计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2000/00(00)-0000-00
doi: 10.3778/j.issn.1673-9418.0000000

多模态协同增强驱动的分布外检测方法

冯鸣旭[1], 徐浩楠[1], 张宇萱[1], 田奇[2], 杨杨[1+]

1. 南京理工大学 计算机科学与工程学院, 南京 210094
2. 中国电子科技集团公司第二十九研究所, 四川 成都 610036
+ 通信作者 E-mail: yyangy@njust.edu.cn

摘 要: 分布外检测是保障深度模型在开放环境下可靠部署的关键。现实应用场景通常呈现多模态特性，单一模态难以提供充分的判别依据，因此如何有效融合多模态信息以提升分布外检测性能已成为研究热点。然而，现有方法多依赖诱导跨模态预测分歧来检测未知样本，存在分歧调控不精准、模态协同信息挖掘不足等问题，导致模型在融合决策阶段的检测性能受限。为应对上述挑战，提出了一种多模态协同增强驱动的分布外检测方法。该方法在分支级协同中引入以均匀分布先验为基准的细粒度分歧引导机制，抑制各模态在非目标类别上的虚假高置信度响应，实现跨模态预测分歧的精准调控；在融合级协同中设计置信度边界约束，促使融合预测的判别置信度受制于单一模态置信度下界，以此挖掘跨模态互补鉴别性信息并缓解联合决策的过度自信问题。在 HMDB51 与 Kinetics-600 等主流基准上的实验结果表明，所提方法显著降低了假正例率并提升了检测精度，在开放环境下展现出优于现有基线的鲁棒性与有效性。

关键词: 分布外检测；多模态学习；跨模态分歧；不确定性估计

文献标志码: A　　**中图分类号**: TP183

Multimodal Collaborative Enhancement-Driven Out-of-Distribution Detection

FENG Mingxu[1], XU Haonan[1], ZHANG Yuxuan[1], TIAN Qi[2], YANG Yang[1+]

1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
2. The 29th Research Institute of China Electronics Technology Group Corporation, Chengdu, Sichuan 610036, China

Abstract: Out-of-distribution (OOD) detection is crucial for the reliable deployment of deep models in open environments. Since real-world scenarios are inherently multimodal and a single modality rarely provides sufficient evidence for discrimination, effectively fusing multimodal information to improve OOD detection has become an active research topic. However, existing methods mostly rely on inducing cross-modal predictive divergence to identify unknown samples, yet they regulate such divergence imprecisely and fail to fully exploit the complementary information across modalities, which limits their performance at the fusion decision stage. To address these issues, this paper proposes a multimodal collaborative enhancement-driven OOD detection method. At the branch level, a fine-grained divergence guidance mechanism is introduced, which anchors a uniform prior as a reference to suppress spurious high-confidence predictions of each modality on non-target classes, thereby enabling precise regulation of cross-modal predictive divergence. At the fusion level, a confidence boundary constraint is introduced to upper-bound the fused prediction confidence by the minimum confidence among the unimodal branches, thereby exploiting complementary cross-modal discriminative information while mitigating overconfidence in the joint decision. Experimental results on mainstream benchmarks including HMDB51 and Kinetics-600 show that the proposed method markedly reduces the false positive rate and improves detection accuracy over existing baselines, demonstrating superior robustness and effectiveness in open environments.

Key words: out-of-distribution detection; multimodal learning; cross-modal divergence; uncertainty estimation

基金项目：国家自然科学基金（62276131）；江苏省自然科学基金（BK20240081）；江苏本科高校人工智能通识课程教学改革研究专项课题（ZNT-10）。

This work was supported by the National Natural Science Foundation of China (No. 62276131), the Natural Science Foundation of Jiangsu Province (No. BK20240081), and the Special Research Project on Teaching Reform of General Artificial Intelligence Courses in Jiangsu Undergraduate Universities (No. ZNT-10).

收稿日期：2026-05-31　　**修回日期**：2000-00-00

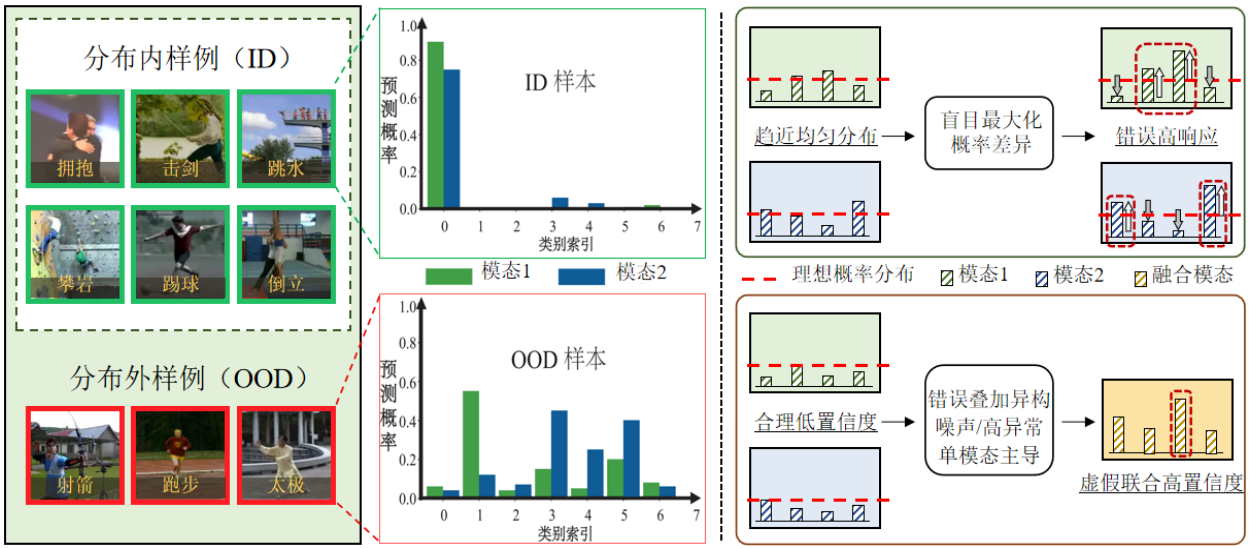


图 1　多模态分布外检测的跨模态响应模式及现有方法局限性

Fig.1　Cross-modal response patterns in multimodal OOD detection and the limitations of existing methods

深度神经网络在计算机视觉[1]、自然语言处理[2-3]及多模态学习[4]等领域已取得显著进展，但其常规训练范式高度依赖封闭世界假设，即测试与训练数据服从独立同分布。在真实开放环境中，模型不可避免地会遭遇分布外（out-of-distribution, OOD）输入，并因缺乏对未知样本的辨别能力而将其高置信度地误判为已知类别。这在自动驾驶感知[5]、医疗辅助诊断[6]等高安全需求场景中尤为危险。因此，分布外检测已成为保障人工智能系统安全可信落地的关键技术。

现有研究主要形成两大技术路线[7]：一是基于预训练模型的事后检测方法，通过分析模型的深层特征或输出响应构建评分函数以量化输入异常程度[8-11]，并辅以特征激活校准、稀疏化等网络调整机制提升检测精度[12-14]；二是训练时正则化方法，引入真实外部数据或合成伪分布外样本进行异常暴露[15]，或重构优化目标施加约束[16-17]，引导模型学习稳健的决策边界。

然而，现有分布外检测研究多局限于图像或视频等单一模态[15]。在自动驾驶多传感器环境感知、医疗影像与电子病历联合诊断等场景中，数据往往呈现多源异构特性[18]。单一模态易受复杂环境干扰而陷入感知盲区，而挖掘多模态间的语义关联与互补优势可为异常鉴别提供更全面的判别线索[19]。为此，Dong 等人[20]率先关注到多模态模型特有响应模式：模型对已知分布内（in-distribution, ID）样本的跨模态预测倾向高度一致，对未知样本则易产生预测分歧（见图 1）。其本质在于分布外数据缺乏训练期的语义支撑，各模态在测试时仅能依据自身表征生成独立预测假设。据此，其提出 A2D（agree to disagree, A2D）方法，在训练时扩大分布内样本在非目标类别的跨模态概率差异来诱导预测分歧，进而提升模型对未知输入的判别不确定性。在此基础上，Li 等人[21]指出静态分歧策略易损害类内特征紧凑性，进一步提出 DPU（dynamic prototype updating, DPU）方法，依据样本表征与类别原型的相似度对分歧强度实施动态自适应调节。

尽管上述策略提升了模型对分布外数据的敏感性，但仍存在两大局限。首先，分歧调控缺乏基准约束，易破坏模型自然的不确定性表达（如图 1 右上所示）。当模型遭遇未知目标时，理想的安全响应是各模态均输出趋近均匀分布的低置信度预测；而现有策略以无界的概率分歧最大化为优化目标，其梯度持续推高部分模态在非目标类别上的响应，从而打破该状态并诱发错误高响应。本文实验印证了这一点：引入均匀分

布基准后，模型显著优于仅扩大模态分歧的策略，而在消融实验中放宽该基准时，虚假共识未被充分抑制，检测性能一致下降。其次，模态协同机制匮乏，加剧了联合决策的过度自信(如图 1 右下所示)。究其根源，融合网络在特征拼接后独立完成分类，其决策与各单模态分支相互解耦；面对分布外输入时，即便单模态分支给出合理的低置信度，融合网络仍会因异构噪声的错误叠加或单一异常模态的主导而产生虚假的高置信度。样本级诊断实验清晰呈现了该失效模式：对典型分布外样本，各单模态响应分散且置信度偏低，融合分支却输出显著的置信度尖峰。现有方法的约束仅作用于单模态分支，致使跨模态鉴别信息难以有效聚合，此类过度自信亦无从校正。

