计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(9): 2166−2182, 2025

置信引导的多模态数据测试时适应框架

尹小龙　詹德川　姜　远
（南京大学人工智能学院　南京　210023）
（计算机软件新技术全国重点实验室（南京大学）　南京　210023）
（yinxl@lamda.nju.edu.cn）

Confidence-Guided Multimodal Data Test-Time Adaptation Framework

Yin Xiaolong, Zhan Dechuan, and Jiang Yuan
（*School of Artificial Intelligence*, *Nanjing University*, *Nanjing* 210023）
（*National Key Laboratory for Novel Software Technology* (*Nanjing University*), *Nanjing* 210023）

Abstract　Test-time adaptation (TTA) seeks to dynamically recalibrate a deployed model through online fine-tuning with unlabeled or sparsely labeled data, addressing performance degradation caused by distribution shift, sensor noise, illumination changes, and other real-world challenges. It has demonstrated great promise in latency- and robustness-critical applications such as autonomous driving, remote healthcare, and video surveillance. However, existing multimodal TTA methods typically overlook the quality variance across views—low-quality or faulty views data can introduce harmful gradients—and lack mechanisms to preserve internal temporal consistency, undermining stability in dynamic scenarios. To tackle these issues, we propose CVPTA (confidence-guided view pruning and temporal contrastive attention), a unified framework comprising three core modules: 1) confidence-guided dynamic attention, which computes each view's confidence via its predictive entropy and uses a softmax over inverted entropy scores to attenuate high-entropy views data during feature fusion; 2) view pruning, which evaluates image-modal quality using the Laplacian variance (for blur) and histogram skewness (for exposure), discards views below a quality threshold, and maintains a constant view count by interpolating adjacent high-quality views data or injecting Gaussian noise—dramatically reducing noise accumulation; 3) temporal contrastive self-supervision, which treats adjacent frames of the same view as positive pairs and non-adjacent frames as negatives, and applies a contrastive loss to enforce temporal feature consistency. We evaluate EVPTA on the public Kinetics-50-C and VGGSound-C perturbation benchmarks under the same online-update settings as prior work. Results show that EVPTA boosts Top-1 accuracy by approximately 2.3% and 0.7% on each benchmark and retains over 0.2% gains even under extreme noise conditions. Ablation studies further confirm the individual and synergistic contributions of all three modules. EVPTA requires no extra annotations, integrates seamlessly with existing multimodal systems, and delivers both efficiency and robustness—offering significant theoretical insights and practical value.

Key words　test-time adaptation; multimodal learning; multi-view learning; view pruning; temporal contrastive learning

摘　要　测试时适应（test time adaptation, TTA）旨在在模型部署后的测试阶段，通过无标注或少量标注的

收稿日期：2025-06-06；**修回日期**：2025-07-16
基金项目：国家自然科学基金项目（62176117）；国家科技重大专项（2024ZD0531802）
This work was supported by the National Natural Science Foundation of China (62176117) and the National Science and Technology Major Project of the Ministry of Science and Technology of China (2024ZD0531802).
通信作者：詹德川（zhandc@nju.edu.cn）

在线微调策略来动态校正模型，以应对由于分布偏移、传感器噪声、光照变化等引发的性能退化问题，并在自动驾驶、远程医疗、视频监控等对实时性和鲁棒性要求极高的领域展现出广泛应用价值. 然而，现有多模态 TTA 方法往往忽视各视图数据的质量差异，容易因低质量或故障视图引入有害梯度，且缺乏对视图内部动态变化的刻画，难以保持时序特征的一致性和稳定性. 为解决上述挑战，提出了一种置信度引导的视图裁剪与时序对比注意力（confidence-guided view pruning and temporal contrastive attention, CVPTA）框架. 该框架包含三大模块：首先，基于模型预测分布计算视图不确定性，自适应削弱低置信度视图数据对特征融合的干扰；其次，视图裁剪策略利用 Laplacian 方差与亮度直方图偏度评估图像质量，丢弃质量分数低于阈值的视图数据，并结合相邻高质量视图重采样，显著降低噪声累积风险；最后，时序对比自监督任务将同一样本的不同视图视为正样本而非同一样本视图视为负样本，通过对比损失强化时序特征一致性. 在 Kinetics-50-C 与 VGGSound-C 两个公开多模态扰动基准上，沿用先前实验配置进行在线更新评测. 结果显示，CVPTA 在 2 组基准上分别将 Top-1 准确率提升约 2.3 个百分点和 0.7 个百分点，在极端噪声场景中依然保持超过 0.2% 的性能增益；消融研究进一步验证了各模块的独立贡献与协同效应. 该方法无需额外标注，可无缝集成现有多模态系统，兼具高效性与鲁棒性，具有重要的理论意义与工程应用前景.

关键词 测试时适应；多模态学习；多视图学习；视图裁剪；时序对比学习

中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202550418 **CSTR:** 32373.14.issn1000-1239.202550418

在传统机器学习范式中，模型在训练阶段完成参数学习与优化后即被固定为“一次训练、终身不变”的静态模型[1]，依赖训练与测试分布一致性的假设，却忽视了测试部署环境中数据分布随时间和场景演进带来的数据变化问题[2]. 这一静态思路的局限性在自动驾驶、视频监控等对实时性与鲁棒性要求极高的场景下尤为明显，当模型遭遇传感器噪声、光照变化或环境干扰时，性能往往会呈现出明显的下降[3]. 此外，传统方法缺乏在线自适应更新机制，使得在测试阶段无法充分利用新采集的无标签或少量标注样本来校正模型，从而增加了训练时与测试时环境差异带来的不匹配风险[4]. 为缓解分布偏移带来的性能下降，域自适应（domain adaptation）被提出，旨在利用源域与目标域数据间的迁移或泛化能力来提高模型在新域中的表现[5]. 然而，这些方法通常需要预先获取目标域的标注或无标注数据进行联合训练，难以满足部署后的模型对实时在线适应的需求[6]. 基于此，测试时适应（test-time adaptation, TTA）作为一种域自适应改良的高级学习范式，能够通过仅使用源域训练的模型和测试阶段的无标签数据，在推理前或推理过程中进行快速自适应. TTA 已在图像分类、语义分割和多模态融合等任务中展现出显著潜力[7]，近期研究表明，通过在测试阶段对预训练模型进行适应，可以在多个任务中实现 5%~10% 的性能提升[8-10].

现有的 TTA 方法主要可以分为三大类：1）基于无监督目标的在线更新. 此类方法无需任何测试标签，直接在测试阶段利用无监督损失对模型进行实时微调. 典型策略包括通过最小化预测熵（entropy minimization）来驱动批归一化参数更新，在批归一化层中重新计算或校正统计量以适应新分布[11]，以及结合自监督辅助任务（如旋转预测、对比学习等）增强更新阶段的稳定性. 这些方法能够在无人干预的情况下，针对每个测试批次或样本进行动态校正，从而缓解分布偏移带来的性能下降. 2）增强适应性的鲁棒 TTA. 针对测试时可能出现的标签分布偏移和自然（或恶意）噪声样本，这类方法在在线更新中引入了额外的鲁棒性机制. 例如，MedBN 通过检测并剔除异常样本来抵御恶意攻击[12]；测试时归一化（test-time normalization, TTN）设计了域敏感的归一化插值策略，有效平衡了源端与目标端的统计量，提升了模型的跨域鲁棒性[13]. 这些方法通过噪声过滤、动态重加权和稳健优化，使得在非理想测试环境下依然可以保持模型表现的稳定性. 3）持续流式的记忆增强 TTA. 在长期部署或流式数据场景中，模型连续在线更新会导致对早期环境的“灾难性遗忘”. 针对这一挑战，这类方法引入了记忆模块和数据缓冲机制：MECTA 使用高效的内存缓存策略保存关键中间结果，以减小在线微调的内存开销并维持旧知识的留存[14]；AR-TTA 通过小规模样本缓冲区与混合增广的结合，在不断吸收新分布信息的同时保留源域知识[15]. 这些方法通过周期性回放或选择性回顾，实现了流式环境的持续自适应与知识保留.

尽管单一模态的 TTA 已在图像分类、语义分割等任务中取得卓越表现[16-19]，但在涉及视觉、语音、文本等多源信息融合的场景下，这类方法难以同时应对各通道的分布偏移问题，不能满足跨通道的动态协同需求[20]. 多模态测试时适应(multimodal test-time adaptation, MM-TTA)需要在无标签的测试阶段，实时校正每种模态的特征分布，并在此基础上实现不同模态间的对齐与自适应融合[21]，充分发挥多源数据的互补优势并保障系统整体的稳健性. 尽管已有研究尝试通过引入跨模态的一致性约束和设计自适应融合网络来增强 MM-TTA 的能力[22]，但在处理模态丢失、噪声干扰以及复杂时序依赖建模等方面依旧存在明显短板. 由此可见，深入探究 MM-TTA 的算法架构与理论原理，对于推动多模态模型在开放、动态环境中的高效部署，具有重要的学术价值和广泛的工程意义.

在开放部署场景中，MM-TTA 不仅要应对传统分布偏移，还需在噪声干扰、设备故障与动态环境变化的多重压力下，确保异构模态信息的高效感知与融合. 具体而言，首先，不同模态或视角在实际测试过程中的可靠度会因遮挡、传感器抖动、通信延迟或模型预测不确定性等因素大幅波动，传统的固定权重融合或简单拼接往往无法及时识别并剔除这些“失效”视图，从而引发关键语义的误判或遗漏，严重影响最终决策；其次，流式数据中经常出现过曝/欠曝帧、模糊帧、背景噪声乃至丢帧断流等低质量输入，如果将它们直接纳入在线微调，会导致有害梯度累积、模型更新方向偏离，进而削弱对其他高质量模态信号的响应能力；最后，对于视频、音频等高维时序模态，现有方法多聚焦于对齐浅层特征，忽略了帧间的上下文关联、语义演进以及长短期依赖，当面临突发场景切换或短时噪声时，模型难以维持编码的一致性，无法充分利用时序结构去过滤干扰和补全缺失信息，从而限制了持续自适应的效果. 面对这三大挑战，构建能够动态感知视图可靠度、智能筛除噪声样本并兼顾时序连贯性的 MM-TTA 机制，已成为实现高鲁棒性在线部署的关键所在.

为全面应对上述三大挑战，本文提出了一种置信度引导的视图裁剪与时序对比注意力(confidence-guided view pruning and temporal contrastive attention, CVPTA)框架. CVPTA 在测试阶段只需进行少量在线更新便能实现跨模态可靠度自适应、低质视图智能剔除与时序一致性增强的协同优化. 具体而言，CVPTA 首先利用模型对每个视图的预测置信度动态生成注意力权重，有效压制由遮挡或预测不确定性带来的低质量视图干扰；随后通过模糊度和曝光度评估自动识别并剔除低质量帧，再结合邻近高质量视图的重采样完成信息补全，切断噪声梯度的累积路径；与此同时，借助时序对比自监督策略，将同一样本的不同视图视为正样本，不同样本的视图视为负样本，强化了视频与音频流的内部时间连贯性. 上述三大模块与主任务目标及跨视图对齐损失在一次端到端优化中紧密融合，无需任何额外标注或外部先验，即可在每个测试批次上完成在线微调. 实验结果显示，CVPTA 在 Kinetics-50-C 和 VGGSound-C 两大多模态扰动基准上分别实现了超过 2.3 个百分点和 0.7 个百分点的准确率提升，并在极端噪声场景中依旧保持超过 0.2 个百分点的增益.

