识仍由 $\mathbf{B}_0$ 的实际数值秩决定。

2.5 源行为偏离度与指纹判定

通过上述正交分解, 可获得候选模型相对于源模型的源正交行为偏离矩阵 $\mathbf{E}_\perp$ 。由于 $\mathbf{E}_\perp$ 同时包含 K 个查询样本和 64 个空间响应维度, 需进一步将其转化为模型级标量, 以比较不同候选模型相对于同一源模型的偏离程度。为此, 本文采用 Frobenius 范数聚合 $\mathbf{E}_\perp$ 在全部查询样本和响应维度上的偏离强度。考虑到源模型自身的行为响应幅值会影响距离尺度, 使用 $\mathbf{R}_0$ 的 Frobenius 范数进行归一化。由此, 将候选模型 M 相对于源模型 M_0 的源行为偏离度 SBD 定义为:

$$D_{\text{SBD}}(M, M_0) = \frac{\|\mathbf{E}_\perp\|_F}{\|\mathbf{R}_0\|_F}, \|\mathbf{R}_0\|_F > 0 \quad (25)$$

公式(25)以源模型自身的行为响应尺度为参照, 度量候选模型行为形变在源行为参考子空间正交方向上的相对强度。SBD 越小, 说明候选模型的行为变化越多地落在源模型能够解释的范围内, 因而越符合源派生模型的行为特征; SBD 越大, 则说明候选模型具有更明显的源正交行为偏离, 更倾向于独立训练模型。

在模型实例级验证中, 设方法开发和校准阶段确定的判别阈值为 τ 。当 $D_{\text{SBD}}(M, M_0) \leq \tau$ 时, 将候选模型判定为源派生模型; 反之, 则判定为独立训练模型。固定阈值用于给出具体模型的来源判定结果, ROC-AUC 等与单一阈值无关的评价指标则用于衡量 SBD 对两类模型的整体区分能力。

3 实验结果与分析

3.1 实验设置

本文实验数据基于欧洲中期天气预报中心发布的 ERA5 单层再分析资料构建^[31], 选取 2010-2014 年数据, 经日尺度聚合和空间切分形成降水超分辨率样本, 包含降水量、2 m 气温、地表气压、10 m 风速、相对湿度和陆海掩膜等变量。在黑盒指纹验证阶段, 预先从候选池中固定 $K = 800$ 个查询样本, 其中训练集样本 605 个、验证集样本 195 个; 查询样本及其顺序、反事实生成规则均在候选模型评分前确定, 且不依据候选模型输出或派生关系标签进行调整。每个查询样本构造正、负两个局部守恒反事实端点, 因此每个候选模型共进行 1600 次预测; 源模型与所有候选模型采用完全一致的查询集合和行为响应构造流程, 以保证模型间比较的一致性。本文以 PACSSRNet^[32]作为源模型, 知识蒸馏模型及其配对的独立训练模型采用与源模型结构相异的 StudentSRNet, 其余派生模型及同主体架构独立训练模型均基于 PACSSRNet 构建。所有方法均在相同的候选模型集合及统一的物理输出定义下进行评估, 所有模型的推理过程均通过 CUDA 统一执行, 以避免候选模型构成及运行接口差异对实验结果造成影响。

候选模型共包含 28 个, 其中 18 个属于本文定义的源派生模型, 10 个为独立训练模型, 具体类型、数量及编号见表 1。派生模型覆盖微调、剪枝恢复^[33]、低比特量化^[34]及知识蒸馏^[35]等典型派生方式。其中, 5 个微调模型均以源模型的已训练参数为初始化, 在保持网络结构不变的情况下对全部可训练参数继续优化, 采用 AdamW 优化器, 学习率为 1×10^{-5} , 沿用源模型训练时的复合损失函数,

并通过不同随机种子生成 5 个独立微调实例。剪枝模型采用 10% 和 20% 的全局 L1 非结构化权重剪枝并进行掩码约束下的恢复训练, 量化模型采用 6 bit 和 4 bit 对称权重伪量化。独立模型则包括与源模型同架构独立训练的 PACSSRNet^[32], 以及与蒸馏学生模型在架构和随机种子上逐一匹配但未使用教师输出训练的 StudentSRNet。5 组蒸馏模型与其对应的独立训练模型共享相同随机种子, 构成严格配对, 用以检验 SBD 方法在区分网络架构相同、训练来源不同的模型时的有效性。判别阈值由 19 个校准模型确定, 其中源派生模型 7 个、独立模型 12 个, 取校准集中正类 SBD 最大值与负类 SBD 最小值的中间点作为固定阈值^[18]。校准模型仅用于阈值确定, 不参与后续 28 个候选模型的性能评价。实验流程首先执行无标签的距离计算与模型级预测, 并进行完整性校验, 随后再载入派生关系标签以计算评价指标, 该流程与严格黑盒完整性验证中预先固定协议、测试后揭示标签的原则相一致。