针对多模态分布外检测中分歧调控不精准与协同机制缺失挑战，本文提出多模态协同增强(multimodal collaborative enhancement, MCE)方法，构建了可即插即用的层次化协同学习框架。该框架首先在分支级引入基于基准约束的一致性抑制机制，通过锚定均匀分布先验阻断各模态在非目标类别的错误高响应，实现跨模态分歧的细粒度引导；其次，在融合级设计置信度边界约束，以单模态置信度下界严格制约融合网络决策，以此缓解模型在复杂场景下的过度自信现象。借此，MCE 实现了多层级协同信息的深度挖掘，强化了模型对多源异构数据的安全感知。

本文的主要贡献如下：

1) 提出了分支级一致性抑制策略。设计基于基准约束的细粒度分歧诱导机制，抑制模态间错误高响应，实现更精准的跨模态分歧调控；

2) 提出了融合级置信度边界约束。强制单模态置信度制约联合决策，实现全局判别信息的深度挖掘；

3) 系统性的基准评估。在涵盖远/近分布外设定的主流多模态视觉基准上验证了 MCE 的有效性，确证该框架优于现有先进算法。

1 相关工作

分布外检测是保障人工智能系统在开放环境下可靠运行的核心技术。当前研究主要沿事后检测与训练时正则化两类范式展开，并正经历由单模态向多模态感知的关键演进。

事后检测方法无需重训模型，旨在通过构建度量指标量化模型认知不确定性。早期研究多聚焦于输出层，如 Hendrycks 等人[8]和 Liu 等人[9]分别提出基于最大预测概率的 MSP 和基于预测计算能量值的 Energy；Liu 等人[11]则引入广义熵 (GEN) 以提升检测精度。随后，研究者发现基于特征空间的方法能更有效地刻画分布差异，如 Lee 等人[22]和 Sun 等人[10]分别利用马氏距离与 K 近邻(KNN)度量样本偏离已知分布的程度。此外，针对深层网络的激活异常现象，Djurisic 等人[13]与 Sun 等人[14]分别提出 ASH 和 ReAct 等事后网络调整技术，在推理阶段对深层特征实施截断或稀疏化校准，显著提升了分布内与分布外样本的特征可分性。然而，此类方法高度依赖预训练模型固有的特征空间，在面对与已知分布高度混淆的复杂未知样本时，其鉴别能力通常受限。

训练时正则化旨在重塑表征空间，引导模型学习具备更强异常判别能力的决策边界[23-24]。例如，Hendrycks 等人[15]提出异常值暴露策略，引入真实外部数据并施加均匀分布约束以抑制模型对未知样本的预测置信度。而为规避穷举真实未知数据的现实困境，Tao 等人[25]与 Du 等人[26]探索了生成合成伪分布外样本的替代方案。此外，部分研究在无外部数据依赖下重构优化目标，如 Ghosal 等人[16]提出 SNN 通过约束已知样本的低维相关子空间来缓解距离度量的维度灾难；Wei 等人[17]提出 LogitNorm 通过解耦概率范数与交叉熵损失来抑制过度自信。然而，上述方法多局限于单一模态，难以应对复杂场景下多源异构数据的环境感知挑战。

多模态分布外检测方法旨在挖掘跨模态协同互补

信息，以提供更丰富的异常判别线索[20-21]。Dong 等人[20]揭示了模型对已知样本跨模态预测一致、对未知样本易生分歧的响应模式，并据此提出 A2D 方法，通过显式拉大样本在非目标类别的概率差异来诱导分歧。然而，统一强度的分歧策略易破坏类内特征的紧凑性。为此，Li 等人[21]提出 DPU 方法，基于样本与原型的相似度实施分歧强度的动态自适应调控。此外，文献[20]提出的 NP-Mix 策略通过多模态原型插值合成伪分布外样本，进一步扩展了模型对未知区域的感知边界。尽管成效显著，此类方法在开放环境中仍面临两大挑战：一是分歧调控机制失准，单纯概率差异最大化易在非目标类别诱发虚假高响应；二是协同信息整合匮乏，过度侧重单模态分歧而缺乏融合层全局约束，致使联合决策易陷入过度自信。

2 研究方法

针对多模态分布外检测中分歧调控不精准与联合决策过度自信的问题，本文提出多模态协同增强（MCE）层次化框架（如图 2 所示）。在分支级，引入基于基准约束的一致性抑制，锚定均匀分布先验阻断跨模态虚假共识，实现分歧的细粒度引导；在融合级，构建置信度边界约束，以单模态置信度下界制约联合预测，有效规避异构噪声融合诱发的过度自信。

2.1 形式化定义

遵循多模态分布外检测的标准评估范式[20]，本文聚焦于多模态分类任务。设输入空间为$\mathcal{X}$，分布内已知标签空间为$\mathcal{Y}_{in}=\{1,2,\dots,K\}$，训练数据集$\mathcal{D}_{\text{train}}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$采样自分布$P_{\mathcal{X}\mathcal{Y}}$。其中$\boldsymbol{x}_i\in\mathcal{X}$包含 M 个不同模态输入，记作$\boldsymbol{x}_i=\{\boldsymbol{x}_i^m\}_{m=1}^{M}$。在开放环境部署时，测试集不可避免地会引入来自未知类别空间$\mathcal{Y}_{out}$的分布外样本，且满足$\mathcal{Y}_{in}\cap\mathcal{Y}_{out}=\emptyset$。

在多模态分类模型$f:\mathcal{X}\to\mathbb{R}^K$中，单模态编码器$g_m(\cdot)$提取特征$\boldsymbol{Z}^m=g_m(\boldsymbol{x}^m)$。融合分类器$h(\cdot)$接收通道拼接后的联合表征，经过 *softmax* 函数$\delta(\cdot)$归一化处理后，输出多模态联合预测概率向量$\hat{\boldsymbol{p}}\in\mathbb{R}^K$：

$$\begin{aligned}\hat{\boldsymbol{p}}=\delta\big(f(\boldsymbol{x})\big)&=\delta\Big(h\big(g_1(\boldsymbol{x}^1),g_2(\boldsymbol{x}^2),\dots,g_M(\boldsymbol{x}^M)\big)\Big)\\&=\delta\big(h([\mathbf{Z}^1,\mathbf{Z}^2,\dots,\mathbf{Z}^M])\big).\end{aligned}\tag{1}$$

此外，设置单模态辅助分类器$h_m(\cdot)$，对于第m个模态其预测概率$\hat{\boldsymbol{p}}^m$表达为$\hat{\boldsymbol{p}}^m=\delta\Big(h_m\big(g_m(\boldsymbol{x}^m)\big)\Big)$。在闭集假设下，模型监督训练目标为最小化联合预测与各单模态预测的交叉熵$\mathcal{L}_{ce}(\cdot,\cdot)$损失均值：

$$\mathcal{L}_{cls}=\frac{1}{M+1}\Big(\mathcal{L}_{ce}(\hat{\boldsymbol{p}},\boldsymbol{y})+\sum_{m=1}^{M}\mathcal{L}_{ce}\left(\hat{\boldsymbol{p}}^m,\boldsymbol{y}\right)\Big),\tag{2}$$

其中$\boldsymbol{y}\in\{0,1\}^K$为真实标签对应的独热编码向量。

推理阶段，分布外检测被建模为基于水平集估计的二分类问题。构建评分函数$S(\boldsymbol{x};f)$以量化模型对输入样本$\mathbf{x}$的置信度，检测决策函数 $G_\tau(\boldsymbol{x})$定义如下：

$$G_\tau(\boldsymbol{x})=\begin{cases}\text{ID}, & \text{if } S(\boldsymbol{x};f)>\tau;\\ \text{OOD}, & \text{if } S(\boldsymbol{x};f)\le\tau,\end{cases}\tag{3}$$

置信度得分$S(\boldsymbol{x};f)$越低，表明模型不确定性越高，越倾向于将其拒绝为分布外样本。阈值τ通常设定为保证验证集 95%的分布内样本被正确召回的临界值。

2.2 分支级协同：基于基准约束的一致性抑制

为有效引导跨模态预测分歧，需抑制各模态分支在同一非目标类别上的同步高响应。结合 2.1 节定义，给定训练样本$(\boldsymbol{x}_i,y_i)$及单模态分类器输出的原始逻辑值向量。在剔除真实标签y_i对应的预测维度后，对其余 K-1 个类别重新进行 *softmax* 归一化，得到非目标类别预测概率向量$\bar{\boldsymbol{p}}^m\in\mathbb{R}^{K-1}$。当模型对非目标类别不存在明显偏好时，该概率向量趋近均匀分布。因此本文将均匀分布设为基准，从而约束多个模态在同一非目标类别上的共同高响应，减少跨模态虚假共识。