本文的主要贡献包括 3 个方面：

1)聚焦 MM-TTA 这一创新且尚未充分探索的研究课题. 该问题在分布偏移与多源异质信息协同等方面具有广泛的实际应用需求，但当前研究仍较为稀缺. 本文从跨模态协同、视图质量和时序一致性 3 个维度，全面剖析在分布偏移和噪声干扰条件下，多模态 TTA 所面临的核心挑战，为该领域提供了清晰的问题定义和研究方向.

2)提出端到端的 CVPTA. CVPTA 集成置信度引导动态注意力、低质量视图裁剪以及时序对比自监督三大模块，实现了在无需额外标注的测试阶段对多模态特征的自适应融合与快速校正.

3)本文在 Kinetics-50-C 和 VGGSound-C 两大扰动基准上统一采用在线更新协议评测，并通过消融研究验证了熵引导注意力、视图裁剪与时序对比三大模块的独立与协同贡献.

1 相关工作

本研究采用递进式分析框架，系统性地梳理 TTA 领域的研究进展. 首先从单模态适应(1.1 节)的基础范式出发，剖析其在动态环境的建模能力与持续适应稳定性方面的固有局限；进而拓展至多模态场景(1.2 节)，深入分析模态间协同效应与信息不一致性带来的适应挑战；在此基础上引入多视图建模(1.3 节)，探讨其在动态环境下的信息增强与特征对齐机制；最终聚焦对比学习(1.4 节)，阐释其如何为多模态 TTA 的多视图特征对齐提供增强手段. 通过“基础范式—扩展挑战—增强机制—优化手段”的递进式分析路径，本研究在现有方法与前沿趋势的交

叉领域，创新性地提出了一个具有结构一致性与多视图融合能力的全流程多模态 TTA 框架，为后续方法部分的展开奠定了坚实的基础. 这一分析框架不仅清晰地勾勒出 TTA 领域的发展脉络，更通过 4 个维度的有机整合，凸显了本研究的创新点与技术突破方向.

1.1 单模态 TTA

近年来，单模态 TTA 技术在解决机器学习模型在测试阶段面临的域偏移问题上取得了突破性进展. 该技术的核心思想是通过在测试过程中动态调整模型参数，使预训练模型能够适应目标域的数据分布变化，从而缓解源域与目标域之间的分布差异问题. 从技术发展脉络来看，当前单模态 TTA 研究主要沿着 3 个重要方向推进：首先是以熵最小化等无监督目标为基础的在线 TTA 方法[23-25]，这类方法通过实时更新批归一化层参数实现了计算高效和部署便捷的优势，但在处理复杂多变的分布偏移时仍存在鲁棒性不足的缺陷；其次是针对实际应用场景中特定挑战的鲁棒 TTA 方法[26-30]，如应对标签偏移问题的样本筛选策略、处理单样本适应场景的隐空间对齐技术等，这些方法通过专门的优化设计显著提升了模型在非理想条件下的适应性能；最后是面向持续域偏移问题的持续 TTA 方法[31-34]，通过引入模型回滚机制和记忆缓冲技术，有效解决了长期适应过程中的灾难性遗忘问题. 值得注意的是，TTA 技术的应用范围已经从最初的图像分类任务扩展到语义分割、人体姿态估计等更复杂的计算机视觉任务[35-39]，极大地拓展了其实际应用价值. 然而，现有单模态 TTA 方法仍存在 2 个关键性局限：一方面，仅依赖单一数据模态的特性限制了模型利用多源信息互补优势的能力；另一方面，对模态内上下文关系的动态建模能力不足，导致在开放环境中的适应性受限. 这些技术瓶颈促使研究界开始将目光转向更具挑战性但也更有前景的多模态 TTA 研究，通过整合视觉、语言、音频等多模态信息，有望构建出更强大、更鲁棒的自适应系统，为人工智能系统在复杂开放环境中的实际部署提供新的技术解决方案.

1.2 MM-TTA

MM-TTA 作为深度学习领域的前沿研究方向，正在重塑传统测试时自适应的技术范式，其核心挑战不仅在于处理模态内部的分布偏移，更涉及跨模态的动态协同、信息互补与特征对齐等复杂问题. 当前研究呈现出多元化发展态势，在跨模态特征对齐方面，基于联合优化的多模态归一化层参数调整和对抗训练方法虽取得一定成效[40-42]，但对模态间语义层级关系的建模仍显不足；多模态提示学习充分利用视觉-语言大模型的泛化能力[43-46]，通过动态调整文本或视觉提示适应分布变化，却在处理模态内部复杂上下文关系时存在局限. 与此同时，模态互补增强技术通过知识迁移和分布校准策略显著提升噪声模态鲁棒性[47-48]，而动态记忆机制则通过在线记忆库存储多模态特征实现增量式适应[49-50]，但都面临各自的技术瓶颈. 最新研究趋势正将自监督学习和元学习等先进技术相融合，致力于构建端到端的多模态自适应框架[51-52]，但这些方法在复杂分布偏移下稳定性不足.

1.3 多视图学习

多视图学习作为机器学习领域的重要分支，其核心价值在于通过挖掘多源数据间的互补性和一致性来增强模型的表征能力. 传统方法主要围绕表征对齐[53-57]和表征融合[58-63] 2 个方向展开，但这些方法普遍存在 2 个关键性局限：一方面受限于线性假设或浅层模型架构，难以捕捉数据间复杂的非线性关系；另一方面主要针对静态数据分布设计，缺乏对动态环境变化的适应能力. 随着深度学习与多模态学习的深度融合，新一代方法通过更先进的网络架构和训练机制显著提升了多视图表征的鲁棒性. 当前研究主要从 3 个维度推动该领域发展：解耦表征学习（如 DVANet[64]）通过分离视图不变特征与视图特定特征，在跨中心医疗影像分割等任务中实现了 12.3% 的性能提升；动态多模态融合方法（如 MMNet[65]）结合时序建模，有效解决了传统方法对时序信息建模不足的问题；隐私保护范式（如 MUTA）则通过测试时自适应框架，避免了医疗等敏感领域的数据集中需求. 值得注意的是，尽管这些创新方法在各自应用场景中取得了显著进展，但普遍未能充分考虑分布偏移问题，这为模型在未知环境中的表现稳定性和泛化能力埋下了隐患，亟待后续研究加以解决.

1.4 对比学习

对比学习作为自监督学习的重要范式，通过构建正负样本对来学习判别性表征，在表征学习领域取得了显著进展. 早期工作如 SimCLR 和 MoCo[66] 通过数据增强和动量编码器技术，证明了无监督对比学习的强大表征能力；随后发展的语义增强方法（如 SupCon[67] 和 PCL[68]）引入类别信息，在细粒度分类等任务中实现了较高的性能提升. 跨模态对比学习方面，CLIP 和 ALIGN[69] 通过大规模图像-文本对训练，建立了跨模态对齐的嵌入空间. 为提升计算效率，

BYOL[70] 和 SimSiam[71] 探索了无需负样本的简化架构. 不同于之前的工作, 我们致力于将对比学习和多视图学习进行融合, 通过拉近同一样本的不同视图的特征而拉远不同样本的特征, 为 MM-TTA 的多视图特征对齐提供增强手段.

2 方 法

本节将系统地阐述本文所研究的 MM-TTA 问题及所提出的框架 CVPTA.

2.1 问题设定

2.1.1 源域与目标域定义

本文以音频模态 a 和视频模态 v 为例, 定义源域有标注样本集:

$$\mathcal{D}_{\mathrm{s}}=\left\{(x_i^a,x_i^v,y_i)\right\}_{i=1}^{N_{\mathrm{s}}},(x_i^a,x_i^v,y_i)\sim Q(x,y), \tag{1}$$

其中 x_i^a 表示音频输入, x_i^v 表示视频输入, 标签 $y_i\in\{1,2,\cdots,K\}$. 同时, 无标注的目标域数据流表示为:

$$\mathcal{D}_{\mathrm{t}}=\{(x_j^a,x_j^v)\}_{j=1}^{N_{\mathrm{t}}},(x_j^a,x_j^v)\sim P(x),P\neq Q. \tag{2}$$

目标域数据以批次方式 $B_{\mathrm{t}}=\{(x_j^a,x_j^v)\}_{j=1}^{N}$ 顺序到达.

2.1.2 源域模型结构

在源域 $\mathcal{D}_{\mathrm{s}}$ 上训练完毕后得到模型参数:

$$\Theta_{\mathrm{s}}=\left\{f_{\Theta_{\mathrm{s}}^a}^a,f_{\Theta_{\mathrm{s}}^v}^v,f_{\Theta_{\mathrm{s}}^m}^m,f_{\Theta_{\mathrm{s}}^c}^c\right\}, \tag{3}$$

其中:

1) $f_{\Theta_{\mathrm{s}}^a}^a$: 音频编码器, 将原始音频信号映射到潜在特征空间;

2) $f_{\Theta_{\mathrm{s}}^v}^v$: 视频编码器, 将单帧或多帧视频输入映射到潜在特征空间;

3) $f_{\Theta_{\mathrm{s}}^m}^m$: 多模态融合层, 对音频与视频特征进行拼接或注意力融合;

4) $f_{\Theta_{\mathrm{s}}^c}^c$: 分类器, 将融合特征映射到类别预测 logits.

2.1.3 MM-TTA 目标

在不访问源域数据 $\mathcal{D}_{\mathrm{s}}$ 的前提下, 仅基于无标签目标域数据流 $\mathcal{D}_{\mathrm{t}}$, 通过在线微调优化 Θ_{s} 中的可更新子模块(主要是融合层 f^m 与分类器 f^c)参数, 使模型能够适应目标域分布 P 的偏移并保持高精度.

2.2 CVPTA 框架

CVPTA 是一种面向多模态数据的在线自适应框架. 如图 1 所示, 其设计思路是: 为每个“视频片段”抽取多视图, 先行剔除劣质视图; 再在融合特征空间中通过多重损失进行在线更新; 最后在推理阶段基于视图置信度动态加权融合得到最终预测.

1)多视图采样与低质量视图裁剪. 在测试时对每个视频片段同时抽取 2 帧视图, 保证仅使用视觉质量高的帧以减少噪声干扰.

2)训练阶段. 无需任何额外标注, 仅利用无标签目标域数据, 通过最小化跨模态对齐损失 L_{align}、分布平衡损失 L_{bal}、可靠性感知损失 L_{ra} 以及时序对比自监督损失 L_{temp} 对模型可更新参数(主要是融合层 f^m 与分类器 f^c)进行在线更新.

3)推理阶段. 对 2 个视图分别计算对数概率 logits 并获得置信度, 再基于置信度动态线性加权融合 logits, 输出最终预测 y_j.