为了综合评估本文所提方法的性能, 本文将 ModelDiff^[4]、DeepJudge^[6]以及面向图像恢复任务的 Fingerprinting-IR^[10]迁移至同一降水超分辨率模型集合中进行对比。其中, SBD 与 ModelDiff 在验证阶段仅依赖模型最终输出, 而 DeepJudge 和 Fingerprinting-IR 需额外访问模型梯度。本文采用准确率 (accuracy, ACC)、平衡准确率 (balanced accuracy, BA)、精确率、召回率 (recall)、特异度 (specificity, Spec)、F1-Score、Matthews 相关系数 (matthews correlation coefficient, MCC)、Point Gap、ROC 曲线下面积 (ROC-AUC) 及平均精确率 (average precision, AP) 作为评价指标。其中, Point Gap 为独立训练模型最小 SBD 距离与源派生

表 1

候选模型组成

模型类别	模型类型	数量	模型编号
源派生	微调	5	C001-C005
源派生	10% 剪枝	3	C006-C008
源派生	20% 剪枝	3	C009-C011
源派生	6 bit 量化	1	C012
源派生	4 bit 量化	1	C013
源派生	知识蒸馏	5	C014,C015,C023-C025
独立训练	同主体架构独立训练	5	C016-C020
独立训练	学生架构 scratch 训练	5	C021,C022,C026-C028

模型最大 SBD 距离之差, 值越大, 表明两类极端样本的分离程度越高。ROC-AUC 用于衡量连续指纹分数对两类模型的整体排序能力。鉴于正负样本数量存在不平衡, 本文同时报告平衡准确率及精确率-召回率 (precision-recall, PR) 相关指标, 以降低类别比例对准确率单一指标解释性的影响。由于 SBD 距离越小表明候选模型越倾向于具有源派生关系, 在计算 ROC 与 PR 曲线时, 采用距离的相反数作为正类得分。

3.2 对比分析

本文将 K 定义为查询样本数, 当 $K = 800$ 时 SBD 方法对 28 个候选模型上得到真正例 15 个、假负例 3 个、真负例 10 个、假正例 0 个, 整体准确率为 89.29%, 平衡准确率为 91.7%, 精确率为 100%, 召回率为 83.33%, 特异度为 100%, F1-Score 为 0.909, MCC 为 0.801。连续指纹分数的 ROC-AUC 为 0.978, 平均精确率 (AP) 为 0.988。结果表明, SBD 在严格黑盒条件下具有较高的来源区分能力, 且当前测试集未出现独立训练模型被误判的情况。

表 2 不同评价指标下的对比结果

方法	ACC /%	BA /%	Recall /%	Spec /%	ROC -AUC	AP
SBD	89.29	91.7	83.33	100	0.978	0.988
DeepJudge ^[6]	78.57	83.3	66.67	100	0.889	0.917
ModelDiff ^[4]	57.14	44.4	88.89	0	0.733	0.888
Fingerprint-IR ^[10]	39.29	50.6	11.11	90	0.617	0.747

表 2 为各模型指纹方法在统一实验设置下的来源验证结果, 当 $K = 800$ 时, ModelDiff 的召回率达到 88.89%, 但将全部 10 个独立训练模型误判为源派生模型, 特异度为 0, 表明基于行为相似度的更易受共同任务映射规律的干扰。DeepJudge 虽能完全拒绝独立模型, 但对派生模型的召回率仅为 66.67%。Fingerprinting-IR 在 18 个派生模型中仅正确识别 2 个, 表明基于自然图像恢复任务关键统计特征设计的指纹方法, 在迁移至连续降水场后适配性受到限制。相比之下, SBD 的准确率较 ModelDiff、DeepJudge 和 Fingerprinting-IR 分别提升 32.15、10.72 和 50.00 个百分点, ROC-AUC 分别提高 0.245、0.089 和 0.361, 且在仅依赖最终输出访问的条件下仍保持较高的综合判别性能。

从派生方式来看, SBD 对 5 个微调模型、6 个

剪枝恢复模型及 6 bit、4 bit 量化模型均实现 100% 识别, 表明参数微调、稀疏化与低比特量化虽改变模型参数及局部数值响应, 但派生模型在反事实扰动下的行为仍保留较强的源结构特征。值得注意的是, 4 bit 量化模型 C013 的 SBD 值为 0.10846, 显著高于同架构其他派生模型, 但仍低于冻结阈值, 说明强量化会加剧行为偏离, 却未完全破坏源模型的行为指纹。相反, 5 个与源模型同结构的独立训练模型, 其 SBD 值介于 0.12567 至 0.14388 之间, 全部高于阈值。上述结果表明, 单纯共享网络结构与训练数据不足以构成有效的指纹证据, 局部扰动下无法由源行为参考空间解释的偏离, 能够提供额外的判别信息。