考虑到上述抑制机制涉及基于阈值的截断，若直接采用非光滑函数，如$\max(\cdot,0)$，低于基准的概率分量将产生零梯度，且约束边界处的梯度变化不连续，不利于边界附近样本的稳定优化。因此我们引入 *Softplus* 函数（简记为 sp）作为平滑松弛，使约束强度能够随概率响应偏离基准的程度连续变化。其定义为：

$$\text{sp}(\boldsymbol{x})=\ln(1+e^{x}).\tag{4}$$

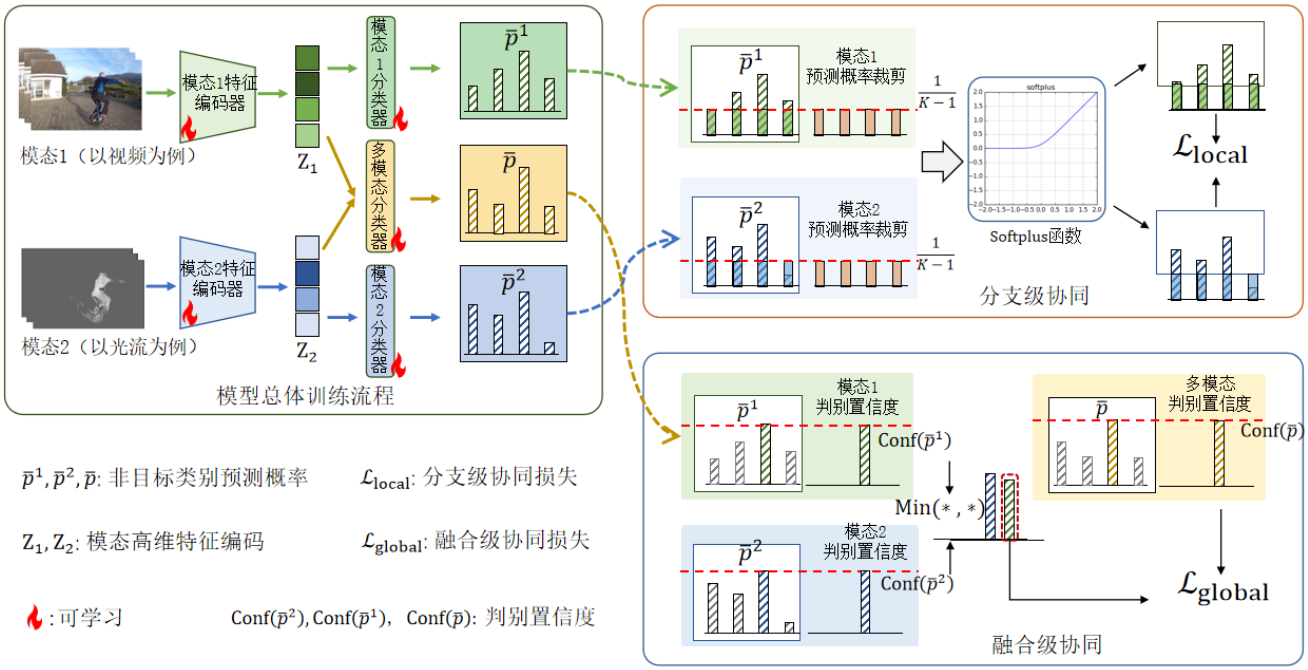


图 2　多模态协同增强驱动的分布外检测框架

Fig.2　The proposed multimodal collaborative enhancement framework for OOD detection

基于该平滑近似，针对 M 个独立模态，定义分支级协同损失$\mathcal{L}_{\text{Local}}$为所有模态两两组合的惩罚项之和：

$$\mathcal{L}_{\text{Local}} = \sum_{1\le u<v\le M}\sum_{k=1}^{K-1} \text{sp}\left(\bar{p}_k^u - \frac{1}{K-1}\right)\cdot \text{sp}\left(\bar{p}_k^v - \frac{1}{K-1}\right). \quad (5)$$

该损失的优化机理在于：仅当多个模态在同一非目标类别的概率均越过均匀分布基准时，惩罚随偏离程度增大；当其中至少一个模态的响应低于基准时，惩罚及梯度平滑减弱，以保留基准附近样本的优化信号。由此，分支级协同约束抑制了异构分支间的偶发虚假共识，实现了对跨模态预测分歧的细粒度调控。

2.3　融合级协同：置信度边界约束

分支级协同侧重于诱导单模态预测分歧，但融合网络仍易因异构噪声错误叠加，在联合决策时陷入过度自信。为深度挖掘跨模态鉴别信息，融合层需具备超越单模态不确定性的全局视野，即融合预测在非目标类别上的置信度应低于任一单模态分支。

为形式化这一目标，本文首先采用最大 *softmax* 概率（MSP）作为置信度度量函数。对于任意给定的非目标类别预测概率向量$\bar{\boldsymbol{p}}$，其置信度得分$\mathcal{C}(\bar{\boldsymbol{p}})$定义为该向量中的最大概率分量：

$$\mathcal{C}(\bar{\boldsymbol{p}}) = \max_{1\le k\le K-1} \bar{p}_k. \quad (6)$$

该得分越低，表示模型对当前输入的不确定性越高。为避免融合网络对单一模态高置信度异常特征的过度依赖，联合预测的置信度被期望严格受制于所有单模态分支的置信度下界。即满足如下条件：

$$\mathcal{C}(\bar{\boldsymbol{p}}) < \min_{1\le m\le M} \mathcal{C}(\bar{\boldsymbol{p}}^m). \quad (7)$$

基于上述准则,我们进一步构建如下融合级协同损失函数:

$$\mathcal{L}_{\text{global}} = sp\left(\mathcal{C}(\bar{\boldsymbol{p}}) - \min_{1\le m\le M} \mathcal{C}(\bar{\boldsymbol{p}}^m)\right). \quad (8)$$

最小化该损失项旨在为融合网络构建动态安全边界，显式引导模型挖掘并聚合跨模态间的互补信息，从而强化融合网络对异常模式的鉴别能力。

2.4　联合训练框架与优化目标

综合上述协同机制，多模态协同增强（MCE）框架在标准已知（ID）闭集数据上的优化目标由分类损失与双层协同约束项构成：

$$\mathcal{L}_{\text{MCE}} = \mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{local}} + \beta\mathcal{L}_{\text{global}}, \quad (9)$$

其中,α与β分别为分支级和融合级的权重参数。MCE 算法流程如算法 1 所示。

算法 1 多模态协同增强驱动的分布外检测算法

Algorithm 1 Flowchart of the multimodal collaborative enhancement-driven out-of-distribution detection algorithm

算法 1 多模态协同增强驱动的分布外检测方法（MCE）

输入：训练集$\mathcal{D}_{\text{train}}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$，模态数 M，类别数 K，单模态编码器 $g_m(\cdot)$，单模态分类器 $h_m(\cdot)$，融合分类器 $h(\cdot)$，损失权重 α、β。

输出：训练后的多模态分类模型 $f(\cdot)$。

1: for each mini-batch $(\boldsymbol{x}_i,y_i)$ do

2: 对每个模态 m=1,...,M，提取单模态特征：

$$\boldsymbol{Z}^m=g_m(\boldsymbol{x}^m)$$

3: 计算融合预测概率与单模态预测概率：

$$\hat{\boldsymbol{p}}=\delta\big(f(\boldsymbol{x})\big)=\delta\big(h([\mathbf{Z}^1,\mathbf{Z}^2,\ldots,\mathbf{Z}^M])\big)$$

$$\hat{\boldsymbol{p}}^m=\delta\left(h_m\big(g_m(\boldsymbol{x}^m)\big)\right)$$

4: 计算闭集分类损失：

$$\mathcal{L}_{cls}=\frac{1}{M+1}\left(\mathcal{L}_{ce}(\hat{\boldsymbol{p}},\boldsymbol{y})+\sum_{m=1}^{M}\mathcal{L}_{ce}\left(\hat{\boldsymbol{p}}^m,\boldsymbol{y}\right)\right)$$

5: 删除真实标签 y 对应维度，并对其余 K-1 个类别重新归一化，得到非目标类别预测概率$\bar{\boldsymbol{p}}^m$。

6: 计算分支级协同损失：

$$\mathrm{sp}(\boldsymbol{x})=\ln(1+e^{x})$$

$$\mathcal{L}_{\text{Local}}=\sum_{1\le u<v\le M}\sum_{k=1}^{K-1}\mathrm{sp}\left(\bar{p}_k^u-\frac{1}{K-1}\right)\cdot\mathrm{sp}\left(\bar{p}_k^v-\frac{1}{K-1}\right)$$

7: 计算融合级协同损失

$$\mathcal{L}_{\text{global}}=sp\left(\mathcal{C}(\bar{\boldsymbol{p}})-\min_{1\le m\le M}\mathcal{C}(\bar{\boldsymbol{p}}^m)\right)$$

8: 得到 MCE 训练目标：

$$\mathcal{L}_{\text{MCE}}=\mathcal{L}_{\text{cls}}+\alpha\mathcal{L}_{\text{local}}+\beta\mathcal{L}_{\text{global}}$$

9: 最小化$\mathcal{L}_{\text{MCE}}$，更新模型参数。

10: end for

为与先前工作保持一致，本文同样引入合成伪分布外样本进行训练时正则化，构建 MCE 的扩展变体框架 MCEN。具体而言，在多模态联合特征空间中提取各已知类别的中心原型 $\boldsymbol{c}$，通过在任意两个异类近邻原型$\boldsymbol{c}_i$和$\boldsymbol{c}_j$间进行线性插值，合成位于类间低密度区域的伪分布外特征：

$$\boldsymbol{Z}_{\text{OOD}}=\lambda\boldsymbol{c}_i+(1-\lambda)\boldsymbol{c}_j. \tag{10}$$

其中，插值比例系数$\lambda\sim Beta(10,10)$。

为促使模型对$\boldsymbol{Z}_{\text{OOD}}$预测概率趋近均匀分布，引入预测熵损失：

$$\mathcal{L}_{\text{ent}}=-\frac{1}{M}\sum_{m=1}^{M}\mathcal{H}\left(\tilde{\boldsymbol{p}}^m\right). \tag{11}$$

其中，$\mathcal{H}$为香农熵。由于$\boldsymbol{Z}_{\text{OOD}}$不属于任何预定义的已知类别，其完整的 K 维预测分布$\tilde{\boldsymbol{p}}\in\mathbb{R}^K$均可视作非目标类别响应，可将前述双层协同增强策略直接推广到伪分布外样本。针对伪分布外样本的分支级与融合级协同损失分别定义为：

$$\mathcal{L}_{\text{local}_{\text{ood}}}=\sum_{1\le u<v\le M}\sum_{k=1}^{K}\mathrm{sp}\left(\hat{p}_k^u-\frac{1}{K}\right)\cdot\mathrm{sp}\left(\hat{p}_k^v-\frac{1}{K}\right), \tag{12}$$

$$\mathcal{L}_{\text{global}_{\text{ood}}}=sp\left(\mathcal{C}(\hat{\boldsymbol{p}})-\min_{1\le m\le M}\mathcal{C}(\hat{\boldsymbol{p}}^m)\right). \tag{13}$$

因此,MCEN 的总体训练目标为:

$$\mathcal{L}_{\text{MCEN}}=\mathcal{L}_{\text{MCE}}+\mathcal{L}_{\text{ent}}+\alpha\mathcal{L}_{\text{local}_{\text{ood}}}+\beta\mathcal{L}_{\text{global}_{\text{ood}}}, \tag{14}$$