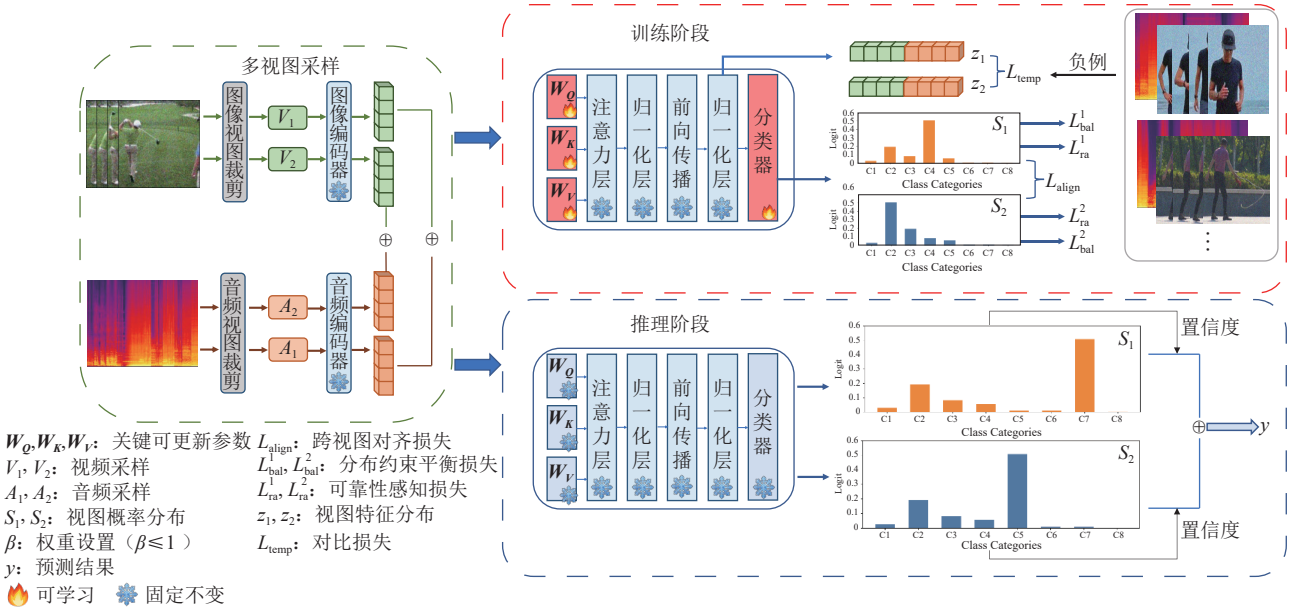


Fig. 1 The pipeline of CVPTA

图 1 CVPTA 流程图

2.2.1 多视图采样与低质视图裁剪

在视频流场景中，单帧图像往往无法完整表达整个片段的动态语义，且某些帧可能因运动模糊、光照过曝/欠曝或遮挡而信息稀少. 直接使用低质量帧参与在线更新，会导致有害梯度累积并损害自适应效果. 因此，本文在测试时对每段视频进行多帧采样并预先裁剪低质量帧，保证后续模块仅对“高质量视图”进行编码与优化，从源头减少噪声干扰. 具体流程分 4 个步骤.

1）候选视图生成

对于批次 B_{t} 中的每个视频片段 x_j^v，均匀抽取 M 帧候选视图：

$$\{x_j^v(k)\mid k=1,2,\cdots,M\}. \tag{4}$$

同时，针对对应的音频信号 x_j^a，本文可直接对齐 2 段时间窗口（对应帧时间）或使用整体音频作为 2 个视图输入，以保证音频和视频的同步性.

2）帧级质量评估

从 2 个角度对第 k 帧图像进行质量评估：

① Laplacian 方差（模糊度量）

$q_j^{\mathrm{lap}}(k)=Var(\nabla^2 x_j^v(k))$，用于评估细节丰富度和画面清晰度，方差越大表示采样图像越清晰、细节越丰富；

②亮度直方图偏度（曝光度量）

$q_j^{\mathrm{lum}}(k)=skewness(hist(x_j^v(k)))$，偏度值用于检测过曝（正偏度）或欠曝（负偏度）情况.

3）合并成整体质量分数

$$q_j(k)=\alpha q_j^{\mathrm{lap}}(k)+(1-\alpha)q_j^{\mathrm{lum}}(k),\alpha\in[0,1]. \tag{5}$$

对 $\{q_j(k)\}_{k=1}^{M}$ 做 min–max 归一化，保证最终结果落在 [0,1]的区间内.

4）低质视图裁剪

按降序选取质量最高的 2 帧索引 $\{k_1,k_2\}=\mathrm{top2}(\{q_j(k)\})$. 检索对应视图帧：

$$x_j^{v_1}=x_j^v(k_1),x_j^{v_2}=x_j^v(k_2). \tag{6}$$

同时选取对应的音频段，或统一使用 x_j^a 作为 2 个视图的音频输入，形成 2 套“高质量”的视图组合：

$$(x_j^{a_1},x_j^{v_1}),(x_j^{a_2},x_j^{v_2}). \tag{7}$$

多视图采样与低质量视图裁剪模块能够在测试阶段同步抽取 2 个高质量帧，既有效避免了因模糊或曝光异常的劣质帧带来的有害梯度，又保留了视频片段前后语义演进的信息，为后续对比学习与特征融合提供了更加丰富的时序样本. 此外，提前完成视图质量筛选，还能在在线更新中减少无效视图的计算开销，从而大幅提升自适应的效率和整体鲁棒性.

2.2.2 跨模态对齐损失 L_{align}

在 2.2.1 节得到 2 组“高质量”多模态视图对 $\{(x_j^{a_1},x_j^{v_1}),(x_j^{a_2},x_j^{v_2})\}$ 后，分别经编码器提取特征并融合.

1）特征提取与融合

$$z_{j,\ell}^a=f_{\Theta_{\mathrm{s}}^a}^a(x_j^{a\ell}),z_{j,\ell}^v=f_{\Theta_{\mathrm{s}}^v}^v(x_j^{v\ell}), \tag{8}$$

$$z_{j,\ell}=concat(z_{j,\ell}^a,z_{j,\ell}^v),z_{j,\ell}^m=f_{\Theta_{\mathrm{s}}^m}^m(z_{j,\ell}), \tag{9}$$

其中 $\ell\in\{1,2\}$. 最终得到 2 组融合特征 $z_{j,1}^m,z_{j,2}^m$.

2）损失函数的计算过程

令 $\boldsymbol{p}_{j,\ell}=Softmax\left(f_{\Theta_{\mathrm{s}}^c}^c\left(z_{j,\ell}^m\right)\right)\in\mathbb{R}^K,\ell=1,2.$

定义 Jensen–Shannon 散度为：

$$D_{\mathrm{JS}}(P\|Q)=\frac{1}{2}D_{\mathrm{KL}}(P\|M)+\frac{1}{2}D_{\mathrm{KL}}(Q\|M), \tag{10}$$

其中 $M=\frac{1}{2}(P+Q)$，则

$$L_{\mathrm{align}}=\frac{1}{|B_{\mathrm{t}}|}\sum_{j\in B_{\mathrm{t}}}D_{\mathrm{JS}}(p_{j,1}\|p_{j,2}). \tag{11}$$

跨模态对齐损失能够显著缓解目标域分布偏移带来的单视图偏颇问题：当某个视图出现偏向时，该损失会强制另一视图对预测概率分布保持一致，迫使模型在更新过程中同时考虑 2 种视图的信息，从而避免片面偏移. 同时，通过对齐概率流，模型在融合特征时能够更好地整合来自音频和视频 2 种模态的互补信息，提升融合后特征的稳健性和泛化能力.

2.2.3 分布约束平衡损失 L_{bal}

为防止模型在多模态融合后对少数类别或单个视图过度依赖，本文对整个批次 B_{t} 中的所有视图 logits 进行归一化统计. 令批次大小为 B_{t}，共有 $2B_{\mathrm{t}}$ 个视图特征 $z_{j,\ell}^m$，其对应的 logits 为

$$o_{j,\ell}=f_{\Theta_{\mathrm{s}}^c}^c(z_{j,\ell}^m),p_{j,\ell}=Softmax(o_{j,\ell}). \tag{12}$$

定义类别累积概率：

$$p_{\mathrm{sum}}^{(c)}=\sum_{j\in B_{\mathrm{t}}}\sum_{\ell\in\{1,2\}}[p_{j,\ell}]_c,\ c=1,2,\cdots,K. \tag{13}$$

则

$$L_{\mathrm{bal}}=-\sum_{c=1}^{K}\frac{p_{\mathrm{sum}}^{(c)}}{\sum_{k=1}^{K}p_{\mathrm{sum}}^{(k)}}\log\left(\frac{p_{\mathrm{sum}}^{(c)}}{\sum_{k=1}^{K}p_{\mathrm{sum}}^{(k)}}\right). \tag{14}$$

分布约束平衡损失能够防止模型在多模态融合时对某些类别或视图过度自信，从而忽略其他有价值的信息；通过最大化整体类别概率分布的熵，它督促模型在各类别之间保持相对均衡，使得次优但仍有意义的模态信息也能得到关注，有效减缓因跨模态不平衡而导致的性能退化，从而使在线自适应过程更加稳定且具有针对性.

2.2.4 可靠性感知损失

当某些视图因噪声或分布偏移导致预测不确定度高时,模型需要识别并放低该视图在在线更新中的权重,以免有害梯度干扰.为此,本文设计了可靠性感知损失 L_{ra}:

1)预测熵与可靠度计算

对每个融合特征 $z_{j,\ell}^{m}$ 计算预测分布:

$$p_{j,\ell}=Softmax(f_{\Theta_{\mathrm{s}}^{c}}^{c}(z_{j,\ell}^{m})),$$
$$H(p_{j,\ell})=-\sum_{c=1}^{K}p_{j,\ell}[c]\log p_{j,\ell}[c], \tag{15}$$

其中 $H(p)$ 为分布熵.定义视图可靠度权重:

$$\sigma\left(p_{j,\ell}\right)=\frac{1}{1+\mathrm{e}^{-H\left(p_{j,\ell}\right)}}\in(0,1). \tag{16}$$

当 $H(p)$ 越大(即不确定程度越高), $\sigma(p)$ 越接近 0; 当 $H(p)$ 越小, $\sigma(p)$ 越接近 1.

2)可靠性感知损失形式

$$L_{\mathrm{ra}}=\frac{1}{2|B_{\mathrm{t}}|}\sum_{j\in B_{\mathrm{t}}}\sum_{\ell=1}^{2}\sigma\left(p_{j,\ell}\right)\left(H\left(p_{j,\ell}\right)+\gamma\right), \tag{17}$$

其中 $\gamma>0$ 为小常数,用于平衡熵的数量级.在线更新阶段,可靠性感知机制能够自动识别高熵(高不确定度)视图并降低其对损失的贡献,此时 $\sigma(p)$ 倾向于接近 0,从而减少低可靠度视图的梯度权重;与之相反,当视图预测熵较低(即置信度较高)时, $\sigma(p)$ 接近 1,会突出该视图的更新信号,充分利用高质量视图信息.通过这种动态区分视图可靠度的方式,模型在分布漂移和噪声干扰环境下能够更加稳健地进行在线自适应.

2.2.5 时序对比自监督损失 L_{temp}

在 2.2.1 节完成“高质量”多视图采样后,得到 2 个融合特征 $z_{j,1}^{m},z_{j,2}^{m}$;在一个批次 B_{t} 中,其他视频样本也各自有 2 个融合特征 $z_{k,1}^{m},z_{k,2}^{m},k\neq j$.利用对比学习,本文定义时序对比自监督损失 L_{temp} 来强化同一视频段内视图的时序连贯性,并推远不同视频样本的特征距离.

$$L_{\mathrm{temp}}=-\frac{1}{2|B_{\mathrm{t}}|}\sum_{j\in B_{\mathrm{t}}}\sum_{\ell=1}^{2}\log\frac{\exp(z_{j,\ell}^{m\mathrm{T}}z_{j,3-\ell}^{m}/\tau)}{\exp(z_{j,\ell}^{m\mathrm{T}}z_{j,3-\ell}^{m}/\tau)+\sum\limits_{(k,u)\neq(j,\ell)}\exp(z_{j,\ell}^{m\mathrm{T}}z_{k,u}^{m}/\tau)}, \tag{18}$$

其中 τ 为温度超参数.

