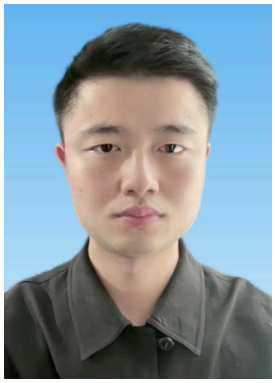
陈昊星

(+86) 153-9937-1933 · 蚂蚁集团 · 算法研究员 ·
hx.chen@hotmail.com · github · Google Scholar



我是陈昊星，毕业于南京大学，现为蚂蚁集团计算机视觉算法研究员，主要研究方向是多模态大模型、表示学习和数据受限下的机器学习，负责支付宝多模态内容风控相关业务，入职至今均获**最高绩效**，**晋升一次**。在人工智能顶级期刊和会议上发表论文 10 余篇，被引 300 余次，包括 NeurIPS/ICML/CVPR/ACM MM/ECCV/AAAI/SCIS 等，开源代码 600+ star，开源数据集下载超千次，相关研究成果在蚂蚁集团开发的国内首个一体化大模型安全解决方案“蚁天鉴”落地。

教育背景

2023.06	**南京大学** · 控制科学与智能工程 · 硕士
2020.09	综合排名第 1，南京大学优秀毕业生，南京大学优秀硕士论文，国家奖学金。
2020.06	**东南大学** · 测控技术与仪器 · 本科
2016.08	综合排名前 10%，推免至南京大学；校三好学生/东南大学优秀本科生党员；2019 美国大学生数学建模竞赛国际级一等奖(M)，2018 第八届中国教育机器人(ERCC)大赛国家级特等奖，2018 全国大学生数学建模竞赛省级一等奖等。

工作经历

蚂蚁集团

2022 年 05 月 – 至今

计算机视觉算法研究员(实习/正式) - 大安全 · 机器智能

负责安全域视觉算法业务与研究，参与蚁天鉴、光鉴等多个产品核心算法研究。

1. AIGC 生成

背景：解决安全域黑样本匮乏的问题，支持物体合成/文字篡改两大类任务，支持内部“光鉴”凭证类篡改/“蚁鉴”大模型测评等需求。

方案 & 成果：

- 构建了安全域 AIGC 生成平台，包含通用生成与文字篡改两大能力。通用生成能力包括人脸/物体生成/IP 生成等 10 余种能力；文字篡改包含去除文字/Copy&Paste/篡改等 10 余种能力。在通用生成方面，我们可合成 deepfake、直播间小物体等数据，为防御和检测算法的研发提供支持；在文字篡改方面，则助力“光鉴”通用篡改检测模型和内部 OCR 模型的训练与迭代。

技术探索与创新：

- 通用生成方向：开发了轻量化的动态和谐化算法 **HDNet@ACM MM Oral**，在减少 80% 参数量的同时提升了 20% 的性能。
- 文字编辑方向：研发了业内首个通用文字篡改模型 **DiffUTE@NeurIPS**，能够对任意带文字图像中任意语种的文字进行编辑，并保持文本风格与背景一致性。
- 可解释检测：构建了首个利用 MLLM 进行可解释文字篡改检测的基准数据集 **ETTD**，包含万条生成数据与真实数据。

2. AIGC 检测

背景：支付宝域内每日新增用户上传图像数据超千万，视频数据超百万，需要对 AIGC 生成内容进行鉴别以方便管控 AIGC 滥用风险。

方案 & 成果：

- 构建了由多模态生成模块和多模态理解模块组成的 AIGC 检测系统:多模态生成模块可覆盖近百种图像、视频的生成算法，可进行批量生成以促进多模态理解模块的训练，同时也可以利用生成样本来评测当前多模态

理解模块的性能，以攻促防。多模态理解模块接收图像、视频输入，它从时序不一致、空间不一致、语义不完善等多个角度进行分析，鉴别输入内容是否是生成内容。

- 目前，该服务已广泛应用于支付宝的短视频、直播、财富社区、兴趣社区等场景，日均调用量超过千万，实时监控和分析上传至平台的图像、音视频内容，保障蚂蚁内容信息的真实性和透明性，维护健康的交流环境。
- 蚂蚁 AIGC 安全检测系统也被整合进蚂蚁大模型安全一体化解决方案“蚁天鉴”，形成针对“AI 滥用风险防御”的产品方案。AIGC 图像/视频服务通过信通院评测，为国内首家单位。相关成果在世界人工智能大会/外滩大会展出，并已开展商业化。

技术与数据开放：

- AIGC 视频检测：构建了 AIGC 视频检测数据集 **GenVideo**，包含百万级真实视频和多种生成方法的视频，设计了针对生成视频时序和空间不一致性的 **DeMamba** 模型，相较于最先进算法，迁移性更强。
- AIGC 图像检测：构建了千万级 AIGC 图像数据集 **WildFake@AAAI**，探索不同生成方法训练的检测模型的迁移性。

3. 安全域图文基座

背景：传统**收集数据-训练模型-上线**的流程，支持需求的周期较长，无法实现对风险的快速响应。为了构建快速风险防控能力，基于十亿级安全域图文数据构建图文模型基座，并设计了零样本/小样本的分类/检测 Adapter 形式的快速布防能力。