式中复用了闭集场景下的协同权重参数α与β。

3 实验与分析

3.1 实验设置

3.1.1 数据集

遵循现有多模态评估范式[20]，本文在 EPIC-Kitchens[27]、HAC[28]、HMDB51[29]、UCF101[30]及 Kinetics-600[31]等 5 个主流视频数据集上开展实验（关键帧见图 3）。上述数据集涵盖第一人称交互至互联网真实动作等多尺度场景，主实验统一提取视频帧与光流构建双模态基准，扩展实验引入音频以验证三模态场景下方法的有效性。为全面验证模型在开放环境下的异常检测能力，构建了以下两类检测基准：

1）远分布外场景：ID 与 OOD 样本源自异构数据集，存在语义与协变量双重偏移。为防止信息泄露并确保类别严格互斥，实验预先剔除了跨数据集的语义重叠类别。具体设置为以 HMDB51 为 ID 集，将剔除 31 个重叠类后的 UCF101（保留 70 类）联同其余数据集作为 OOD 测试集。

图 3 视频数据集关键帧

Fig.3 Sample keyframes of the video datasets

2）近分布外场景：ID 与 OOD 样本划分自同一数据集的互斥类别。由于该场景仅存在语义偏移而无协变量偏移（即背景、画质等数据域一致），极具细粒度判别挑战。具体划分与规模详见表 1。

表 1　近分布外数据集划分与规模

Table 1　Partitioning and statistics of near-OOD datasets

数据集	ID 类别数	OOD 类别数	视频数
EPIC-Kitchens	4	4	4871
HMDB51	25	26	6766
UCF101	50	51	13320
Kinetics-600 子集	129	100	57205

3.1.2　评价指标

本文采用三种广泛用于评估分布外检测性能的指标进行实验:(1)FPR95,在 ID 数据的真正例率 95%时,OOD 数据的假正例率;(2)AUROC,表示接收者操作特征曲线(ROC 曲线)下的面积;(3)ID 准确率(ID ACC),在 ID 测试集上的分类准确率。对于模态退化与对抗实验，进一步固定由干净 ID 验证集 95%真正例率确定的阈值，并报告该阈值下的 OOD 误接受率或攻击成功率(ASR),从而避免针对每种扰动重新调节阈值。

3.1.3　对比基线

实验中选取五类代表性分布外检测算法作为评分函数基准:涵盖概率空间的 MSP[8]、GEN[11]与 Energy[9]、

表 2　多模态远分布外检测基准实验结果(FPR95↓/AUROC↑)

Table 2　Experimental results of multimodal far-OOD detection on benchmark datasets (FPR95↓/AUROC↑)

方法	OOD 数据集				ID 准确率↑
	Kinetics-600	UCF101	EPIC-Kitchens	HAC	
MSP	39.11 / 88.78	46.64 / 86.40	17.33 / 95.99	39.91 / 89.10	87.23
+A2D	37.40 / 88.92	45.95 / 86.28	17.10 / 95.96	33.87 / 90.90	87.00
+MCE	**34.21** / **91.37**	**40.71** / **87.28**	**11.97** / **97.40**	**23.95** / **93.67**	**87.80**
+AN	29.42 / 90.65	40.59 / 88.00	13.45 / 96.43	28.62 / 91.57	86.89
+DPU	27.59 / 93.17	36.72 / 90.52	**12.31** / 97.52	27.25 / 93.30	87.34
+MCEN	**26.91** / **93.93**	**32.38** / **91.36**	12.66 / **97.54**	**23.03** / **94.86**	**87.45**
Energy	32.95 / 92.48	44.93 / 87.95	**08.10** / **97.70**	32.95 / 92.28	87.23
+A2D	31.36 / 92.47	39.57 / 88.13	10.38 / 96.46	25.20 / 93.41	87.00
+MCE	**29.87** / **93.25**	**37.97** / **88.87**	09.81 / 96.73	**20.52** / **94.71**	**87.80**
+AN	24.52 / 93.96	36.49 / 89.67	06.96 / 97.53	22.92 / 94.41	86.89
+DPU	**21.89** / **95.81**	33.30 / 92.76	**04.33** / 98.46	20.64 / 95.39	87.34
+MCEN	22.57 / 95.70	**28.51** / **93.18**	05.13 / **98.86**	**17.10** / **96.44**	**87.45**
ASH	51.20 / 87.81	53.93 / 84.22	19.95 / 95.21	42.99 / 90.23	86.20
+A2D	33.64 / 91.47	**38.31** / 88.05	18.59 / 95.40	26.45 / 93.86	86.66
+MCE	**22.01** / **94.95**	38.43 / **89.60**	**07.87** / **97.46**	**12.66** / **96.11**	**87.91**
+AN	27.82 / 93.17	38.43 / 89.52	06.84 / 98.23	23.03 / 94.45	86.20
+DPU	26.45 / 86.62	37.04 / 89.95	04.33 / 98.46	20.64 / 95.39	**87.34**
+MCEN	**17.45** / **95.81**	**25.88** / **93.07**	**03.76** / **98.68**	**16.19** / **95.92**	86.55
GEN	41.51 / 90.34	46.18 / 87.91	08.21 / 98.26	38.31 / 91.28	87.23
+A2D	33.52 / 92.18	**38.65** / **89.45**	09.81 / 97.21	23.95 / 94.22	87.00
+MCE	**31.70** / **93.37**	41.73 / **89.45**	**04.47** / **98.97**	**21.78** / **95.50**	**87.80**
+AN	25.66 / 93.50	37.40 / 91.19	05.25 / 98.98	24.63 / 94.28	86.89
+DPU	**07.18** / **98.43**	36.37 / 90.93	03.19 / 99.30	21.43 / 95.66	87.34
+MCEN	19.73 / 96.23	**26.80** / **94.15**	**02.62** / **99.44**	**15.51** / **96.94**	**87.45**
KNN	22.69 / 95.01	39.34 / 89.28	09.97 / 97.92	20.75 / 96.02	87.23
+A2D	20.41 / 95.07	**36.15** / 89.48	09.69 / 97.46	16.19 / 96.30	87.00
+MCE	**20.18** / **95.95**	37.29 / **90.24**	**05.93** / **98.69**	**12.77** / **97.50**	**87.80**
+AN	15.05 / 96.96	33.06 / 91.92	**05.47** / **98.97**	**13.45** / 97.25	86.89
+DPU	**03.08** / **99.50**	31.81 / 92.52	05.87 / **98.97**	14.48 / **97.26**	87.34
+MCEN	22.35 / 95.26	**26.34** / **92.53**	12.77 / 97.49	16.76 / 96.37	**87.45**

特征空间的 KNN[10]及激活校准策略 ASH[13]。为验证方法先进性，本文对比了近期主流的多模态检测基线，包括基于预测分歧诱导的 A2D[20]、引入异常样本合成（NP-Mix）的变体 AN[20]以及利用动态原型更新实现分歧强度自适应调控的 DPU[21]。并应用冻结的 LanguageBind-ZS[32]作为多模态基础模型对比基线。为确保评估公平性，所有对比结果均引自公开的 MultiOOD[20]基准，并严格遵循其实验设置。

3.1.4 实现细节

本文实验基于 PyTorch 2.0.1 与 MMAction2 开源

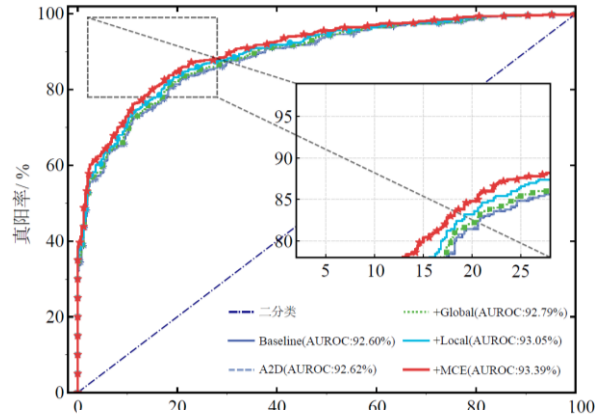


图 4 消融实验 AUROC 曲线

Fig.4 AUROC curves of ablation study

表 3 多模态近分布外检测基准实验结果(FPR95↓/AUROC↑/ID ACC↑)

Table 3 Experimental results of multimodal near-OOD detection on benchmark datasets (FPR95↓/AUROC↑/ID ACC↑)