1)正样本对:同一视频段的 2 个融合特征 $z_{j,1}^{m},z_{j,2}^{m}$.

2)负样本对:批次中所有其他视频的视图融合特征 $\{z_{k,u}^{m}\mid k\neq j,u\in\{1,2\}\}$.

时序对比自监督通过拉近同一视频段中 2 个视图的特征距离,强化了短期时序语义的一致性,即使某一帧受到噪声干扰,另一帧依然可以提供稳定信号,从而提升模型对连续语义的捕捉能力;与此同时,它又将不同视频段的视图特征相互分离,增强了模型对不同样本间差异的判别能力,避免在线更新过程中被异类信息所干扰.

2.2.6 置信度引导动态注意力融合

在推理和微调过程中,本文对 2 个融合特征 $z_{j,1}^{m},z_{j,2}^{m}$ 分别计算分类器输出 logits 及预测概率,并使用置信度作为视图融合的动态权重.

1)计算 logits 与预测概率:

$$o_{j,\ell}=f_{\Theta_{\mathrm{s}}^{c}}^{c}\left(z_{j,\ell}^{m}\right),p_{j,\ell}=Softmax\left(o_{j,\ell}\right),\ell=1,2. \tag{19}$$

2)置信度定义 $c_{j,\ell}=\max\limits_{k}p_{j,\ell}[k]$,即第 ℓ 个视图对样本 j 的最高预测概率.

3)动态加权融合

$$Logits\tilde{o}_{j}=c_{j,1}o_{j,1}+c_{j,2}o_{j,2}, \tag{20}$$
$$\tilde{p}_{j}=Softmax(\tilde{o}_{j}),y_{j}=\arg\max_{k}\tilde{p}_{j}[k]. \tag{21}$$

动态注意力机制会根据每个视图在当下帧的预测置信度动态分配权重,将更可靠的视图置于更高优先级,弱化那些因噪声干扰而置信度较低的视图,从而显著提升融合结果的稳健性,确保在测试阶段遇到分布漂移时,模型不会仅凭单一的失真视图做出错误判断.

2.2.7 联合优化与算法伪代码

将各子损失按权重线性组合,得到总体训练损失:

$$L_{\mathrm{train}}=L_{\mathrm{align}}+\lambda_{\mathrm{bal}}L_{\mathrm{bal}}+\lambda_{\mathrm{ra}}L_{\mathrm{ra}}+\lambda_{\mathrm{temp}}L_{\mathrm{temp}}, \tag{22}$$

其中

1) L_{align}:使用 Jensen–Shannon 散度衡量 2 个视图预测概率分布一致性;

2) L_{bal}:通过最大化整体类别概率熵保证预测分布均衡, λ_{bal} 为该损失对应的权重;

3) L_{ra}:利用预测熵与 Sigmoid 抑制高不确定度视图梯度, λ_{ra} 为该损失对应的权重;

4) L_{temp}:拉近同视频段不同视图特征、推远不同视频段特征, λ_{temp} 为该损失对应的权重.

以上各项损失共同驱动融合层 f^{m} 和分类器 f^{c} 在测试阶段进行少量迭代梯度更新,保持学习率 η 固定.

算法 1. CVPTA:多视图不确定性引导与时序对比在线自适应.

输入:源域模型 $\Theta_{\mathrm{s}}=\{f^{a},f^{v},f^{m},f^{c}\}$,无标注测试流

$D_{\mathrm{t}}=\{x_j^a,x_j^v\}_{j=1}^{N_{\mathrm{t}}}$;

输出：实时预测t_j以及更新之后的f^m,f^c.

for B_{t} in D_{t} do:

①多视图采样与低质视图裁剪

for $\left(x_j^a,x_j^v\right)$ in B_{t} do:

抽取M帧候选视图，并计算质量分数$q_j(k)$;

选取质量最高的2帧作为$x_j^{v_1},x_j^{v_2}$;

生成对应的2组音频视图$x_j^{a_1},x_j^{a_2}$;

end for

②特征提取和多模态融合

for l in {1, 2} do:

$z_{\{j,l\}}^a=f^a(x_j^{\{a,l\}})$;

$z_{\{j,l\}}^v=f^v(x_j^{\{v,l\}})$;

$z_{\{j,l\}}=concat(z_{\{j,l\}}^a,z_{\{j,l\}}^v)$;

$z_{\{j,l\}}^m=f^m(z_{\{j,l\}})$;

end for

③L_{align}的计算

f^c通过计算$p_{\{j,1\}},p_{\{j,2\}}$;

基于式(11)得到L_{align};

④L_{bal}的计算

收集当前batch所有的$p_{\{j,l\}}$;

计算每个y对应的概率的和;

基于式(14)得到L_{bal};

⑤L_{ra}的计算：

计算$H(p_{\{j,l\}})$和$\sigma(p_{\{j,l\}})$;

基于式(17)得到L_{ra};

⑥L_{temp}的计算：

for j in B_{t} do:

for l in {1, 2} do:

计算正对相似度;

计算负对相似度;

基于式(18)得到L_{temp};

end for

end for

⑦联合优化更新基于式(22)：

$L_{\mathrm{train}}=L_{\mathrm{align}}+\lambda_{\mathrm{bal}}L_{\mathrm{bal}}+\lambda_{\mathrm{ra}}L_{\mathrm{ra}}+\lambda_{\mathrm{temp}}L_{\mathrm{temp}}$;

⑧置信度引导动态注意力融合与预测：

for j in B_{t} do:

计算$o_{\{j,1\}},o_{\{j,2\}}$和$p_{\{j,1\}},p_{\{j,2\}}$;

$c_{\{j,1\}}=\max p_{\{j,1\}}$;

$c_{\{j,2\}}=\max p_{\{j,2\}}$;

$\tilde{o}_j=c_{\{j,1\}}\cdot o_{\{j,1\}}+c_{\{j,2\}}\cdot o_{\{j,2\}}$;

$y_j=\arg\max(softmax(\tilde{o}_j))$;

end for

end for

3 实验结果及分析

本节在模态可靠性存在偏差的MM-TTA场景下，评估所提出的CVPTA框架在视听联合行为识别和事件分类任务中的性能.

3.1 基准构建与实验设置

为系统研究MM-TTA中由于视图质量不稳定导致的模型表现下降，本研究基于广泛使用的Kinetics-50[72]和VGGSound[73]多模态数据集构建了2个基准测试集. 参照Hendrycks & Dietterich的方法[74]，为视频模态设计了11种、为音频模态设计了6种不同类型的干扰. 其中，针对视频模态设计的干扰分为噪声添加(Noise)、动态模糊(Blur)以及数码干扰(Digital)三个大方向，并在每个方向下选择了最具代表性的几种具体干扰形式，如噪声添加方向选择了高斯模糊(Gaussian)等3种干扰类型作为最终的任务设置；针对音频模态的干扰设置在思路上与视频类似，且因为音频模态本身的数据特性，只设置了Noise和天气影响(Weather)两类干扰方向，并在其中设置了以Gaussian和交通噪声(Traffic)为代表的6种经典干扰环境. 在每个具体的干扰任务中本文都按照干扰强度由低到高设置了5个强度等级，在本节中使用的所有实验结果具体使用的干扰强度都是最高等级的五级干扰强度，以此为依据最终形成包含模态损坏的Kinetics-50-C和VGGSound-C基准测试集.

本文将所提出的CVPTA框架与当前5种具有代表性的TTA方法(Tent, MMT, EATA, SAR, READ)进行了系统对比分析，涵盖单模态与多模态TTA的典型范式，具体说明如下：

1) Tent[75]. 在测试阶段通过最小化模型预测输出的熵值，提升模型对分布偏移样本的适应能力，广泛适用于图像分类、语义分割等任务，并在多个标准数据集上取得了优异性能.

2) MMT[76]. 基于多模态3D语义场景构建，利用互补模态间的信息融合增强自监督信号，从而提升多模态自适应分割与表示学习能力.

3) EATA[77]. 引入不确定性感知的样本选择机制，动态识别可靠样本并引导模型在稳定子集上进行测试时更新，有效缓解分布偏移引起的灾难性遗忘问题.

4) SAR[78]. 通过抑制大梯度噪声样本并引导模型优化方向靠近平坦解，实现对混合分布、非平稳环境

下的鲁棒适应.

5)READ[79]. 设计模态级的可靠性评估机制, 根据各模态在不同噪声条件下的表现动态调整其在模型更新中的权重, 从而实现多模态噪声环境下的自适应增强.

上述方法分别代表了当前 TTA 研究在熵最小化、自监督增强、样本筛选、梯度平滑与模态可靠性建模等不同方向的主流策略. 后续实验将从准确率、鲁棒性与跨模态协同能力等维度对比分析上述方法与 CVPTA 的性能差异.

本研究提出的 CVPTA 是一个通用框架, 以当前最先进的 CAV-MAE 模型[80]作为基础架构, 该模型已在网络规模的视听数据上进行预训练. 本文在 Kinetics-50 和 VGGSound 训练集上对其进行微调,获得相应的源域模型. 其中, 以 Kinetics-50 和 VGGSound 的训练集作为源域, 而其对应的干扰版本 Kinetics-50-C 和 VGGSound-C 作为目标域. 在 TTA 阶段, 该方法采用 Adam 优化器对源模型的特定参数进行更新: 设置初始学习率为 0.000 1, 批处理大小为 32, 单次迭代周期; 置信度阈值 γ 固定为 e^{-1}. 所有实验均在配备 NVIDIA 2080 GPU 的 Ubuntu 20.04 平台上完成.

参考现有的 TTA 方法, 本文选择平均准确率 (*MA*) 作为模型的衡量指标, 并独立计算每种噪声添加情况下的模型表现, 每种设置下均进行多轮测试计算平均得分作为最终的结果, 计算公式为:

$$MA = \frac{\sum_{i=1}^{n} Accuracy_i}{n},$$

其中 $Accuracy_i$ 为某一噪声添加下的模型单次测试准确率, n 为测试轮次.