为进一步考察相同学生网络架构下不同训练来源对 SBD 判别结果的影响, 本文对 5 组知识蒸馏学生模型及其对应随机种子配对的独立从头训练 (scratch) 学生模型进行了对比分析, 其 SBD 距离及冻结阈值 τ 的关系如图 2 所示。

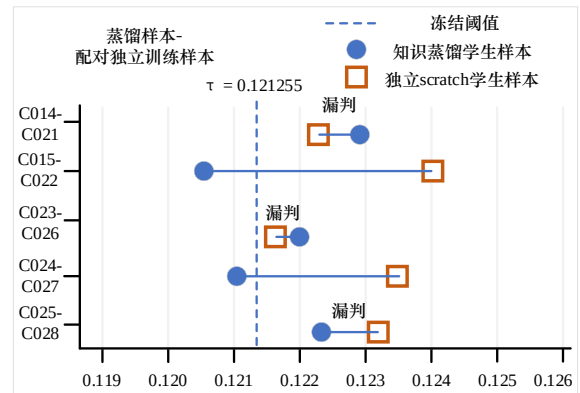


图 2 知识蒸馏学生与配对独立学生模型的 SBD 距离比较

在全部 5 组架构与随机种子均一致的配对模型中, 独立从头训练的学生模型, 其 SBD 距离均位于冻结阈值 τ 之上, 表明其行为形变中存在稳定的、无法由源模型行为结构解释的正交偏离成分。相比之下, 知识蒸馏学生模型的 SBD 距离整体上更接近判别边界, 其中 2 组位于阈值下方, 被正确判别为源派生模型; 其余 3 组则略高于阈值, 构成当前实验中的 3 个假负例。上述结果表明, 知识蒸馏能够在一定程度上将源模型的行为结构迁移至学生模型, 使其 SBD 响应更趋近于源派生区域。

然而, 即便学生网络架构和随机种子与蒸馏过程严格对齐, 在未借助教师输出监督条件下独立训

训练的模型仍整体位于判别阈值的另一侧,进一步验证共享网络结构与训练数据分布本身并不足以构成有效的指纹依据,而局部扰动下无法由源行为参考空间解释的行为偏离,才是区分派生关系的关键。进一步分析表明,当前3个假负例均来源于知识蒸馏模型,而微调、剪枝与量化模型均被正确识别。该现象可能与知识蒸馏的跨架构知识迁移方式有关:蒸馏主要约束学生模型对教师输出映射的逼近,但无法保证其在局部守恒反事实扰动下的响应结构能够被完整继承。同时,StudentSRNet与源模型PACSSRNet的网络结构不同,表达能力及归纳偏差差异也可能进一步改变局部扰动响应。因此,蒸馏学生模型的SBD整体虽较配对scratch模型更接近源派生区域,但部分模型仍位于冻结阈值之外,说明跨架构知识蒸馏是当前方法的主要识别难点。

为考察行为响应空间分辨率对SBD验证性能的影响,在保持查询预算 $K=800$ 和扰动强度 $\varepsilon=0.2$ 不变的条件下,分别采用 4×4 、 8×8 和 16×16 自适应平均池化对行为响应场进行压缩,对应得到16维、64维和256维行为表示。各分辨率设置下的实验结果如图3所示。

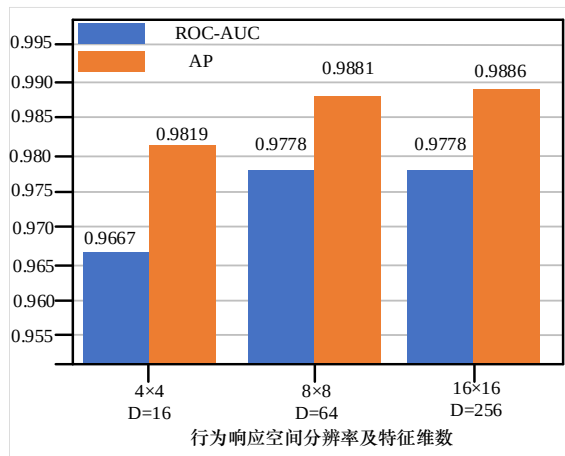


图3 不同行为响应空间分辨率下SBD性能与特征维数比较

从排序性能来看, 4×4 分辨率下的ROC-AUC和AP分别为0.9667和0.9819,明显低于 8×8 分辨率下的0.9778和0.9881,说明过度压缩行为响应会损失与模型来源相关的空间结构信息。相比之下,将分辨率进一步提高至 16×16 后,ROC-AUC仍为0.9778,AP仅由0.9881小幅提升至0.9886,增益非常有限。从表示复杂度来看,行为特征维数