方法	数据集			
	HMDB51 25/26	UCF101 50/51	EPIC-Kitchens 4/4	Kinetics-600 129/100
MSP	44.66 / 87.74 / 89.32	22.14 / 95.73 / 99.22	76.31 / 67.59 / **71.46**	64.08 / 76.06 / **80.11**
+A2D	38.78 / 88.37 / 90.63	**07.09** / 98.19 / 99.61	66.23 / 67.91 / **71.46**	63.04 / 76.47 / 79.50
+MCE	**36.17 / 89.73 / 90.85**	07.38 / **98.24 / 99.90**	**65.11 / 73.06** / 71.27	**62.89 / 77.63** / 79.91
+AN	33.99 / 88.79 / 89.98	07.96 / 98.24 / 99.71	67.91 / 71.52 / 71.64	62.91 / 76.92 / 80.52
+DPU	34.20 / 89.15 / **92.16**	07.57 / 98.17 / **99.81**	**63.81** / 71.47 / **72.39**	61.59 / 77.50 / **81.07**
+MCEN	**32.03 / 90.36** / 89.76	**07.28 / 98.26 / 99.81**	64.74 / **73.83** / 72.01	**61.53 / 77.03** / 80.56
Energy	43.36 / 87.46 / 89.32	22.52 / 96.06 / 99.22	76.68 / 68.29 / **71.46**	68.75 / 75.49 / **80.11**
+A2D	39.22 / 88.84 / 90.63	09.81 / 98.16 / 99.61	66.98 / 72.45 / **71.46**	64.59 / 76.45 / 79.50
+MCE	**35.51 / 89.32 / 90.85**	**07.28 / 98.39 / 99.90**	**64.18 / 73.51** / 71.27	**62.53 / 77.26** / 79.91
+AN	36.38 / 88.91 / 89.98	06.50 / 98.48 / 99.71	67.91 / 73.79 / 71.64	63.69 / 77.11 / 80.52
+DPU	35.07 / 89.52 / **92.16**	07.48 / 98.43 / **99.81**	63.62 / 74.13 / **72.39**	62.56 / 77.40 / **81.07**
+MCEN	**31.59 / 90.49** / 89.76	**06.12 / 98.49 / 99.81**	**63.43 / 74.84** / 72.01	**60.75 / 77.58** / 80.56
ASH	53.59 / 87.16 / 89.54	32.14 / 94.02 / 99.22	76.87 / 67.92 / **70.15**	69.24 / 76.16 / 79.62
+A2D	42.05 / 87.72 / 90.41	**12.52 / 97.43** / 99.42	**63.25 / 74.73** / 67.72	64.28 / 77.28 / 79.15
+MCE	**37.25 / 89.67 / 91.29**	14.37 / 97.15 / **99.90**	64.74 / 73.60 / 68.84	**62.24 / 77.43 / 79.82**
+AN	36.17 / 89.30 / 89.32	10.68 / 97.80 / **99.81**	66.23 / 73.06 / 66.23	63.77 / 77.44 / 79.44
+DPU	**35.08** / 89.51 / **92.16**	**07.48 / 98.43 / 99.81**	71.46 / 71.24 / 69.78	62.57 / 77.41 / **81.07**
+MCEN	36.82 / **89.55** / 91.07	10.49 / 98.00 / **99.81**	**59.33 / 75.59 / 72.01**	**60.57 / 77.59** / 80.56
GEN	43.79 / 87.49 / 89.32	23.79 / 95.54 / 99.22	76.87 / 68.52 / **71.46**	69.03 / 75.33 / **80.11**
+A2D	38.56 / 88.61 / 90.63	08.83 / 98.12 / 99.61	65.86 / 73.05 / **71.46**	62.28 / 77.08 / 79.50
+MCE	**36.82 / 89.69 / 90.85**	**07.67 / 98.21 / 99.90**	**65.67 / 74.39** / 71.27	**61.91 / 77.40** / 79.91
+AN	35.95 / 89.78 / 89.32	07.67 / 98.28 / 99.71	65.30 / 75.17 / 71.64	62.95 / 76.95 / 80.52
+DPU	34.64 / 89.71 / **92.16**	**06.41** / 98.41 / **99.81**	**63.06** / 74.22 / **72.39**	**62.10 / 78.05 / 81.07**
+MCEN	**31.81 / 90.78** / 89.76	08.16 / **98.65 / 99.81**	63.99 / **75.56** / 72.01	62.20 / 77.58 / 80.56
KNN	42.92 / 88.46 / 89.32	15.63 / 96.93 / 99.22	75.93 / 63.60 / **71.46**	68.67 / 74.64 / **80.11**
+A2D	**33.33** / 89.59 / 90.63	07.48 / 98.39 / 99.61	72.39 / 67.83 / **71.46**	65.71 / 76.23 / 79.50
+MCE	34.42 / **89.98 / 90.85**	**06.21 / 98.54 / 99.90**	**68.66 / 70.57** / 71.27	**63.18 / 77.14** / 79.91
+AN	33.77 / 90.05 / 89.98	12.04 / 97.65 / 99.71	71.27 / 69.94 / 71.64	66.81 / 74.19 / 80.52
+DPU	32.90 / 89.36 / **92.16**	11.84 / 97.74 / **99.81**	71.64 / 68.03 / **72.39**	66.71 / 75.34 / **81.07**
+MCEN	**31.81 / 91.30** / 89.76	**06.80 / 98.65 / 99.81**	**68.10 / 72.13** / 72.01	**64.32 / 76.16** / 80.56

算法库构建，硬件加速平台搭载 NVIDIA RTX 4090 GPU 与 Intel Xeon Gold 5220R CPU。在多模态网络结构设计上，视觉分支采用标准的 SlowFast 三维卷积架构作为特征编码器，该网络由以低帧率捕获高层空间语义的 Slow 路径和以高帧率提取瞬态运动特征的 Fast 路径构成，并通过侧向连接实现双分辨率特征交互；光流分支则针对密集时序运动特性，精简为单一的 Slow-only 架构。输入视频序列经统一的帧采样与空间裁剪后送入网络，提取的各模态高维特征经全局平均池化后沿通道维度拼接。所有编码器均使用 Kinetics-400 预训练权重初始化。训练时采用 Adam 优化器，初始学习率设为 0.0001，训练期间保持固定学习率，未额外采用学习率衰减。同时权重衰减设为 0.0001，批大小设为 16，分支级抑制系数α与融合级约束系数β均设为 0.5。训练轮次为 50，按验证集性能保存最优检查点。所有新增训练与复测均用随机种子 0。

3.2 主实验结果与分析

3.2.1 多模态远分布外检测评估

表 2 报告了本文方法与现有基准在远分布外场景下的对比实验结果。数据表明，MCE 及其扩展框架 MCEN 在融合不同的评分函数时，均取得了显著的性能提升。具体而言：(1) 常规检测性能：在标准训练范式下，MCE 一致优于同类多模态基线。以 HAC 数据集为例，相较于基础 MSP，MCE 将其 FPR95 大幅抑制了 15.96%，AUROC 提升 4.57%，有力验证了分支与融合双层协同机制对多模态判别表征的有效强化；(2) 异常暴露正则化增益：引入合成分布外样本进行训练时正则化（阴影区域）后，MCEN 进一步规范了开放特征空间的判别边界，在 Kinetics-600 和 UCF101 等复杂评测集上展现出极具竞争力的表现；(3) 异常鉴别与闭集分类的稳健权衡：在提升开放环境检测能力的同时，MCE 与 MCEN 均严格维持甚至略微提升了分布内数据的分类精度。

3.2.2 多模态近分布外检测评估

表 3 呈现了近分布外场景的评估结果。面对与分布内语义高度重叠的困难样本，所提框架依然展现出显著的鉴别优势。具体而言：(1) 困难样本细粒度判别：在基线极易误判的近分布外设定下，MCE 取得稳健增益。以 HMDB51 为例，结合 MSP 基线评分，MCE 将 FPR95 下降 8.49%，AUROC 提升 1.99%。这表明双层协同约束有效抑制了模型对未知样本的过度自信；(2) 模糊边界精细刻画：引入异常暴露策略后，MCEN 进一步强化了模型对边缘样本的拒识能力（如 HMDB51 上的 FPR95 进一步降至 32.03%），证实伪样本合成与约束机制能有效锐化近分布的模糊决策边界；(3) 闭集分类能力：在更具挑战性的近分布外设定下，MCE 与 MCEN 不仅未牺牲闭集泛化能力，反而使 ID 准确率有所提升（例如 HMDB51 提升至 90.85%）。这确证了该框架在深度挖掘多模态异常信息的同时保持了分布内样本的分类能力。

3.2.3 结果波动性分析

表 2 和表 3 显示，所提方法在不同数据集与评分函数上的增益幅度虽存在差异，但整体提升趋势稳定。MCE 在近分布外的 20 组实验中均同时降低 FPR95 并提高 AUROC，在远分布外的 20 组实验中有 19 组实现一致改进，表明其性能优势不依赖于特定数据集或评分函数。局部波动主要与数据集的分布偏移程度、原始评分函数的性能上限以及特征调整机制的适配性有关：当基线性能已接近饱和时，进一步提升空间有限；对于依赖局部邻域结构的 KNN，异常暴露引起的特征分布变化也可能削弱其局部度量优势。此外，MCEN 的收益还受到合成异常样本与真实未知分布匹配程度的影响，因此并非在所有设置下均单调优于 MCE。总体而言，MCE 提供了稳定的跨场景增益，MCEN 则在多数实验中进一步提升检测性能，同时保持相近的分布内分类精度，验证了所提框架在不同分布偏移和评分机制下的有效性与稳定性。

表 4 多模态协同策略消融分析（实验结果基于 Far-OOD 基准上 4 个 OOD 测试数据集的平均值）

Table 4 Ablation analysis of multimodal collaborative strategies (experimental results averaged across four OOD test datasets on far-OOD benchmarks)

方法	Local	Global	FPR95↓	AUROC↑	ID ACC↑
Energy	×	×	29.73	92.60	87.23
+A2D	×	×	26.63	92.62	87.00
+Local	√	×	26.31	93.05	87.39
+Global	×	√	27.36	92.79	87.34
+MCE	√	√	**24.54**	**93.39**	**87.80**