3.2 实验结果

从表 1 可见, 在 Kinetics-50 数据集上, CVPTA 在 11 种设置中的 9 个任务上实现了性能领先, 平均准确率显著优于 Tent, MMT, EATA, SAR 和 READ 等主流 TTA 方法, 展现出更强的跨模态协同能力和分布鲁棒性. 表中的"Stat."和"Dyn."分别是静态(statical)和动态(dynamic)的缩写, LN 表示层归一化(layer normalization)机制, LF 表示后融合(late fusion)机制, AF 表示基于注意力的融合(attention-based fusion)机制, SAF 表示自适应的基于注意力的融合(self-adaptive attention-based fusion)机制. 尤其在 Gaussian 噪声场景中, CVPTA 达到 63.2%, 较当前最佳结果(READ)提高了 13.8 个百分点, 在 Shot 和 Impulse 噪声下均领先 2.6 个百分点. 在模糊扰动设置(如 Glass, Zoom)中, CVPTA 在 2 个子任务中分别取得 65.3% 和 68.9% 的最高准确率, 相较 READ 和 Tent 展现出更强的泛化能力. 与此形成对比, 基于 Late Fusion 或伪标签生成策略的方法(如 MM-TTA 和 READ)在模态信号质量不一致的条件下表现出明显不稳定性, 难以有效利用低质量模态的信息. 在 VGGSound 数据集中尤为典型, 由于音频模态受损严重, 这类方法在 Gaussian、散焦模糊(Defocus)及 JPEG 压缩(JPEG)场景下性能下降明显, 反映出其对模态间可靠性偏差的高度敏感性. 总之, CVPTA 通过构建跨模态协同优化机制和保证视图质量与时序一致性, 在高噪声、高偏移环境中保持稳定性能, 验证了多视图建模在 MM-TTA 中的有效性.

Table 1 video-corruption Experimental Data on Kinetics-50 Dataset

表 1 Kinetics-50 数据集上的 video-corruption 实验数据 %

方法	Noise			Blur				Digital			
	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Contract	Elas.	Pix.	JPEG
源 ((Stat.LN)& LF)	31.8	33.4	31.7	64.0	54.3	67.5	61.9	44.0	60.2	61.7	56.4
MM-TTA(Dyn. LN)	46.2	46.6	46.1	58.8	55.7	62.6	58.7	49.3	57.6	56.4	54.6
Tent(Dyn. LN)	28.6	29.8	28.3	63.4	51.1	67.7	61.7	38.6	60.7	61.8	54.9
EATA(Dyn. LN)	31.8	33.3	31.6	64.2	54.6	67.7	62.2	44.2	60.4	62.0	57.0
SAR(Dyn. LN)	31.9	33.3	31.7	63.8	54.0	67.7	61.8	44.0	60.3	62.0	56.5
READ(Dyn. LN)	34.0	34.5	33.8	65.3	57.7	68.7	64.9	48.7	62.9	64.6	59.2
源 (Stat.(LN& AF))	46.8	48.0	46.9	67.5	62.2	70.8	66.7	52.1	65.7	66.5	61.9
Tent(Dyn. LN)	46.3	47.0	46.3	67.2	62.5	71.0	67.6	51.6	66.8	67.2	62.7
EATA(Dyn. LN)	46.8	47.6	47.1	67.2	62.7	70.6	67.2	52.4	65.9	66.8	62.5
SAR(Dyn. LN)	46.7	47.4	46.8	67.0	61.9	70.4	66.4	52.1	65.7	66.4	62.0
READ(SAF)	49.4	49.7	49.0	**68.0**	65.1	**71.2**	68.2	53.3	68.2	70.1	68.3
CVPTA	**63.2**	**52.3**	**51.6**	67.0	**65.3**	71.1	**68.9**	**54.4**	**69.3**	**71.1**	**70.7**

注: 黑体数值为最优值, "Elas."是"Elastic-transform"的缩写, "Pix."是"Pixelate"的缩写.

从表 2 可以观察到，CVPTA 框架在 Kinetics-50 数据集的 6 种音频扰动设置下均取得最优性能，全面优于 Tent, MMT, EATA, SAR 和 READ 等主流 TTA 方法. 其中，在 Gaussian 噪声场景下，CVPTA 达到 74.9% 的准确率，较 READ 提高 0.8 个百分点，为所有噪声设置中最高. 在 Crowd 与 Thunder 两类高噪环境中分别取得 71.0% 与 72.4% 的准确率，均显著优于现有方法. 在 Traffic 场景下，虽然 CVPTA 相较 READ 提升 0.1 个百分点，但考虑到任务本身已接近饱和，这一提升依然具有代表意义. 从整体平均准确率来看，CVPTA 达到 71.6%，在所有方法中排名第一，较 READ 高出 0.5 个百分点，相较 Tent 和 MM-TTA 等传统方法分别提升了 2.0 个百分点与 2.2 个百分点. 这一结果不仅说明本文方法在复杂音频扰动下具有更强的泛化能力，也验证了所提出的多模态协同机制在分布不确定性条件下的鲁棒性优势. 表 3 结果进一步验证了 CVPTA 框架在多模态分布扰动下的适应性优势. 在 VGGSound 数据集上的 video-corruption 实验中，CVPTA 在 11 个测试场景中的全部 11 项任务中实现了 SOTA 结果，覆盖 Noise, Blur 和 Digital 三大类扰动设定. 在部分任务中，例如 Gaussian、Defocus 和 玻璃模糊（Glass）三种典型高噪声场景，准确率相较 READ 都有一定的提升，展示了该方法对低质量模态输入的鲁棒建模能力. 相比之下，其他方法（如 Tent 和 SAR）虽在某些任务中表现接近，但整体缺乏一致性，且在模态质量严重不均衡的任务设定中易出现性能退化. 而 CVPTA 借助其跨模态的协同优化设计，联合构建动态注意力机制与跨模态对比损失，有效缓解了模态间可靠性偏差对测试时推理带来的影响. 综合多个任务维度的对比分析结果，CVPTA 表现出稳定和显著的性能优势，说明跨模态协同优化与动态注意力机制和对比损失的协同设计是解决多模态动态适应问题的关键.

3.3 对比分析

3.3.1 消融实验

为了验证框架中每个设计的重要性和有效性，本文会对相关方法的消融实验进行分析.

1）去除熵引导的动态注意力机制. 为验证动态注意力机制的有效性，本文在 Kinetics-50 数据集上进行了系统对比实验. 如图 2 所示，包含动态注意力机制的模型在绝大多数损坏场景下均优于包含无注意力机制的对比模型，尤其在运动模糊（Motion）（71.1% vs. 70.6%）、Defocus（66.9% vs. 66.4%）和雪花噪声（Snow）（64.5% vs. 63.9%）等复杂干扰下表现突出，平均准确率提升 0.5 个百分点.

值得注意的是，在脉冲噪声（Impulse）这一特定损坏类型下，单视图通路略占优势（差异小于 0.1 个百分点），这可能由于极端噪声破坏了跨模态的协同关系，导致动态注意力机制暂时失效. 然而，在更常见的自然干扰（如 Motion、Defocus、Snow 等）中，动态注意力机制通过实时协同不同模态间的信息，显著提升模型鲁棒性.

2）去除低质视图裁剪与伪重采样. 为验证重采样机制的有效性，本文在多种损坏场景下进行了系统

Table 2 audio-corruption Experimental Data on Kinetics-50 Dataset

表 2 Kinetics-50 数据集上的 audio-corruption 实验数据 %

方法	Noise			Weather			平均值
	Gaussian	Traffic	Crowd	Rain	Thunder	Wind	
源 ((Stat.LN)& LF)	71.1	67.8	67.4	67.4	70.6	68.6	68.8
MM-TTA(Dyn. LN)	70.8	69.2	68.5	69.0	69.8	69.4	69.4
Tent(Dyn. LN)	71.1	68.6	67.8	67.4	71.2	68.9	69.2
EATA(Dyn. LN)	71.2	67.9	67.5	67.8	70.9	68.7	69.0
SAR(Dyn. LN)	71.1	67.5	67.4	67.4	70.6	68.6	68.8
READ(Dyn. LN)	71.3	68.5	68.5	68.4	71.8	69.0	69.6
源 (Stat.(LN& AF))	73.7	65.5	67.9	70.3	67.9	70.3	69.3
Tent(Dyn. LN)	73.9	67.4	69.2	70.4	66.5	70.5	69.6
EATA(Dyn. LN)	73.7	66.1	68.5	70.3	67.9	70.1	69.4
SAR(Dyn. LN)	73.7	65.4	68.2	69.9	67.2	70.2	69.1
READ(SAF)	74.1	69.0	69.7	71.1	71.8	70.7	71.1
CVPTA	**74.9**	**69.1**	**71.0**	**71.4**	**72.4**	**71.0**	**71.6**

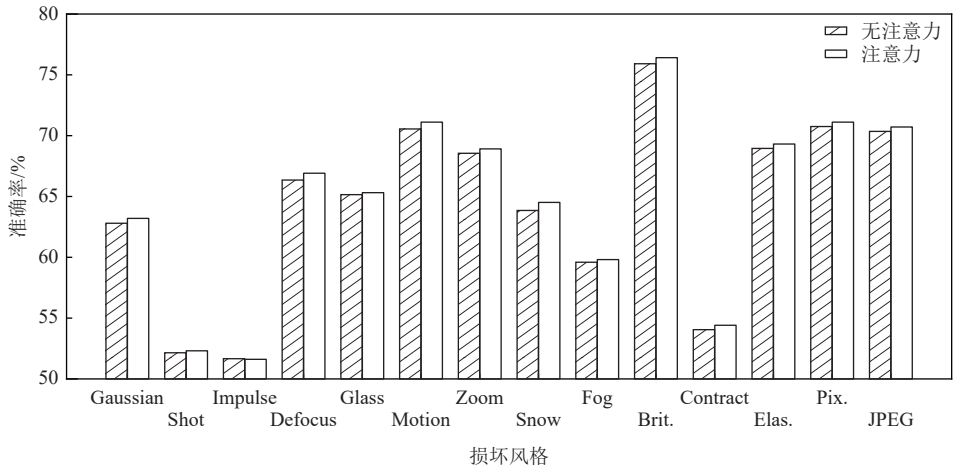
注：黑体数值表示最优值.

Table 3 video-corruption Experimental Data on VGGSound Dataset

表 3 VGGSound 数据集上的 video-corruption 实验数据

方法	Noise			Blur				Digital			
	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Contract	Elas.	Pix.	JPEG
源 ((Stat.LN)& LF)	37.7	36.5	37.8	52.7	51.3	55.2	53.7	45.2	52.5	51.7	52.3
MM-TTA(Dyn. LN)	7.1	7.3	7.3	44.8	41.5	48.0	45.5	24.2	40.3	40.7	45.7
Tent(Dyn. LN)	7.6	6.8	7.2	53.1	52.1	55.5	54.5	16.6	52.6	54.2	53.1
EATA(Dyn. LN)	37.7	36.5	37.7	53.2	52.3	56.0	54.4	45.2	53.5	52.3	52.7
SAR(Dyn. LN)	37.7	36.4	37.7	52.8	51.5	55.5	53.9	44.8	52.7	51.8	52.3
READ(Dyn. LN)	42.1	41.5	42.1	49.3	50.9	53.5	52.5	46.2	52.5	49.1	50.2
源 (Stat.(LN& AF))	52.8	52.7	52.7	57.2	57.2	58.7	57.6	53.7	56.9	55.8	56.9
Tent(Dyn. LN)	52.7	52.7	52.7	56.7	56.5	57.9	57.2	54.0	57.4	56.2	56.7
EATA(Dyn. LN)	53.0	52.8	53.0	57.2	57.1	58.6	57.8	54.1	57.4	56.1	57.0
SAR(Dyn. LN)	52.9	52.8	52.9	57.2	57.1	58.6	57.6	54.0	57.0	56.0	57.0
READ(SAF)	52.8	52.9	52.6	57.1	57.0	58.8	58.0	54.6	58.0	56.6	57.2
CVPTA	**53.8**	**54.1**	**53.7**	**58.1**	**57.9**	**59.7**	**59.0**	**55.2**	**58.9**	**57.7**	**58.4**

注：黑体数值为最优值.“Elas.”是“Elastic-transform”的缩写，“Pix.”是“Pixelate”的缩写.