由 8×8 的64维增加到 16×16 的256维,同时源行为参考子空间的数值秩也由65增加到257,说明更高分辨率会显著增加表示复杂度,却未带来实质性的整体排序性能提升。综合来看, 4×4 分辨率的信息保留不足,而 16×16 分辨率虽然保留了更细的空间细节,但性能提升已基本饱和,同时显著增加了行为表示维数和源参考空间复杂度。因此,最终采用 8×8 行为响应网格作为默认设置,在验证性能和表示复杂度之间取得较好的平衡。

为了分析局部反事实扰动位置选择方式对SBD指纹判别能力的影响,在保持查询预算 $K=800$ 、扰动系数 $\varepsilon=0.2$ 、 8×8 行为响应网格及其他计算流程不变的条件下,比较最大共享容量选择(Max-Capacity)、首个可行位置选择(First-Feasible)和随机可行位置选择(Random-Feasible)三种策略。其中,Max-Capacity按本文正式方法选择共享可转移容量 $c=\min(P_a|P_b)$ 最大的水平相邻格点对,其余两种分别选择固定扫描顺序中的首个可行格点对以及从全部可行格点对中确定性随机选择一个位置。

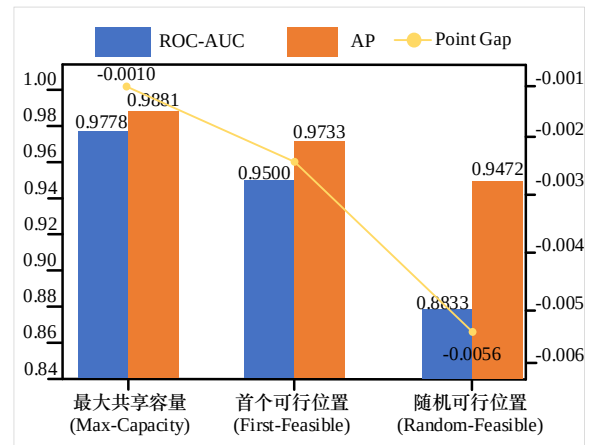


图4 局部扰动位置选择策略对SBD验证性能的影响

如图4所示,Max-Capacity策略的ROC-AUC和AP分别为0.9778和0.9881,均高于First-Feasible(0.9500和0.9733)及Random-Feasible策略(0.8833和0.9472)。同时,Max-Capacity的极端样本Point Gap为-0.000997,其重叠程度也小于First-Feasible的-0.002571和Random-Feasible的-0.005552,说明最大共享容量策略能够获得更稳定的模型来源排序。

为进一步分析性能差异的来源,对三种策略的

查询统计量也进行比较。三种策略的有效查询比例均为 99.5%，表明性能差异并非由有效查询样本数量的不同所致。Max-Capacity 所选择格点对的平均共享容量为 10.4046 mm，对应平均转移量为 2.0809 mm，明显高于 First-Feasible 的 1.4330 mm 和 0.2866 mm，以及 Random-Feasible 的 1.8827 mm 和 0.3765 mm。上述结果表明，在维持局部非负性与降水总量守恒的条件下，Max-Capacity 策略能够实现更充分的局部空间重新分配，从而更有效地激发模型间的行为响应差异。

3.3 查询预算与稳定性分析

为分析查询预算对 SBD 验证性能及结果稳定性的影响，本文在固定 $K = 800$ 的查询集合内，分别设置 $K = 50$ 、100、200、400 和 800 进行实验。为降低查询样本选择随机性带来的偶然影响，对 $K < 800$ 的设置采用 30 次确定性随机排列的嵌套前缀进行重采样，其余行为表示和距离计算过程保持不变。考虑到查询数量变化会影响 SBD 的距离尺度，小预算实验不重新校准判别阈值，主要采用 ROC-AUC、AP 以及与 $K = 800$ 完整距离排序的 Spearman 相关系数评价不同查询预算下的判别性能和排序稳定性，实验结果如图 5 所示。

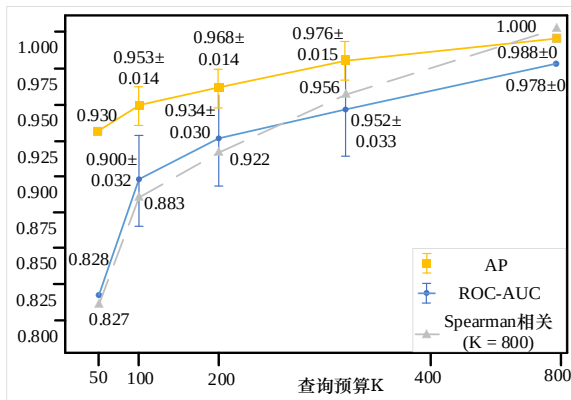


图5 不同查询预算下 SBD 的性能与排序稳定性

当查询预算过小，SBD 会受到源参考空间可辨识条件的约束；而当查询预算增加至 $K = 100$ 后，30 次重复均能够满足可辨识条件。从图 5 可以直观看出，ROC-AUC、AP 以及与 $K = 800$ 完整距离排序的 Spearman 相关系数总体均随查询预算增加而提高。 $K = 100$ 时，平均 ROC-AUC 和 AP 分别为 0.900 和 0.953，Spearman 相关系数为 0.883；当 K 增至 200 时，三项指标分别提高至 0.934、0.968