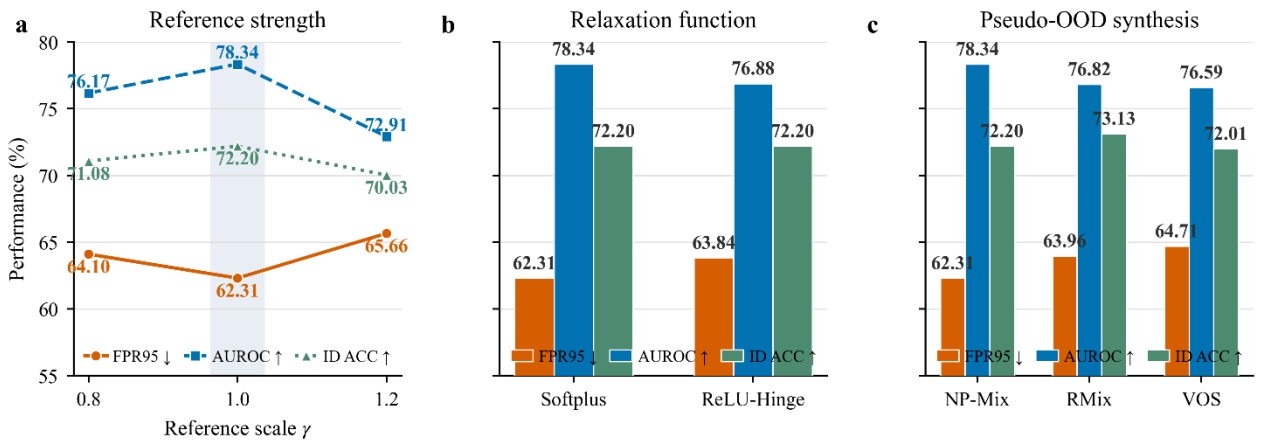


图 5 组件设计消融结果

Fig.5 Ablation results for the component designs

3.3 方法有效性与鲁棒性分析

3.3.1 协同策略消融分析

表 4 和图 4 基于 Energy 评分报告了远分布外场景下的消融结果，确证了各协同模块的有效性与互补增益：(1) 分支级约束 (+Local)：在四个测试集上均优于 A2D（无约束扩大模态分歧的策略）的同时较基线显著抑制了 FPR95 (-3.42%)，表明基于基准约束的一致性抑制能精准阻断跨模态虚假共识，从而提升模型分布外检测能力；(2) 融合级约束 (+Global)：独立应用亦获稳健增益，证实置信度边界约束可有效规避融合决策时的过度自信风险；(3) 双层协同 (+MCE)：联合部署后性能达全局最优（FPR95 降至 24.54%，AUROC 达 93.39%），验证了分支级约束与融合级约束的高度互补性，实现了多模态鉴别能力的有效强化。

3.3.2 组件设计消融分析

为进一步考察关键设计选择的必要性，本节在 EPIC-Kitchens 三模态近分布外基准上，围绕分支级基准强度、约束的平滑松弛形式与伪分布外样本合成策略开展消融实验。结果如图 5 所示。

(1) 基准强度。将分支级约束的均匀分布基准乘以缩放系数γ，$\gamma\in\{0.8,1.0,1.2\}$。$\gamma=1$时各项指标均为最优，向两侧偏离均导致性能一致下降，其中$\gamma=1.2$，FPR95 提升 3.35 个百分点，AUROC 降低 5.43 个百分点。偏小的γ会扩大约束范围，可能抑制正常的非目标响应；偏大的γ则减弱对跨模态一致高响应的约束，虚假共识未被充分抑制，检测与闭集分类性能同步受损。这表明过强或过弱的基准约束均会影响检测与分类性能，而均匀先验为分歧引导提供了恰当的参考基准。

(2) Softplus 松弛。将 Softplus 替换为 ReLU-Hinge 后，FPR95 上升 1.53 个百分点，AUROC 下降 1.46 个百分点，ID ACC 保持不变。ReLU-Hinge 在约束边界处不可导，且在约束满足侧梯度恒为零，边界附近样本难以获得有效的优化信号；Softplus 提供连续可导的近似，使约束项梯度平滑过渡。二者仅在检测指标

表 5 EPIC-Kitchens 三模态近分布外检测结果及基础模型参考

Table 5 Tri-modal near-OOD detection results on EPIC-Kitchens with a foundation-model reference

方法	训练协议	FPR95↓	AUROC↑	ID ACC↑
LanguageBind-ZS	零样本	96.74	45.59	26.12
Baseline	监督训练	69.22	72.39	73.13
A2D	监督训练	67.54	70.80	72.39
DPU	监督训练	68.47	72.71	71.46
MCEN	监督训练	**62.31**	**78.34**	**72.20**

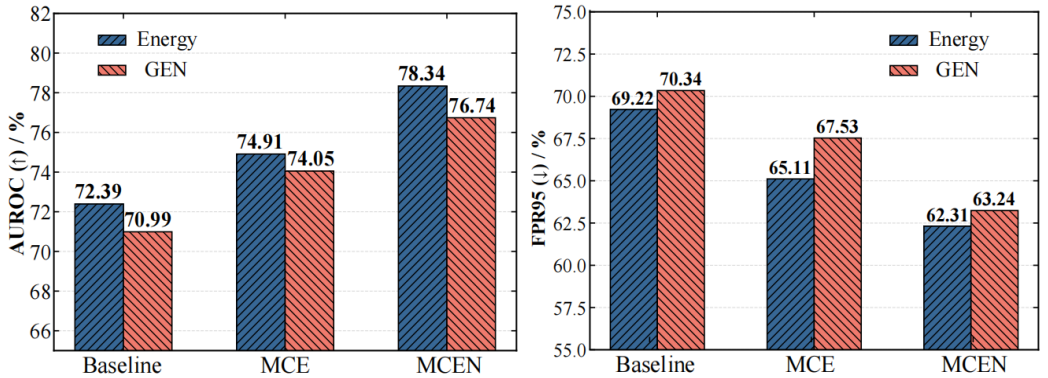


图 6 EPIC-Kitchens 三模态近分布外检测性能

Fig.6 Tri-modal Near-OOD detection performance on EPIC-Kitchens

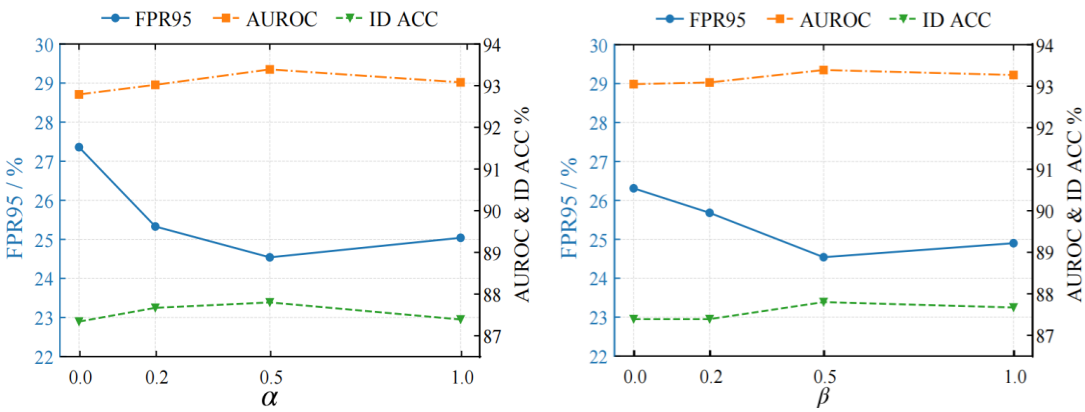


图 7 协同损失权重的敏感性分析

Fig.7 Sensitivity analysis of the collaborative loss weights

上产生差异，表明平滑松弛的作用集中于分布外判别边界的学习，且不以牺牲闭集分类性能为代价。

(3) 伪分布外合成策略。本文分别采用随机跨类特征插值 RMix 和类条件高斯低密度采样 VOS 替换 NP-Mix。NP-Mix 取得最佳 FPR95 和 AUROC，表明邻近异类原型插值更适合刻画近分布外类别之间的局部过渡区域。相比之下，RMix 覆盖更广泛的跨类组合，VOS 侧重类条件分布的低密度区域，而 NP-Mix 能够更有针对性地在语义邻近类别之间构造边界样本，从而强化模型对近边界未知模式的识别能力。

3.3.3 三模态检测评估

表 5 与图 6 报告了引入音频后的三模态评估结果：(1) 在干净测试条件下，MCEN 较基线在 Energy 评分下实现 FPR95 下降 6.91%、AUROC 提升 5.95%，且性能由 Baseline 到 MCE 再到 MCEN 依次提升，表明层次化协同约束可自然扩展至更多模态；(2) 不同

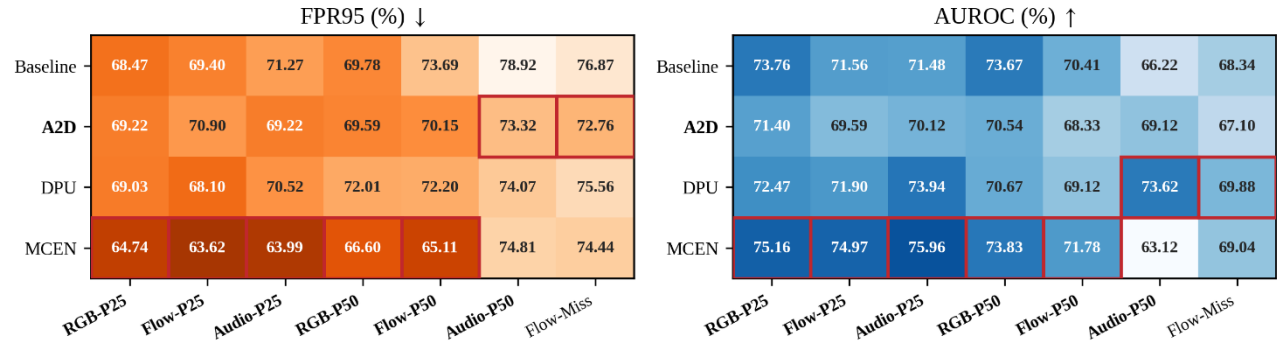


图 8 模态损坏与缺失条件下的三模态近分布外检测结果

Fig.8 Tri-modal Near-OOD detection under modality corruption and missingness

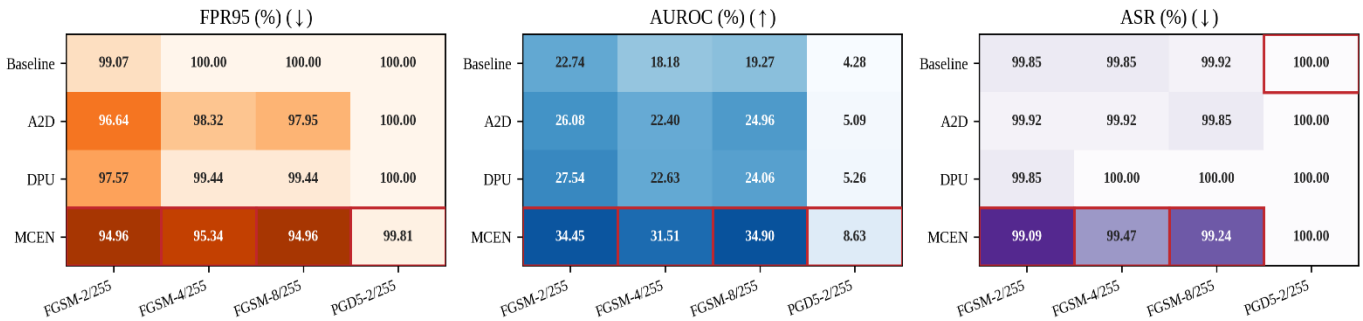


图 9 RGB 单模态对抗扰动下的分布外检测结果

Fig.9 OOD detection under RGB-only adversarial perturbations

评分函数下均获得一致增益，说明该机制与评分函数解耦，兼容性和稳定性良好；(3) 为提供跨范式参照，表 5 同时列出多模态基础模型 LanguageBind 的零样本检测结果。尽管受益于大规模跨模态预训练，检测性能仍显著落后于任务特定训练的 MCEN（FPR95 差距达 34.43%），表明通用对齐表征尚不能替代面向已知类别结构的判别约束，二者的结合是值得探索的方向。