注：“Brit.”是“Brightness”的缩写，“Elas.”是“Elastic-transform”的缩写，“Pix.”是“Pixelate”的缩写.

Fig. 2 Ablation experiment for entropy-guided dynamic attention mechanisms

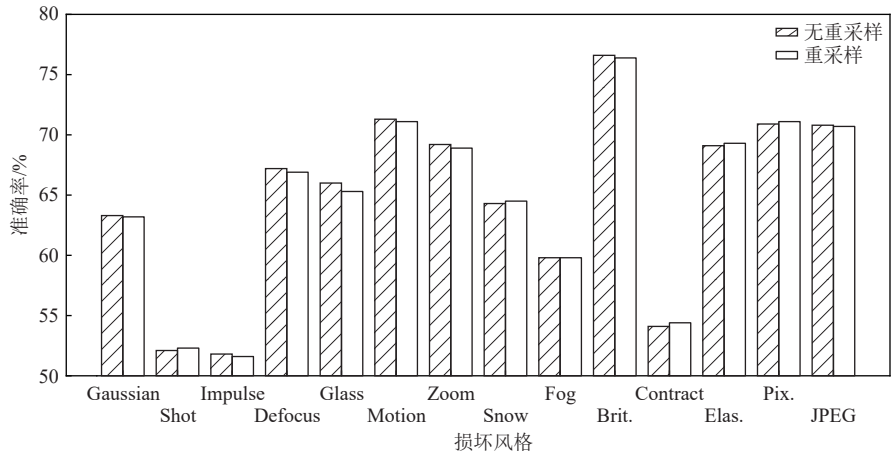
图 2 熵引导的动态注意力机制的消融实验

对比实验. 实验数据如图 3 所示，加入重采样机制的方法在绝大多数任务场景下并未优于未使用重采样机制的方法，特别是在 Defocus(66.9% vs. 67.2%)、Glass(65.3% vs. 66%)和 Motion(71.1% vs. 71.3%)等复杂干扰条件下表现突出，平均性能差距约 0.3 个百分点.

值得注意的是，在 Gaussian 和冲击噪声(Shot)等特定损坏类型下，2 种方法表现相当接近 (差异小于 0.2%)，这可能表明在某些特定噪声分布下，重采样机制的优化效果有限. 然而，在更广泛的自然干扰(如 Defocus, Glass, Motion 等)以及 JPEG(70.7% vs. 70.8%)等常见图像失真情况下，重采样机制通过优化特征分布，显著提升了模型对各类干扰的鲁棒性(平均增益 0.25%). 特别是在光照变化(Brightness)和像素化(Pixelate)等任务中，重采样机制展现出了稳定的性能优势，验证了其在多样化场景下的有效性.

3)去除对比损失. 为验证对比损失的有效性，本文在多种损坏场景下进行了系统对比实验. 实验数据如图 4 所示，加入对比损失机制的方法在多数任务场景下与未使用对比损失的方法在预测准确率上持平，但在特定干扰条件下仍展现出轻微优势，平均性能差距约 0.1 个百分点.

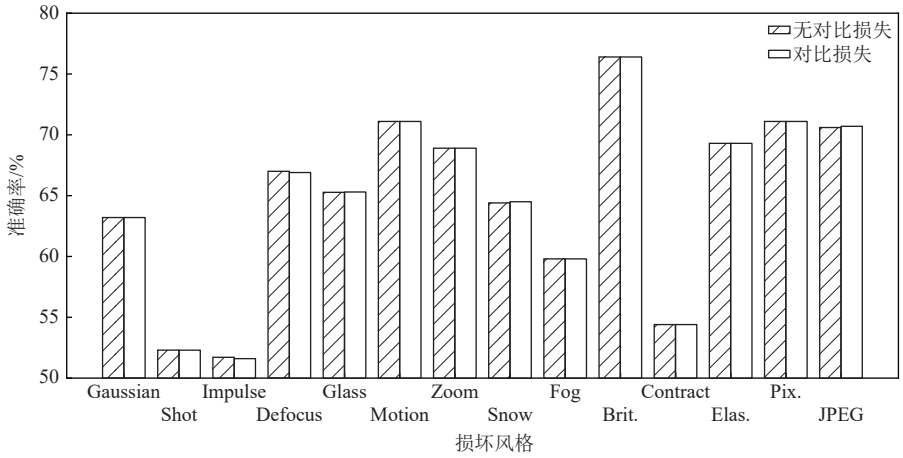
值得注意的是，在 JPEG 任务中，对比损失机制表现出优势(70.7% vs. 70.6%)，这表明对比损失对于处理压缩伪影这类结构化噪声具有独特效果. 而在



注：“Brit.”是“Brightness”的缩写，“Elas.”是“Elastic-transform”的缩写，“Pix.”是“Pixelate”的缩写

Fig. 3 Ablation experiment for low-quality view clipping and pseudo-resampling

图 3 低质视图裁剪与伪重采样的消融实验



注：“Brit.”是“Brightness”的缩写，“Elas.”是“Elastic-transform”的缩写，“Pix.”是“Pixelate”的缩写.

Fig. 4 Ablation experiment for contrast loss

图 4 对比损失的消融实验

Snow（64.5% vs. 64.4%）场景下，对比损失也展现出稳定的性能提升. 虽然整体性能差异较小，但对比损失在保持模型对各种干扰的稳定处理能力方面显示出其价值，特别是在处理结构化噪声和特定类型干扰时. 这种一致性表现验证了对比损失作为辅助训练目标的可靠性，能够在不影响主要任务性能的前提下，为模型提供额外的正则化效果.

3.3.2 参数敏感度分析

1）低质视图裁剪与伪重采样阈值敏感度分析. 为探究重采样过程中质量阈值 q 对模型性能的影响，本文系统分析了不同阈值设置下模型在各类损坏场景的表现.

实验数据如图 5 所示，随着质量阈值 q 从 0.1 逐步提升至 0.9，模型整体性能保持相对稳定，但在特定干扰类型下呈现出规律性变化. 在散焦模糊 Defocus 等空间失真类干扰下，较高阈值（q=0.7~0.9）反而带来轻微性能提升 (0.3~0.5 个百分点)，这表明适度放宽质量要求有助于保留更多有效训练样本. 值得注意的是，在结构化噪声场景中，不同阈值设置下的性能差异最小，反映出模型对这些干扰具有较好的鲁棒性. 特别地，当阈值设置过低 (q<0.3) 时，虽然在 Gaussian 等简单干扰下表现稳定，但在 Snow 等复杂干扰中出现了较大的性能波动，这可能是由于过度严格的样本筛选导致训练数据不足. 而当阈值设置过高 (q>0.5) 时，在极端干扰下的性能标准差明显增大 (达 0.5~0.6 个百分点).

综合分析表明，阈值 q 在 0.3~0.7 区间能取得最佳平衡，且在不同干扰类型间展现出最稳定的性能表现. 这一发现为实际应用中的参数选择提供了重要参考.

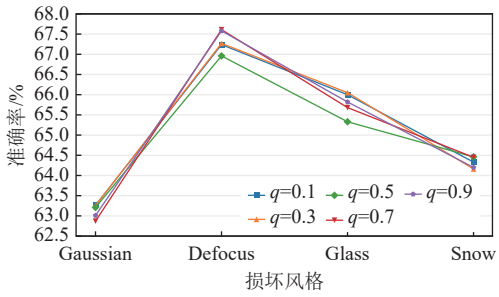


Fig. 5 Low-quality view clipping and pseudo-resampling threshold sensitivity analysis

图 5 低质视图裁剪与伪重采样阈值敏感度分析

2)时序对比自监督损失权重敏感度分析. 为探究对比损失权重 d 对模型性能的影响, 本文系统分析了权重系数从 0.1 到 0.9 范围内模型在各类损坏场景的表现. 实验数据如图 6 所示, 对比损失权重的变化对模型整体性能影响较小, 但在特定干扰类型下仍展现出一定的规律性特征. 当对比损失权重在 0.3~0.5 时, 模型在多数任务场景下达到最佳平衡状态. 值得注意的是, 在极端干扰条件下, 适当增大对比损失权重(d=0.7~0.9)可带来轻微性能提升, 这表明对比学习对这些特定噪声类型具有独特的正则化效果.

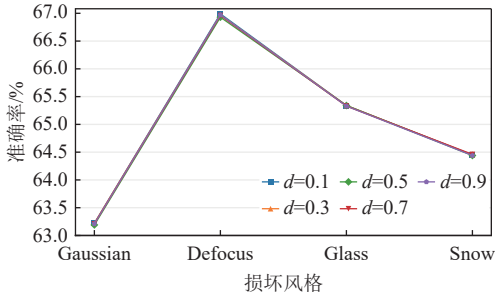


Fig. 6 Time series comparative self-supervised loss weights sensitivity analysis

图 6 时序对比自监督损失权重敏感度分析

而在结构化噪声场景中, 不同权重设置下的性能差异最小, 反映出模型对这些干扰具有较好的适应性. 特别地, 在 Glass 空间失真类干扰下, 模型性能对权重变化表现最为敏感, 当权重超过 0.5 时出现轻微的性能波动. 而在明暗条件干扰下, 模型展现出最强的稳定性, 说明对比损失对这些场景的优化效果较为有限. 综合分析表明, 对比损失权重为 0.3~0.5 能取得最佳平衡, 此时各项任务的平均标准差最小(约 0.5%). 这一发现为实际应用中的参数选择提供了重要参考: 虽然绝对性能差异不大, 但适度的对比损失权重能帮助模型在保持主要任务性能的同时, 增强对特定噪声类型的鲁棒性. 值得注意的是, 过高的对比损失权重(d>0.7)可能导致在部分空间失真任务中出现性能波动.

3)温度参数敏感度分析. 为探究对比学习中温度参数 (Temperature) 对模型性能的影响, 本文系统分析了温度系数 t 为 0.07~0.20 时模型在各类损坏场景的表现. 实验数据如图 7 所示, 温度参数的变化对模型整体性能影响较小, 表明模型对温度参数的选择具有较好的鲁棒性. 在温度参数从 0.07 增加到 0.20 的过程中, 各任务性能保持高度稳定, 平均波动范围小于 0.05 个百分点. 值得注意的是, 在 Glass 等空间失真类干扰下, 不同温度设置下的性能标准差相差较大, 这表明温度参数可能对这些特定干扰类型的特征对比学习产生轻微影响. 而在结构化噪声场景中, 模型展现出最强的稳定性. 特别地, 当温度参数设置为 0.10 时, 模型在多数任务场景下达到最佳平衡状态, 此时各项任务的平均标准差最小. 在极端噪声场景中, 温度参数的调整对性能影响最为有限, 这表明对比学习对这些噪声类型的特征表达具有较强的适应性.