和 0.922； $K = 400$ 时达到 0.952、0.976 和 0.956；完整 $K = 800$ 设置下 ROC-AUC 和 AP 最终达到 0.978 和 0.988，候选模型距离排序与完整结果完全一致。需要注意的是，当 $K = 50$ 时，30 次重复中仅有 1 次满足源正交补可辨识条件，因此该设置仅能反映低查询预算下的退化情况，不足以提供稳定的性能评估。

3.4 不同源模型下的泛化验证

为验证 SBD 对不同降水超分辨率模型的适用性，并验证结论是否依赖 PACSSRNet 特定网络结构，本文进一步选取 DeepSD^[36]、气候降尺度模型 (Climate Downscaling, CD)^[37] 及基于扩散生成模型 EDM^[38-39] 的扩散降尺度模型作为扩展源模型。上述三个模型分别代表卷积式确定性降尺度、注意力增强超分辨率与扩散生成等不同建模范式，并在数据集、数据划分及降尺度任务设置方面均与 PACSSRNet 保持一致。针对每个扩展源模型，分别构建 3 个全参数微调模型、10% 和 20% 剪枝恢复模型各 1 个、6 bit 和 4 bit 量化模型各 1 个，以及 3 个同架构独立训练模型，共形成 7 个源派生模型和 3 个独立训练模型。验证阶段统一采用原实验冻结的 $K = 800$ 查询集合及相同的反事实行为响应构造流程，并采用 ROC-AUC、AP 和 Point Gap 评价不同源模型条件下的整体区分能力。对于 EDM 扩散模型，为避免生成随机性干扰反事实响应计算，对同一查询的正、负端点采用相同的采样随机量，并在源模型及全部候选模型间复用一致的查询级随机噪声流。

表3 SBD 方法在不同降水超分模型上的泛化性能

源模型	派生/独立	ROC-AUC	AP	Point Gap
CD ^[37]	7/3	1.000	1.000	0.091
EDM ^[38-39]	7/3	0.905	0.968	-0.002
DeepSD ^[36]	7/3	0.714	0.910	-0.314

由表 3 可知，SBD 在不同源模型下表现出不同程度的来源区分能力。对于 CD 模型，其 ROC-AUC 与 AP 均达到 1.000，Point Gap 为 0.091，源派生模型与独立训练模型实现完全分离；EDM 扩散模型的 ROC-AUC 和 AP 分别达到 0.905 和 0.968，Point Gap 仅为 -0.002，表明两类模型在判别边界附近虽存在轻微重叠，但整体排序仍保持较高区分

度。相比之下, DeepSD 上的 ROC-AUC 下降至 0.714, 但 AP 仍达到 0.910。上述结果表明, SBD 的来源判别能力并非仅依赖 PACSSRNet 特定网络结构, 在确定性卷积、注意力增强及扩散生成等不同模型下均具备一定的迁移性, 但其验证效果会受到具体源模型及派生操作对行为结构保持程度的影响。为进一步分析 DeepSD 上区分性能下降的原因, 图 6 给出了各候选模型相对于源模型的验证集 PSNR 变化及对应 SBD 距离。