3.3.4 参数敏感性分析

图 7 展示了分支级权重α与融合级权重β的敏感性分析实验结果。可以看到，两类权重的变化趋势整体一致：随着权重从 0 增加到 0.5，FPR95 持续下降、AUROC 与 ID ACC 稳步提升；当权重继续增大到 1.0 时，检测性能略有回落。结果表明，适中的协同约束强度能够在增强分布外判别能力的同时较好保持分布内分类性能，其中α=β=0.5 时综合表现最佳。这说明所提出的层次化协同机制对超参数变化具有较好的稳定性与鲁棒性。

3.3.5 真实环境鲁棒性：模态损坏与缺失

实际部署中传感器噪声、压缩失真与设备故障会造成模态质量下降乃至完全缺失。为模拟传感器短时丢帧与完全失效，本文在归一化后对单一模态施加连续时间遮蔽，其余模态保持不变。Partial-25 和 Partial-50 分别将 25%和 50%的连续时间片段置零，Missing 将整个模态置为零。所有模型均直接使用既有检查点，不引入掩码标记、重构模块或针对性微调。

由图 8 可以观察到：(1) Partial-25 轻度损坏下，MCEN 在三种模态上均保持明显优势，FPR95 平均为 64.12%，较次优方法 DPU 低 5.10 个百分点，AUROC 亦全面领先。这表明当受损模态仍保留部分判别信息时，融合级边界约束能有效抑制模态噪声向联合决策的传播；(2) 损坏比例升至 50%后，MCEN 在视觉与光流损坏下仍取得最低的 FPR95(分别为 66.60%与 65.11%)，但在音频重度损坏下优势不再。其原因在于，边界约

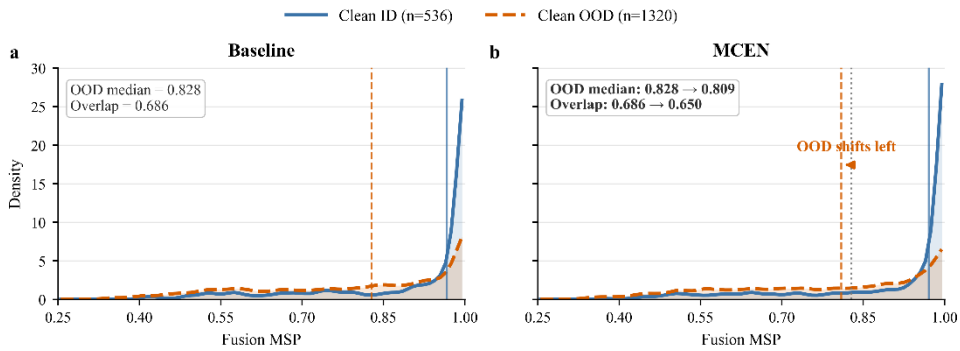


图 10　Baseline 与 MCEN 的融合置信度分布

Fig.10　Fusion-confidence distributions of Baseline and MCEN

束以单模态置信度为参考，当某一模态的置信度因严重损坏而失真时，该约束的校准作用随之减弱；（3）在 Audio-P50 与 Flow-Miss 等极端条件下，各方法性能普遍大幅退化且差距缩小。这说明现有基于置信度的多模态检测范式均隐含模态信息完整的前提，模态缺失自适应检测仍是有待解决的问题。

3.3.6　对抗扰动下的检测实验

本文进一步在 OOD 测试样本的 RGB 帧上实施白盒置信度规避攻击。攻击以推动 OOD 样本的 Energy 评分向 ID 样本评分区域偏移为目标，并报告 FPR95、AUROC 及干净阈值 τ 下的 ASR。FGSM 采用 ε=2/255、4/255、8/255；PGD 采用 5 步迭代、ε=2/255。

图 9 表明：（1）对抗扰动对所有方法均造成严重退化：FGSM 攻击下各方法 FPR95 普遍升至 95%以上，AUROC 降至 35%以下；PGD 攻击下 AUROC 进一步跌至 10%以下。值得注意的是，AUROC 远低于 50%的随机水平，意味着对抗样本的评分被系统性抬高至 ID 样本之上，说明基于置信度的评分函数本身即构成攻击面，现有多模态检测范式对该类威胁缺乏内在防御；（2）相对而言，MCEN 在 FGSM 各扰动强度下保持最低的 FPR95 与最高的 AUROC（领先次优方法 7 个百分点以上），可能得益于均匀基准约束压低了非目标类别 logit 的幅值，缩小了梯度攻击可利用的置信度上升空间。但该优势随攻击增强而收窄，PGD 下与其余方法趋同，不足以支撑对抗条件下的可靠部署；（3）需要说明的是，本文并未针对对抗鲁棒性进行专门设计，上述实验旨在客观界定方法的适用边界。

3.3.7　融合边界诊断与典型样本分析

为进一步理解融合级约束的作用机理与失效模式，本节从两个角度进行实验。

（1）低置信度子集与置信度分布分析。针对融合级约束在极端条件下的有效性，本节从两个互补角度进行诊断。其一，在各方法的 OOD 样本中选取最大单模态置信度位于最低四分位的 330 个样本构成困难子集——该子集对应所有单模态分支均高度不确定、无任何模态提供可靠判别的情形。如表 6 所示，MCEN 将该子集的误接受率降至 67.58%，较次优方法 DPU 降低 12.72 个百分点，表明 MCEN 的融合级约束在单模态线索普遍失效时依然存在性能提升，增益正集中于此类模糊样本。其二，图 10 给出全体测试样本的融合置信度分布。相较基线，MCEN 使 OOD 分布向低置信度方向移动：中位数由 0.828 降至 0.809，高置信度密度峰收缩，ID–OOD 重叠面积由 0.686 降至 0.650。两项结果共同表明，该约束选择性地压制了 OOD 样本的错误高响应，分布层面未出现其置信度被推高的迹象。

表 6　低单模态置信度 OOD 子集误接受率统计

Table 6　Statistics on the false acceptance rate of OOD subsets with the low-unimodal-confidence

方法	样本数	误接受率↓
Baseline	330	85.45
A2D	330	80.61
DPU	330	80.30
MCEN	330	**67.58**

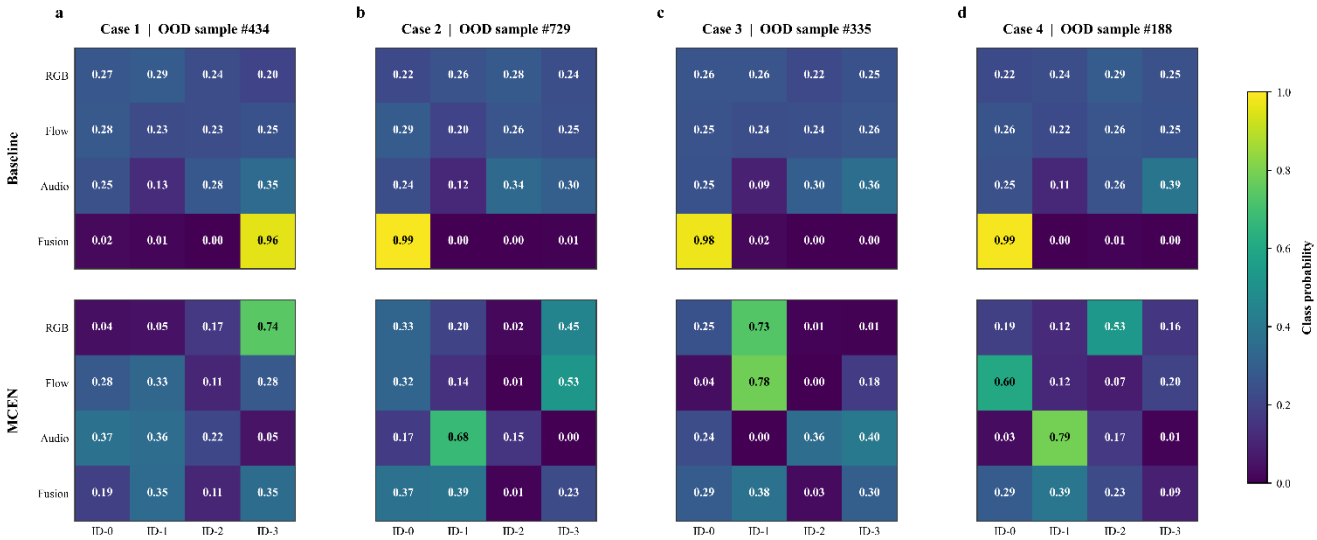


图 11 代表性 OOD 样本的跨模态类别概率热力图

Fig.11 Cross-modal class-probability heatmaps for representative OOD samples

(2)典型样本分析。图 11 和图 12 展示了 Baseline 误接受而 MCEN 正确拒绝的代表性 OOD 样本。Baseline 的单模态类别响应较为分散，最大概率均不超过 0.35，但融合分支却产生超过 0.96 的置信度尖峰，表明融合过程放大了单模态中缺乏一致支持的类别证据。MCEN 通过融合级约束抑制此类过度自信，同时保留不同模态间的预测分歧。四个样本的融合置信度均降至 0.40 以下，Energy 评分降至 0.9 以下，并由误接受转为正确拒绝。上述样本级证据表明融合级约束的校正可以作用在缺乏单模态支撑的融合过度自信这类失效模式上。