方案 & 成果：

- 基于十亿级安全域图像数据，构建“数量-重复率-低质性-多样性”的图像质量筛选策略，筛选得到高质量图像；利用 Qwen/InternVL 等先进算法对图像进行 caption；基于 cluster 对图像进行 mask，并结合文本进行预训练得到安全域图文基座。
- 基于 LLM 对类别名做增强，即生成多维度描述来增强文本原型，对图像数据也做增强，通过一致性约束提升泛化性，训练得到 Adapter。

技术探索与创新：

- **Conditional Prototype Rectification：**构建基于训练集文本原型/视觉原型来在推理时根据输入图像生成不同的用于分类原型的策略，并利用无标签数据再对分类原型进行进一步细化以提升分类性能。

4. MLLM for 内容理解

背景：支付宝内富含直播、短视频、广告等多种场景的各种内容来源，需要对广告/短视频进行质量评估/内容理解，以助力推荐，优化用户体验。

方案 & 成果：

- 构建了以大模型兜底，小模型优先支持的技术体系。针对需求进行可行性分析，对于较难的任务利用大模型来提升对规则的理解，对于规则复杂、人审也难以理解的内容，构建了 CoT 制造-过滤-MPO 的训练流程，有效提升模型的推理能力。
- 在广告 AIGC 审核场景，审核自动化率由 0 提升至 90%；在短视频/直播场景，构建音视文基座，基于 QwenVL2，通过嫁接音频 encoder，实现三模态对齐结构，在短视频场景提升质量审核自动化率 0 至 30%，准确率 90%+。

部分科研经历

统一文字篡改/生成 2023 年 03 月 – 2023 年 06 月

DiffUTE: Universal Text Editing Diffusion Model @ NeurIPS 2023, CCF-A, 1st author

构建业内首个通用文字编辑模型 DiffUTE，该模型可以对任意图像中的文字进行编辑，保持文字风格/文字角度/文字位置以及背景的和谐性。

动态图像和谐化 2023 年 01 月 – 2023 年 05 月

Hierarchical Dynamic Image Harmonization @ ACM MM 2023 Oral, CCF-A, 1st author

基于前景物体和背景图像之间的语义相关性，在全局和局部两个维度进行动态调制学习，从而使合成图像更加的和谐。

对比学习视觉预训练改进 2022 年 12 月 – 2023 年 01 月

Model-Aware Contrastive Learning: Towards Escaping the Dilemmas @ ICML 2023, CCF-A, corresponding/co-1st author

为了解决 uniformity-tolerance 和 gradient reduction 困境，我们提出了一种模型感知对比学习策略，该策略的温度可根据对齐度的大小进行调整，从而反映出实例判别任务的基本可信度，从而使对比学习能够自适应地调整对硬否定的惩罚力度。

小样本学习 2023 年 01 月 – 2023 年 05 月

Sparse Spatial Transformers for Few-Shot Learning @ SCIS 2023, CCF-A, 1st author

针对全局特征和像素级特征可能都不能有效地进行小样本学习，全局特征会丢失局部信息，而像素素级特征会丢失图像的上下文信息。此外，普通的 embedding 无法生成任务特定的视觉表示。为此，提出了一种新颖的稀疏空间 Transformer 用于小样本学习，它通过基于 Transformer 的架构生成特定任务的原型。

实例分割扩散模型 2022 年 11 月 – 2022 年 12 月

DiffusionInst: Diffusion Model for Instance Segmentation @ ICASSP 2024, CCF-B, corresponding/2nd author

首个利用扩散模型进行实例分割的工作 DiffusionInst，它将实例表示为实例感知滤波器，并将实例分割表述为噪声到滤波器的去噪过程。

SAM4 图像和谐化 2023 年 07 月 – 2023 年 09 月

Segment Anything Model Meets Image Harmonization @ ICASSP 2024, CCF-B, 1st author

利用 SAM 生成语义先验，设计语义引导的正则化层，利用背景的语义和颜色信息来对前景特征进行对齐，从而实现更和谐的合成图像。

多视图聚类 2022 年 05 月 – 2022 年 07 月

Learning Latent Distangled Embeddings and Graphs for Multi-view Clustering @ PR 2024, CCF-B, 2nd author

多视图聚类，我们探索了多视图聚类中的解藕问题，通过构建解藕图来提升聚类效果。

视频文本模型高效迁移 2024 年 09 月 – 2024 年 11 月

Efficient Transfer Learning for Video-language Foundation Models, 1st author

提出了时空 Adapter 和一致性约束，有效增强 video-language model 的迁移能力，在多个小样本和迁移任务展现出了有效性。

ViM 训练策略 2024 年 05 月 – 2024 年 09 月

Stochastic Layer-Wise Shuffle: A Good Practice to Improve the Vision Mamba Training, corresponding/1st author

提出了 SLWS 方法，通过对 ViM 的 token 进行分层随机 shuffle，避免 ViM 过拟合，可将现有模型扩展至 ViM-L。

比赛经历

AFAC 金融文档验真 **2/717** 2023

检测各种凭证中是否出现文字篡改，并定位相关区域。基于安全域 AIGC 生成平台，在提供的训练集白图样本上进行多种方式篡改，并利用多种模型进行训练和 emsembel 以得到更好的模型。

ICDAR 文档篡改检测 **3/1267** 2023

检测各种文档中是否出现文字篡改进行分类，构建基于 self-blending 数据增强策略来训练模型，并通过自蒸馏策略来提升模型效果。