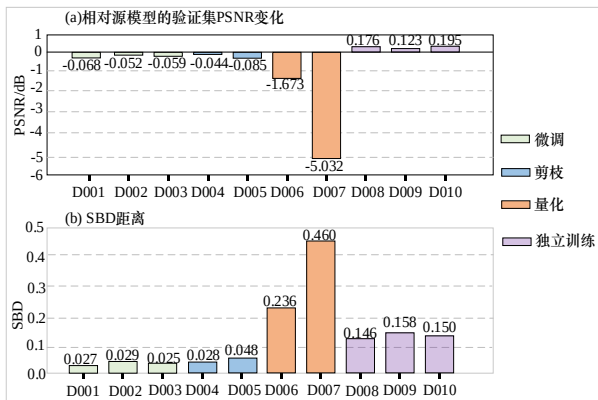


图 6 DeepSD 候选模型的任务性能与 SBD 距离

可以看到, 3 个微调模型和 2 个剪枝恢复模型的 PSNR 相较源模型仅变化 $-0.085 \sim -0.044$ dB, 其 SBD 亦保持在 $0.025 \sim 0.048$ 的范围。相比之下, 6 bit 和 4 bit 量化模型的 PSNR 分别下降 1.673 dB 和 5.032 dB, SBD 同步升高至 0.236 和 0.460, 超过 3 个独立训练模型的 $0.146 \sim 0.158$, 从而构成 DeepSD 组主要的排序反转。若仅考察微调 and 剪枝派生模型, 其最大 SBD 为 0.048, 仍明显低于独立训练模型的最小 SBD 0.146。该结果说明, DeepSD 对强低比特量化具有较高敏感性, 当量化操作已显著损害模型预测性能时, 其局部反事实行为结构也可能发生较大改变, 使源行为指纹难以继续稳定保持。另一方面, 3 个独立训练 DeepSD 模型在验证集上的 PSNR 与源模型相当甚至略高, 但其 SBD 仍明显高于微调和剪枝派生模型, 这说明 SBD 并非单纯的刻画任务性能, 而是候选模型相对于特定源模型的行为结构偏离。综上, DeepSD 上的性能下降主要集中于强量化派生模型, 反映出本文方法在源行为结构受到显著破坏条件下的识别边界。

3.5 消融实验

为分析 SBD 各组成部分对模型来源验证性能

的影响, 本文分别对源正交分解、仿射偏移建模、残差聚合方式和局部反事实查询进行消融实验, 并在每次实验中仅改变一个组成部分。具体设置如下: DirectDifference-F (DDF) 去除源行为参考子空间上的正交分解, 直接采用候选模型与源模型行为响应之差计算 Frobenius 距离; NoConstantColumn-F (NCC) 在构造源参考空间时去除全 1 常数列, 用于分析统一响应偏移建模的影响; NuclearNorm (NN) 保留完整的源正交分解过程, 仅将 Frobenius 范数替换为核范数进行残差聚合; NoCounterfactual-F (NOCF-F) 则取消局部守恒反事实成对查询, 直接使用普通未扰动输入的模型输出构造行为表示。前三种变体按照与完整 SBD 相同的校准流程分别确定判别阈值, 其实验结果见表 4。

表 4 SBD 核心结构消融结果

方法	ACC/%	BA/%	Recall/%	F1	MCC	ROC-AUC	AP
SBD	89.29	91.7	83.33	0.909	0.801	0.978	0.988
DDF	82.14	86.1	72.22	0.839	0.694	0.989	0.994
NCC	89.29	91.7	83.33	0.909	0.801	0.978	0.988
NN	85.71	88.9	77.78	0.875	0.745	0.972	0.985

通过表 4 可知, 当去除源正交分解后, DDF 的 ROC-AUC 和 AP 分别取得 0.989 和 0.994, 略高于完整 SBD, 表明直接行为差异仍具有较强的整体排序能力; 但其准确率和平衡准确率分别下降至 82.14% 和 86.1%, 召回率由 83.33% 降至 72.22%, MCC 下降至 0.694。两种方法均能够正确拒绝全部独立训练模型, 性能差异主要来自源派生模型的漏判。由此可见, 源正交分解的主要作用并非单纯提高连续距离的整体排序能力, 而是在固定判别条件下过滤可由源模型行为结构解释的变化, 从而降低源派生模型的漏判。接着, 将 Frobenius 范数替换为核范数后, NN 的准确率、平衡准确率和召回率分别下降至 85.71%、88.9% 和 77.78%, MCC 为 0.745, 整体低于完整 SBD, 表明采用 Frobenius 范数聚合全部查询样本和响应维度上的正交偏离, 更适合当前行为响应表示。NCC 与完整 SBD 的各项指标保持一致, 表明统一响应偏移在当前候选模型集合中并非主要判别因素, 但保留常数列能够使源参考空间同时描述线性变化和统一平移, 从而保持源仿射行为建模形式的完整性。

为进一步验证局部守恒反事实查询的作用，还将完整 SBD 与 NOCF-F 进行对比，结果见表 5。采用局部守恒反事实响应时，源派生模型最大 SBD 距离为 0.12265，独立训练模型最小距离为 0.12165，表明两类极端样本之间仍有一定的重叠。取消反事实查询后，两者分别变为 0.02133 和 0.01764，正负类样本重叠程度明显增大。同时，NOCF-F 的 ROC-AUC 和 AP 分别下降至 0.817 和 0.917，明显低于完整 SBD 结果。

需要指出的是，Point Gap 反映两类极端样本之间的局部分离程度，而 ROC-AUC 和 AP 衡量全部候选模型的整体排序性能。因此，完整 SBD 的 Point Gap 虽略小于 0，但较 NOCF-F 已明显改善，同时其 ROC-AUC 和 AP 保持较高水平，说明少量边界样本重叠并未改变整体排序优势。上述结果表明，在普通未扰动输入条件下，不同模型易受共同任务映射规律的影响，仅依赖其最终输出时，两类模型的极端样本重叠更加明显。局部守恒反事实查询通过主动改变局部降水空间分布，在保持非负性与局部降水总量守恒的前提下，激发模型响应差异，从而增强与模型来源相关的行为特征。综合各项消融实验结果，局部反事实查询能够有效增强与模型来源相关的行为差异；源正交分解进一步改善固定判别下对源派生模型的识别稳定性；Frobenius 范数适用于聚合正交行为偏离度量；常数列则用于表示所有查询样本共同发生的响应偏移。