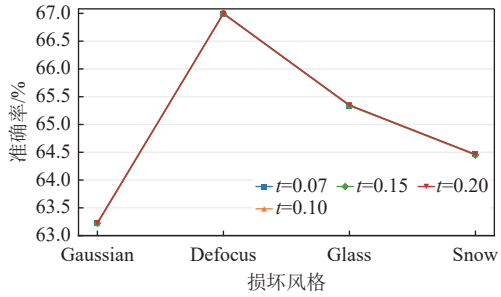


Fig. 7 Temperature parameters sensitivity analysis

图 7 温度参数敏感度分析

综合分析表明, 温度参数在 0.07~0.15 时都能取得良好效果, 模型性能对温度选择表现相对不敏感. 这一发现为实际应用中的参数调优提供了重要参考: 在保证温度参数处于合理范围内的前提下, 可以优先考虑其他对模型性能影响更大的超参数进行优化. 值得注意的是, 过低的温度参数 (t<0.07) 可能导致特征对比过于尖锐, 而过高的温度参数 (t>0.20) 则可能使对比学习过于平滑, 但在本实验的参数范围内均未观察到明显的性能下降.

4)跨模态对齐损失敏感度分析. 为探究跨模态对齐损失权重对模型性能的影响, 本文系统分析了权重系数从 1.0 到 0.2 范围内模型在 2 类最具代表性的损坏场景中的表现. 实验数据如图 8 所示, 可以看到权重系数的变化对模型整体性能影响较小, 表明模型对联合损失权重的敏感度较低、鲁棒性较强. 在

权重系数从 1.0 逐渐降低到 0.2 的过程中，模型性能保持高度稳定，Gaussian 场景准确率波动范围仅 0.01 个百分点（63.6%~63.59%），Shot 场景波动范围为 0.03 个百分点（51.79%~51.82%）.

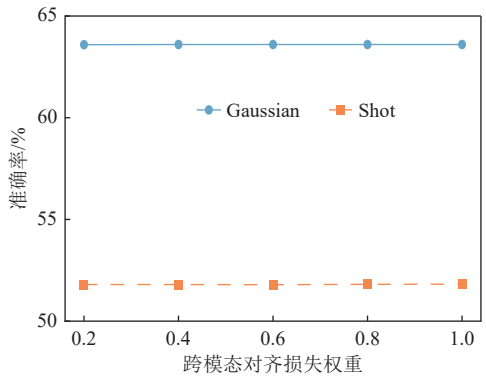


Fig. 8　Cross modal alignment loss

图 8　跨模态对齐损失

综合分析表明，联合权重系数在 0.2~1.0 都能取得良好效果，模型性能对联合损失权重变化的表现相对不敏感. 这一发现为实际应用中的参数调优提供了重要参考：在保证联合损失权重处于合理范围内的前提下，可以优先考虑其他对模型性能影响更大的超参数进行优化，标准差的稳定也进一步验证了该结论.

5）平衡损失权重敏感度分析. 为探究分布约束平衡损失权重对模型性能的影响，我们系统分析了权重系数从 0.2 到 0.5 范围内模型在 2 类典型噪声场景中的表现. 实验数据如图 9 所示，可以看到权重系数的变化对模型整体性能影响有限，Gaussian 场景下准确率波动范围为 0.13 个百分点（63.54%~63.67%），Defocus 场景波动为 0.17 个百分点（51.79%~51.96%）. 在权重系数从 0.2 增加到 0.5 的过程中，模型性能呈现平缓下降趋势.

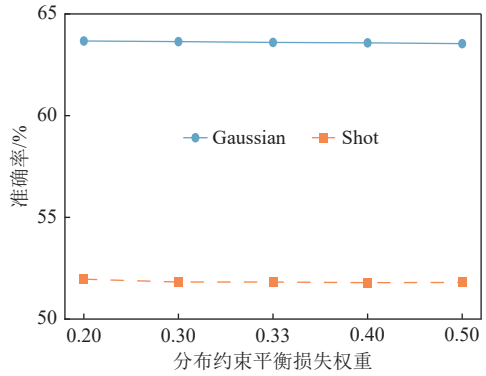


Fig. 9　Distribution constrains balance loss

图 9　分布约束平衡损失

综合分析表明，平衡损失权重系数在 0.2~0.5 时均能维持较好效果，其中权重系数 0.2 时取得最佳性能（Gaussian：63.67%，Shot：51.96%）. 模型对平衡损失权重变化表现较为稳健，最大性能波动不足 0.2 个百分点. 这一发现说明在实际应用中，可将平衡损失权重设置在 0.2~0.3 以获得最优性能.

6）信赖损失权重敏感度分析. 为探究信赖损失权重对模型性能的影响，本文系统分析了权重系数从 1.0 到 0.2 范围内模型在 2 类最具代表性的损坏场景中的表现. 实验数据如图 10 所示，可以看到权重系数的变化对模型整体性能影响较小，表明模型对信赖度损失权重的敏感度较低、鲁棒性较强. 在权重系数从 1.0 逐渐降低到 0.2 的过程中，模型性能保持高度稳定，平均波动范围小于 0.05 个百分点.

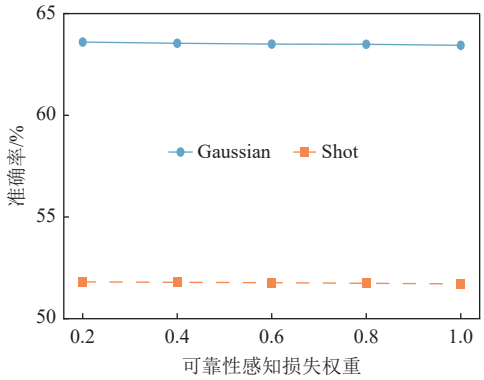


Fig. 10　Perceived loss of reliability

图 10　可靠性感知损失

综合分析表明，信赖度权重系数在 0.2~1.0 时能取得良好效果，模型性能对信赖损失权重变化的表现相对不敏感. 这一发现为实际应用中的参数调优提供了重要参考：在保证信赖损失权重处于合理范围内的前提下，可以优先考虑其他对模型性能影响更大的超参数进行优化.

4　结　　论

本文提出了一种基于跨模态协同机制的测试时自适应方法 CVPTA，解决了 MM-TTA 中一个非常现实且极具挑战的问题：多模态数据在实际应用时面对分布偏移问题无法实现信息的充分利用和模型的稳定运行. CVPTA 通过引入跨模态协同机制，强化了模态内部结构建模与模态间特征交互的鲁棒性；创新性地设计了低质量视图裁减和筛选模块，结合对比损失优化策略，从训练到测试的全流程确保模型在适应过程中具备稳定且可泛化的特征感知能力.

在此基础上,该方法基于视图不确定性集成多个视图的预测结果,通过一致性融合策略实现更稳健的模态感知与最终决策.在2个公开数据集上的大量实验,证明了CVPTA的有效性以及全流程多视图系统的可行性.

尽管CVPTA框架在多个数据集上取得了优异成绩,但是现实世界中TTA方法需要面对的挑战还不止于此.随着模态数量继续增加,基于多视图的全流程框架也将受到严峻的考验:跨模态协同机制需要处理更复杂的模态间关系,低质量视图裁减和筛选模块面临更高维度的计算压力,对比损失优化策略也需要适应更动态的特征空间.如何准确衡量持续复杂化的模态可靠性,在此基础上如何构建更加完善的多模态协同体系,以及在更加开放和复杂的任务中怎么保证模型的性能,都是亟待解决的问题.

作者贡献声明: 尹小龙提出了CVPTR方法流程和框架,完成实验和论文撰写;詹德川完善方法,给出了参数调整方面的建议;姜远指导论文写作,并规范了排版.

参 考 文 献

[1] Gama J, Žliobaitė I, Bifet A, et al. A survey on concept drift adaptation[J]. ACM Computing Surveys, 2014, 46(4): 1−37

[2] Moreno-Torres J G, Raeder T, Alaíz-Rodríguez R, et al. A unifying view on dataset shift in classification[J]. Pattern Recognition, 2012, 45(1): 521−530

[3] Wang Mei, Deng Weihong. Deep visual domain adaptation: A survey[J]. Neurocomputing, 2018, 312: 135−153

[4] Sun Yu, Wang Xiaolong, Liu Zhuang, et al. Test-time training with self-supervision for generalization under distribution shifts[C]//Proc of the 37th Int Conf on Machine Learning. Online, JMLR. org, 2020: 9229−9248

[5] Pan S J, Yang Qiang. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345−1359

[6] Zhou Kaiyang, Liu Ziwei, Qiao Yu, et al. Domain generalization: A survey[J/OL]. 2021[2022-08-08]. https://arxiv.org/abs/2103.02503

[7] Liang Jian, He Ran, Tan Tieniu. A comprehensive survey on test-time adaptation under distribution shifts[J]. International Journal of Computer Vision, 2023, 133: 31−64

[8] Wang Dequan, Shelhamer E, Liu Shaoteng, et al. Tent: Fully test-time adaptation by entropy minimization[C/OL]//Proc of the 9th In Int Conf on Learning Representations. Vienna: OpenReview. net, 2021 [2021-03-18]. https://openreview.net/forum?id=uXl3bZLkr3c

[9] Wilson G, Cook D J. A survey of unsupervised deep domain adaptation[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(5): 1−46

[10] Panagiotakopoulos T, Dovesi P L, Härenstam-Nielsen L, et al. Online domain adaptation for semantic segmentation in ever-changing conditions[C]// Proc of the 17th European Conf on Computer Vision (ECCV 2022). Berlin: Springer, 2022: 128−146

[11] Yang Tao, Zhou Shenglong, Wang Yuwang, et al. Test-time batch normalization[J/OL]. 2022[2022-05-20]. https://arxiv.org/abs/2205.10210.

[12] Park H, Hwang J, Mun S, et al. MedBN: Robust test-time adaptation against malicious mest mamples[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 5997−6007

[13] Lim H, Kim B, Choo J, et al. TTN: A domain-shift aware batch normalization in test-time adaptation[C/OL]// Proc of the 9th In Int Conf on Learning Representations. Kigali, OpenReview.net, 2023 [2023-02-18]. https://arxiv.org/abs/2302.05155

[14] Hong Junyuan, Lyu Lingjuan, Zhou Jiayu, et al. MECTA: Memory-economic continual test-time model adaptation[C/OL]// Proc of the 9th In Int Conf on Learning Representations. Kigali, OpenReview. net, 2023[2023-03-04]. https://openreview.net/pdf?id=N92hjSf5NNh

[15] Sójka D, Cygert S, Twardowski B, et al. AR-TTA: A simple method for real-world continual test-time adaptation[C/OL]//Proc of British Machine Vision Conf. Glasgow: BMVA. org, 2024[2024-10-14]. https://arxiv.org/pdf/2309.10109

[16] Yu Yongcan, Sheng Lijun, He Ran, et al. Benchmarking test-time adaptation against distribution shifts in image classification[J]. arXiv preprint, arXiv: 2307.03133, 2023

[17] Chen Dian, Wang Dequan, Darrell T, et al. Contrastive test-time adaptation[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 295−305

[18] Boudiaf M, Mueller R, Ayed I B, et al. Parameter-free online test-time adaptation[C]// /Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 8334−8343

[19] Zhou Zhi, Zhang Dingchu, Li Yufeng, et al. Robust test-time adaptation method for open-set recognition[J]. Journal of Software, 2024, 35(4): 1667−1681(in Chinese)
(周植, 张丁楚, 李宇峰, 等. 面向开集识别的稳健测试时适应方法. 软件学报, 2024, 35(4): 1667−1681)

[20] Baltrušaitis T, Ahuja C, Morency L. Multimodal machine learning: A survey and taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2): 423−443

[21] Radford A, Kim J, Hallacy C, et al. Learning transferable visual models from natural language supervision[C/OL]// Proc of Int Conf on Machine Learning. New York: ACM, 2021[2021-02-26]. https://arxiv.org/abs/2103.00020

[22] Lei Jixiang, Pernkopf F. Two-level test-time adaptation in multimodal learning[C/OL]//Proc of ICML 2024 Workshop on Foundation Models in the Wild. New York: ACM, 2024[2025-05-10]. https://icml.cc/virtual/2024/39426.