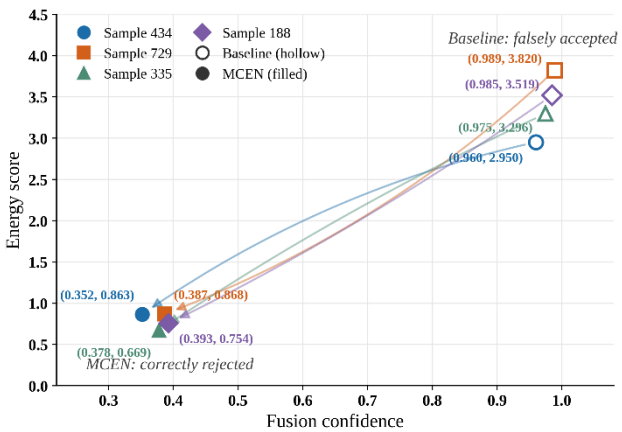


图 12 融合置信度和 Energy 评分对比图

Fig.12 Comparison of fused confidence and energy scores

4 结束语

针对多模态分布外检测中分歧调控不精准及模态协同信息挖掘不足的挑战，本文提出多模态协同增强方法 (MCE)。该方法通过分支级一致性抑制与融合级置信度边界约束，对跨模态分歧和融合判别进行联合建模，从而提升模型对未知样本的识别能力。多基准实验表明，MCE 及其扩展框架 MCEN 全面优于现有先进算法，确证了其在保障系统感知可靠的优越性。未来研究可从两个方向展开：其一，面向大规模视觉语言基础模型与具身智能系统，研究可迁移的多模态分布外评分函数，提升模型在开放场景中的异常感知能力；其二，面向自动驾驶、医疗辅助诊断等场景，进一步考虑模态缺失、传感器噪声、时序漂移等问题，构建应用于实际部署环境的鲁棒检测框架。

参考文献：

[1] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//Proceedings of the International Conference on Learning Representations. OpenReview.net, 2021.

[2] 陈德光, 马金林, 马自萍, 等. 自然语言处理预训练技术综述[J]. 计算机科学与探索, 2021, 15(8): 1359-1389. CHEN D G, MA J L, MA Z P, et al. Review of pre-training techniques for natural language processing[J]. Journal of

Frontiers of Computer Science and Technology, 2021, 15(8): 1359-1389.

[3] 桂韬, 奚志恒, 郑锐, 等. 基于深度学习的自然语言处理鲁棒性研究综述[J]. 计算机学报, 2024, 47(1): 90-112.
GUI T, XI Z H, ZHENG R, et al. Recent researches of robustness in natural language processing based on deep neural network[J]. Chinese Journal of Computers, 2024, 47(1): 90-112.

[4] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述[J]. 软件学报, 2021, 32(2): 327-348.
DU P F, LI X Y, GAO Y L. Survey on multimodal visual language representation learning[J]. Journal of Software, 2021, 32(2): 327-348.

[5] 张燕咏, 张莎, 张昱, 等. 基于多模态融合的自动驾驶感知及计算[J]. 计算机研究与发展, 2020, 57(9): 1781-1799.
ZHANG Y Y, ZHANG S, ZHANG Y, et al. Multi-modality fusion perception and computing in autonomous driving[J]. Journal of Computer Research and Development, 2020, 57(9): 1781-1799.

[6] 田娟秀, 刘国才, 谷珊珊, 等. 医学图像分析深度学习方法研究与挑战[J]. 自动化学报, 2018, 44(3): 401-424.
TIAN J X, LIU G C, GU S S, et al. Deep learning in medical image analysis and its challenges[J]. Acta Automatica Sinica, 2018, 44(3): 401-424.

[7] YANG J K, WANG P Y, ZOU D J, et al. OpenOOD: benchmarking generalized out-of-distribution detection[C]//Proceedings of the 36th Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2022: 32598-32611.

[8] HENDRYCKS D, GIMPEL K. A baseline for detecting misclassified and out-of-distribution examples in neural networks[C]//Proceedings of the International Conference on Learning Representations. OpenReview.net, 2017.

[9] LIU W, WANG X, OWENS J D, et al. Energy-based out-of-distribution detection[C]//Proceedings of the 34th Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2020: 1069-1078.

[10] SUN Y Y, MING Y, ZHU X Y, et al. Out-of-distribution detection with deep nearest neighbors[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022: 20827-20840.

[11] LIU X X, LOCHMAN Y, ZACH C. GEN: pushing the limits of softmax-based out-of-distribution detection[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 23946-23955.

[12] XU H N, YANG Y. ITP: instance-aware test pruning for out-of-distribution detection[C]//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2025: 21743-21751.

[13] DJURISIC A, BOZANIC N, ASHOK A, et al. Extremely simple activation shaping for out-of-distribution detection[C]//Proceedings of the International Conference on Learning Representations. OpenReview.net, 2023.

[14] SUN Y Y, GUO C, LI Y X. ReAct: out-of-distribution detection with rectified activations[C]//Proceedings of the 35th Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2021: 144-157.

[15] HENDRYCKS D, MAZEIKA M, DIETTERICH T. Deep anomaly detection with outlier exposure[C]//Proceedings of the International Conference on Learning Representations. OpenReview.net, 2019.

[16] GHOSAL S, SUN Y Y, LI Y X. How to overcome curse-of-dimensionality for out-of-distribution detection?[C]//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2024: 19849-19857.

[17] WEI H X, XIE R C, CHENG H, et al. Mitigating neural network overconfidence with logit normalization[C/OL]//Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022: 23631-23644 [2026-08-12]. https://arxiv.org/abs/2205.09310.

[18] JIAO Y, JIE Z, CHEN S X, et al. MSMDFusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 21643-21652.

[19] WANG Y, PENG J L, ZHANG J N, et al. Multimodal industrial anomaly detection via hybrid fusion[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 8032-8041.

[20] DONG H, ZHAO Y, CHATZI E, et al. MultiOOD: scaling out-of-distribution detection for multiple modalities[C]//Proceedings of the 38th Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2024: 129250-129278.

[21] LI L, GONG H X, DONG H, et al. DPU: dynamic prototype updating for multimodal out-of-distribution detection[C]//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 10193-10202.

[22] LEE K, LEE K, LEE H, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2018: 7167-7177.

[23] ZHOU Z, GUO L Z, CHENG Z Z, et al. STEP: out-of-distribution detection in the presence of limited in-distribution labeled data[C]//Proceedings of the 35th Conference on Neural

Information Processing Systems. Red Hook: Curran Associates, 2021: 29168-29180.
[24] ZHANG Z H, XIANG X. Decoupling MaxLogit for out-of-distribution detection[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 3388-3397.
[25] TAO L T, DU X F, ZHU J, et al. Non-parametric outlier synthesis[C]//Proceedings of the International Conference on Learning Representations. OpenReview.net, 2023.
[26] DU X F, WANG Z N, CAI M, et al. VOS: learning what you don't know by virtual outlier synthesis[C]//Proceedings of the International Conference on Learning Representations. OpenReview.net, 2022.
[27] MUNRO J, DAMEN D. Multi-modal domain adaptation for fine-grained action recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 119-129.
[28] DONG H, NEJJAR I, SUN H, et al. SimMMDG: a simple and effective framework for multi-modal domain generalization[C]//Proceedings of the 37th Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2023: 78674-78695.
[29] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society, 2011: 2556-2563.
[30] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[EB/OL]. [2026-08-12]. https://arxiv.org/abs/1212.0402.
[31] CARREIRA J, NOLAND E, BANKI-HORVATH A, et al. A short note about Kinetics-600[EB/OL]. [2026-08-12]. https://arxiv.org/abs/1808.01340.
[32] ZHU B, LIN B, NING M N, et al. LanguageBind: extending video-language pretraining to N-modality by language-based semantic alignment[C]//Proceedings of the International Conference on Learning Representations. OpenReview.net, 2024.

冯鸣旭（2001—），女，河北定州人，博士研究生，研究方向为分布外检测、多模态学习等。

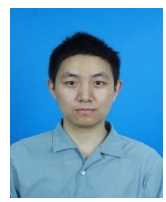
FENG Mingxu, born in 2001, Ph.D. candidate. Her research interests include out-of-distribution detection, multimodal learning, etc.

徐浩楠（2001—），男，浙江温州人，硕士研究生，研究方向为分布外检测。

XU Haonan, born in 2001, M.S. candidate. His research interests include out-of-distribution detection.

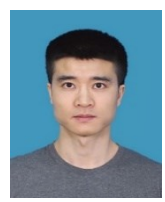
张宇萱（1998—），男，江苏南通人，博士研究生，研究方向为分布外检测、多模态大模型等。

ZHANG Yuxuan, born in 1998, Ph.D. candidate. His research interests include out-of-distribution detection, multimodal large language models, etc.



田奇（1994—），男，四川成都人，博士，高级工程师，研究方向为信号处理。

TIAN Qi, born in 1994, Ph.D., senior engineer. His research interests include signal processing.



杨杨（1991—），男，安徽宁国人，博士，教授，研究方向为多模态学习、分布外检测等。

YANG Yang, born in 1991, Ph.D., Professor. His research interests include multimodal learning, out-of-distribution detection, etc.