4 结论

本文面向仅可获取最终预测输出的降水超分辨率模型黑盒指纹验证问题，提出一种基于源行为偏离度的验证方法。该方法通过局部守恒反事实成对查询提取模型行为响应，利用源模型响应构建源行为参考子空间，并以候选模型行为变化中无法由该子空间解释的正交分量作为结构性偏离的度量依据。实验结果表明，在包含 28 个候选模型的测试集合上，SBD 能够有效识别绝大多数源派生模型，并准确拒绝全部独立训练模型，整体准确率、平衡准确率及 ROC-AUC 均处于较高水平。消融实验进

一步表明，局部反事实查询机制能够有效削弱共同任务规律所造成的行为相似干扰，减小源派生模型与独立训练模型之间的极端样本重叠，并显著改善连续指纹分数的整体排序性能。

表 5

局部反事实查询消融结果

行为响应形式	最大正类距离	最小负类距离	Point Gap	ROC-AUC	AP
局部反事实响应（本文方法）	0.12265	0.12165	-0.00100	0.978	0.988
普通未扰动输出	0.02133	0.01764	-0.00369	0.817	0.917



袁程胜 (1989-), 男, 博士, 南京信息工程大学副教授, 主要研究方向为人工智能安全、信息隐藏与人工智能+气象。



孔令志 (2001-), 男, 南京信息工程大学硕士生在读, 主要研究方向为大模型版权保护、图像超分。



李欣亭 (1991-), 女, 博士, 国防科技大学外国语学院副教授, 主要研究方向为数据挖掘, 多媒体内容安全与认知战。



宋甫元 (1991-), 男, 博士, 南京信息工程大学副教授, 主要研究方向为隐私保护、应用密码学与区块链安全。



刘庆程 (1987-), 男, 南京信息工程大学博士生, 主要研究方向为多媒体内容安全、电力大数据挖掘与分析。



付章杰 (1983-), 男, 博士, 南京信息工程大学教授, 主要研究方向为信息安全、区块链安全与权属保护。

参考文献:

- [1] Price I, Rasp S. Increasing the accuracy and resolution of precipitation forecasts using deep generative models[C]//Proceedings of the 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, 2022, 151: 10555-10571.
- [2] Reddy P J, Matear R, et al. A precipitation downscaling method using a super-resolution deconvolution neural network with step orography[J]. Environmental Data Science, 2023, 2: e17.
- [3] He Z, Zhang T, Lee R B. Sensitive-Sample Fingerprinting of Deep Neural Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4729-4737.
- [4] Li Y, Zhang Z, Liu B, et al. ModelDiff: Testing-based DNN similarity comparison for model reuse detection[C]//Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis. 2021: 139-151.
- [5] Peng Z, et al. Fingerprinting Deep neural networks globally via universal adversarial perturbations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13430-13439.
- [6] Chen J, Wang J, Peng T, et al. Copy, Right? A Testing Framework for Copyright Protection of Deep Learning Models[C]//2022 IEEE Symposium on Security and Privacy. 2022: 824-841.
- [7] Guan J, Liang J, He R. Are You Stealing My Model? Sample Correlation for Fingerprinting Deep Neural Networks[C]//Advances in Neural Information Processing Systems. 2022, 35: 36571-36584.
- [8] Xu T, Wang C, et al. United We Stand, Divided We Fall: Fingerprinting Deep Neural Networks via Adversarial Trajectories[C]//Advances in Neural Information Processing Systems. 2024, 37: 69299-69328.
- [9] Bai X, Li S, et al. SDBF: Steep-Decision-Boundary Fingerprinting for Hard-Label Tampering Detection of DNN Models[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 29278-29287.
- [10] Quan Y, Teng H, Xu R, et al. Fingerprinting Deep Image Restoration Models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 13285-13295.
- [11] 袁程胜, 郭强, 付章杰. 基于差分隐私的深度伪造指纹检测模型版权保护算法[J]. 通信学报, 2022, 43(9): 181-193.
Yuan C S, Guo Q, Fu Z J. Copyright protection algorithm based on differential privacy deep fake fingerprint detection model[J]. Journal on Communications, 2022, 43(9): 181-193.
- [12] Kim B, Lee S, Lee S, et al. Margin-based neural network watermarking [C]//Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, 2023, 202: 16696-16711.
- [13] Lv P, Ma H, Chen K, et al. MEA-Defender: A robust watermark against model extraction attack[C]//2024 IEEE Symposium on Security and Privacy. IEEE, 2024: 2515-2533.
- [14] Zhou H, Liu X, Shen Z. Frequency-Embedded Watermarking: Defending vision models from extraction attacks[J]. Neurocomputing, 2025, 647: 130456.
- [15] Yuan Z, Li L, et al. Watermarking for Stable Diffusion Models[J]. IEEE Internet of Things Journal, 2024, 11(21): 35238-35249.
- [16] Yuan C, Cao L, et al. Zero-Knowledge Proof-Based IP Protection of Visual Large Models of Autonomous Driving[J]. IEEE Transactions on Information Forensics and Security, 2026, 21: 6973-6984.
- [17] Bai X, He C, Ma X, et al. Intersecting-boundary-sensitive fingerprinting for tampering detection of DNN models[C]//Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, 2024, 235: 54402-54413.
- [18] He C, Bai X, et al. Towards stricter black-box integrity verification of deep neural network models[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 9875-9884.
- [19] Sinita S, Fried O. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 4067-4076.
- [20] Song H J, Khayatkhoei M, AbdAlmageed W. ManiFPT: Defining and analyzing fingerprints of generative models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 10791-10801.
- [21] Teng H, Quan Y, et al. Fingerprinting denoising diffusion probabilistic models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025: 28811-28820.
- [22] Yamabe S, Waseda F K, Takahashi T, et al. MergePrint: Merge-resistant fingerprints for robust black-box ownership verification of large language models[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Vienna: Association for Computational Linguistics, 2025: 6894-6916.
- [23] Li Q, Xu G. Deep neural network fingerprinting by general examples [J]. Neurocomputing, 2026, 665: 132184.
- [24] 王延文, 雷为民, 等. 基于生成模型的视频图像重建方法综述[J]. 通信学报, 2022, 43(9): 194-208.
Wang Y W, Lei W M, Zhang W, et al. Survey on video image reconstruction method based on generative model[J]. Journal on Communications, 2022, 43(9): 194-208.
- [25] 寇旗旗, 刘规, 等. 基于多域信息增强的轻量级图像超分辨率网络 [J]. 通信学报, 2025, 46(4): 144-159.
Kou Q Q, Liu G, Jiang H, et al. Lightweight image super-resolution network based on multi-domain information enhancement[J]. Journal on Communications, 2025, 46(4): 144-159.
- [26] Xia H, Wang K. PreciDBPN: A customized deep learning approach for hourly precipitation downscaling in eastern China[J]. Atmospheric Research, 2024, 311: 107705.
- [27] Srivastava P, Yang R, Kerrigan G, et al. Precipitation downscaling with spatiotemporal video diffusion[C]//Advances in Neural Information Processing Systems. 2024, 37: 56374-56400.
- [28] Glawion L, Polz J, Kunstmann H, et al. Global spatio-temporal ERA5 precipitation downscaling to km and sub-hourly scale using generative AI[J]. npj Climate and Atmospheric Science, 2025, 8: 219.
- [29] Yi C, Yu M, Qian W, et al. Efficient kilometer-scale precipitation downscaling with conditional wavelet diffusion[J]. Journal of Geophysical Research: Machine Learning and Computation, 2026, 3: e2025JH000941.
- [30] 齐佩汉, 丁瀚磊, 等. 异构联邦双向知识蒸馏传递半监督调制类型识别[J]. 通信学报, 2023, 44(11): 183-200.
Qi P H, Ding Y L, Yin K, et al. Heterogeneous federated bidirectional knowledge distillation transfer semi-supervised modulation recognition [J]. Journal on Communications, 2023, 44(11): 183-200.

- [31] Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis[J]. Quarterly Journal of the Royal Meteorological Society, 2020, 146 (730): 1999-2049.
- [32] Yuan C S, Kong L Z, Liu Q C, Tang Z F, Cao Y, Li X T. PACSSRNet: Physics-Aware Cross-Scale Spectral Network for Precipitation Downscaling Super-Resolution[C]//2026 IEEE International Symposium on AI and Cybersecurity (ISAICS). 2026. Accepted.
- [33] Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C]//Advances in Neural Information Processing Systems. 2015: 1135-1143.
- [34] Jacob B, Kligys S, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2704-2713.
- [35] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[EB/OL]. arXiv:1503.02531, 2015.
- [36] Vandal T, et al. DeepSD: generating high-resolution climate change projections through single-image super-resolution[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 1663-1672.
- [37] Chiang C H, Huang Z H, et al. Climate downscaling: a deep-learning based super-resolution model of precipitation data with attention block and skip connections[EB/OL]. arXiv:2403.17847, 2024.
- [38] Karras T, Aittala M, Aila T, et al. Elucidating the design space of diffusion-based generative models[C]//Advances in Neural Information Processing Systems. 2022, 35: 26565-26577.
- [39] Watt R A, Mansfield L A. Generative diffusion-based downscaling for climate[EB/OL]. arXiv:2404.17752, 2024.