[23] Mao Tingting, Li Chang, Zhao Yuchang, et al. Online test-time adaptation for patient-independent seizure prediction[J]. IEEE Sensors Journal 2023, 23(19): 23133−23144

[24] Yuan Longhui, Binhui Xie, Shuang Li. Robust test-time adaptation in dynamic scenarios[C]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 15922−15932

[25] Varsavsky T, Orbes-Arteaga M, Sudre C H, et al. Test-time

unsupervised domain adaptation[J]. arXiv preprint, arXiv: 2010.01926, 2020

[26] Tsai Y Y , Chen F C , Chen A Y C, et al. GDA: Gneralized diffusion for robust test-time adaptation[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 23242−23251

[27] Zhang M, Levine S, Finn C, et al. Memo: Test time robustness via adaptation and augmentation[J]. Advances in Neural Information Processing Systems 2022, 35: 38629−38642

[28] Li Shuang, Yuan Longhui, Xie Binhui, et al. Generalized robust test-time adaptation in continuous dynamic scenarios[C/OL]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023[2023-05-08]. https://arxiv.org/abs/2303.13899

[29] Gong T, Kim Y, Lee T, et al. Sotta: Robust test-time adaptation on noisy data streams[J]. Advances in Neural Information Processing Systems 2023, 36: 14070−14093

[30] Park H, Hwang J, Mun S, et al. Medbn: Robust test-time adaptation against malicious test samples[C] // Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 5997−6007

[31] Wang Qin, Fink O, Gool L, et al. Continual test-time domain adaptation[C/OL] // Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022[2022-06-01]. https://openaccess.thecvf.com/content/CVPR2022/papers/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.pdf

[32] Liu Jiaming, Xu Ran, Yang Senqiao, et al. Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation [C]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 28653−28663

[33] Döbler M, Marsden R A, Yang Bin. Robust mean teacher for continual and gradual test-time adaptation[C]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7704−7714

[34] Song J, Lee J, Kweon I S, et al. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization[C]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11920−11929

[35] Tang Yushun, Chen Shuoshuo, Kan Zhehan, et al. Learning visual conditioning tokens to correct domain shift for fully test-time adaptation[J]. IEEE Transactions on Multimedia 2024, 10(1145): 6260−6269

[36] Karmanov A, Guan Dayan, Lu Shijian, et al. Efficient test-time adaptation of vision-language models[C] // Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. New York: ACM, 2024: 14162−14171

[37] Zhou Lihua, Ye Mao, Li Shuaifeng, et al. Bayesian test-time adaptation for vision-language models[J]. arXiv preprint, arXiv: 2503.09248, 2025

[38] Döbler M, Marsden R A, Raichle T, et al. A lost opportunity for vision-language models: A comparative study of online test-time adaptation for vision-language models[J]. arXiv preprint, arXiv: 2405.14977, 2024

[39] Zhang Yunbei, Mehra A, Hamm J. OT-VP: Optimal transport-guided visual prompting for test-time adaptation[C] // Proc of IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2025: 1122−1132

[40] Weijler L, Mirza M J, Sick L, et al. TTT-KD: Test-time training for 3D semantic segmentation through knowledge distillation from foundation models[J]. arXiv preprint, arXiv: 2403.11691, 2024

[41] Bar Y, Shaer S, Romano Y. Protected test-time adaptation via online entropy matching: A betting approach[J]. Advances in Neural Information Processing Systems 2024, 37: 85467−85499

[42] Dong Hao, Liu Moru, Zhou Kaiyang, et al. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models[J]. arXiv preprint, arXiv: 2501.18592, 2025

[43] Shu Manli, Nie Weili, Huang Dean, et al. Test-time prompt tuning for zero-shot generalization in vision-language models[J]. Advances in Neural Information Processing Systems 2022, 35: 14274−14289

[44] Xiao Zehao, Shen Jiayi, Derakhshani M M, et al. Any-shift prompting for generalization over distributions[C] // Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 13849−13860

[45] Sun Jiachen, Ibrahim M, Hall M, et al. VPA: Fully test-time visual prompt adaptation[C/OL]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 5796−5806

[46] Tsai Yun-Yun, Mao Chengzhi, Yang Junfeng. [J]. Advances in Neural Information Processing Systems, 2023, 36: 27897−27921

[47] Hassan J, Gani H, Hussein N, et al. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization[J]. Advances in Neural Information Processing Systems, 2023, 36: 80396−80413

[48] Wang Yanshuo, Cheraghian A, Hayder Z, et al. Backpropagation-free network for 3D test-time adaptation[C]//Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 23231−23241

[49] Osowiechi D, Noori M, Hakim G A V, et al. WATT: Weight average test-time adaptation of clip[J]. arXiv preprint, arXiv: 2406.13875, 2024

[50] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C/OL]//Proc of Int Conf on Machine Learning. New York: ACM, 2020[2020-07-01]. https://arxiv.org/abs/2002.05709

[51] Xiao Zehao, Zhen Xiantong, Liao Shengcai, et al. Energy-based test sample adaptation for domain generalization[J]. arXiv preprint, arXiv: 2302.11215, 2023

[52] Xiao Zehao, Zhen Xiantong, Shao Ling, et al. Learning to generalize across domains on single test samples[J]. arXiv preprint, arXiv: 2202.08045, 2022

[53] Li Yingming, Yang Ming, Zhang Zhongfei. A survey of multi-view representation learning[J]. IEEE Transactions on Knowledge and Data Engineering 2018, 31(10): 1863−1883

[54] Trosten D J, Løkse S, Jenssen R, et al. Reconsidering representation alignment for multi-view clustering[C]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1255−1265

[55] Zhang Qingheng, Sun Zequn, Hu Wei, et al. Multi-view knowledge graph embedding for entity alignment[J]. arXiv preprint, arXiv: 1906.02390, 2019

[56] Wang Ren, Sun Haoliang, Ma Yuling, et al. Metaviewer: Towards a

unified multi-view representation[C]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11590−11599

[57] Guo Chenfeng, Wu Dongrui. Canonical correlation analysis (CCA) based multi-view learning: An overview[J]. arXiv preprint, arXiv: 1907.01693, 2021

[58] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]// Proc of the 11th Annual Conf on Computational Learning Theory. New York: ACM, 1998: 92−100

[59] Chen Mansheng, Lin Jiaqi, Li Xianglong, et al. Representation learning in multi-view clustering: A literature review[J]. Data Science and Engineering 2022, 7(3): 225−241

[60] Qin Zequn, Chen Jingyu, Chen Chao, et al. Unifusion: Unified multi-view fusion Transformer for spatial-temporal representation in bird's-eye-view[C]// Proc of IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 8656−8665

[61] Ke Guanzhou, Chao Guoqing, Wang Xiaoli, et al. A clustering-guided contrastive fusion for multi-view representation learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(4): 2056−2069

[62] Cap Fuyuan, Chen Xiaohui. Multi-view attributed graph clustering with shared and specific representations[J]. Journal of Software, 2025, 36(3): 1254−1267

[63] Zhao Xingwang, Wang Shujun, Liu Xiaolin, et al. Joint spectral embedding for multi-view clustering via bipartite graph[J]. Journal of Software, 2024, 35(9): 4408−4424

[64] Siddiqui N, Tirupattur P, Shah M. DVANet: Disentangling view and action features for multi-view action recognition[J]. Proceedings of the AAAI Conf on Artificial Intelligence, 2024, 38(5): 4873−4881

[65] Bruce X B, Liu Yan, Zhang Xiang, et al. MMNet: Amodel-based multimodal network for human action recognition in RGB-D videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3522−3538

[66] He Kaiming, Fan Haoqi, Wu Yuxin, et al. Momentum contrast for unsupervised visual representation learning[C] //Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 9726−9735

[67] Khosla P, Teterwak P, Wang Chen, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems 2020, 33: 18661−18673

[68] Li Junnan, Zhou Pan, Xiong Caiming, et al. Prototypical contrastive learning of unsupervised representations[J]. arXiv preprint, arXiv: 2005.04966, 2021

[69] Jia Chao, Yang Yinfei, Xia Ye, et al. Scaling up visual and vision-language representation learning with noisy text supervision [C/OL]//Proc of Int Conf on Machine Learning. New York: ACM, 2021[2021-06-11]. https://arxiv.org/abs/2102.05918

[70] Grill J B, Strub F, Altché F, et al. Bootstrap your own latent-a new approach to self-supervised learning[C]// Proc of the 34th Int Conf on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2020: 21271−21284

[71] Chen Xinlei, He Kaiming. Exploring simple siamese representation learning[C]// Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 15745−15753

[72] Kay W, Carreira J, Simonyan K, et al. The Kinetics human action video dataset[J]. arXiv preprint, arXiv: 1705.06950, 2017

[73] Chen Honglie, Xie Weidi, Vedaldi A, et al. Vggsound: A large-scale audio-visual dataset[C] // Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 721−725

[74] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations[J]. arXiv preprint, arXiv: 1903.12261, 2019

[75] Wang Dequan, Shelhamer E, Liu Shaoteng, et al. Tent: Fully test-time adaptation by entropy minimization[C/OL]// Proc of the 9th Int Conf on Learning Representations. Vienna: OpenReview. net, 2021[2021-03-18]. https://openreview.net/forum?id=uXl3bZLkr3c

[76] Shin I, Tsai Y, Zhuang Bingbing, et al. MM-TTA: Multi-modal test-time adaptation for 3D semantic segmentation[C] // Proc of IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16907−16916

[77] Niu Shuaicheng, Wu Jiaxiang, Zhang Yifan, et al. Efficient test-time model adaptation without forgetting[C/OL]//Proc of Int Conf on Machine Learning, New York: ACM, 2022[2022-05-31]. https://arxiv.org/abs/2204.02610

[78] Niu Shuaicheng, Wu Jiaxiang, Zhang Yifan, et al. Towards stable test-time adaptation in dynamic wild world[J]. arXiv preprint, arXiv: 2302.12400, 2023

[79] Gong Yuan, Rouditchenko A, Liu A H, et al. Contrastive audio-visual masked autoencoder[J]. arXiv preprint, arXiv: 2210.07839, 2023

[80] Yang Mouxing, Li Yunfan, Zhang Changqing, et al. Test-time adaptation against multi-modal reliability bias[C/OL]// Proc of the 12th Int Conf on Learning Representations. Vienna: OpenReview. net, 2024[2024-01-16]. https://openreview.net/forum?id=TPZRq4FALB



Yin Xiaolong, born in 1998. PhD candidate. His main research interests include test-time adaptation and reinforcement learning.

尹小龙, 1998 年生. 博士研究生. 主要研究方向为测试时适应、强化学习.



Zhan Dechuan, born in 1982. PhD, professor. His main research interests include machine learning and data mining.

詹德川, 1982 年生. 博士, 教授. 主要研究方向为机器学习、数据挖掘.



Jiang Yuan, born in 1976. PhD, professor. Her main research interests include machine learning and data mining.

姜　远, 1976 年生. 博士, 教授. 主要研究方向为机器学习、数据挖掘.