杨靖懿

WeChat 13082309660 yangjingyi@mail.ustc.edu.cn https://yjyddq.github.io

教育经历

本科: 大连海事大学(211)　信息科学技术学院 2018 – 2022

专业：电子信息科学与技术　排名:1/86(1%)　GPA:4.39/5.0 成绩单　英语水平: CET6

研究生(研二 to 研三): 中国科学技术大学(985/211)　电子工程与信息科学系 2022 – 至今

专业：信息与通信工程　排名: 4/247(1.5%)　GPA: 3.96/4.3 成绩单

奖项荣誉

两次获得优秀学生一等奖学金(2%)、一次三等奖学金(10%) 2020,2021,2022

大连市优秀毕业生 2022

竞赛专项奖学金 2021

MCM/ICM Honorable Mention 2021

第十一届全国大学生数学竞赛（非数学类）辽宁省三等奖 2019

大连市第二十八届大学生数学竞赛理工类一等奖 2019

研究经历

- 人脸反欺骗(Face Anti-spoofing) - 一篇ECAI (CCF-B Oral)会议论文发表，AAAI25 UnderReview
- 视频理解/动作识别 - ICLR25 UnderReview
- 生物体特征识别 - 项目

研究兴趣

- (图像/视频)视觉文本/多模态学习
- 视频理解/生成，生成模型
- 多媒体内容安全，AI生成/伪造内容检测

人脸反欺骗检测 (Face Anti-Spoofing, FAS)

会议论文(一作): Jingyi Yang, Zitong Yu, Xiuming Ni, Jia He, Hui Li. Generalized Face Anti-spoofing via Finer Domain Partition and Disentangling Liveness-irrelevant Factors (ECAI CCF-B Oral) 2024 2022 – 2023

研究动机：人脸防欺骗检测系统是帮助人脸识别系统抵御欺骗攻击的一种防御系统。FAS系统追求安全性，算法的设计以泛化能力，鲁棒性为目的。目前，该领域的主导研究趋势域泛化(Domain Generalization，DG)，DG的假设贴近实际中存在数据域/攻击类型在训练中不可见的场景。之前的大多数工作集中在研究学习域不变的人脸活性表征。根据数据集进行域划分，并采用Adversarial feature learning/Meta learning学习数据集不变的表征。但以数据集作为域划分比较粗略。

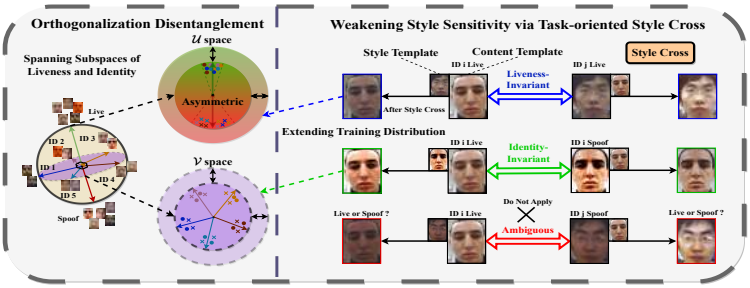

Figure 1: Motivation.

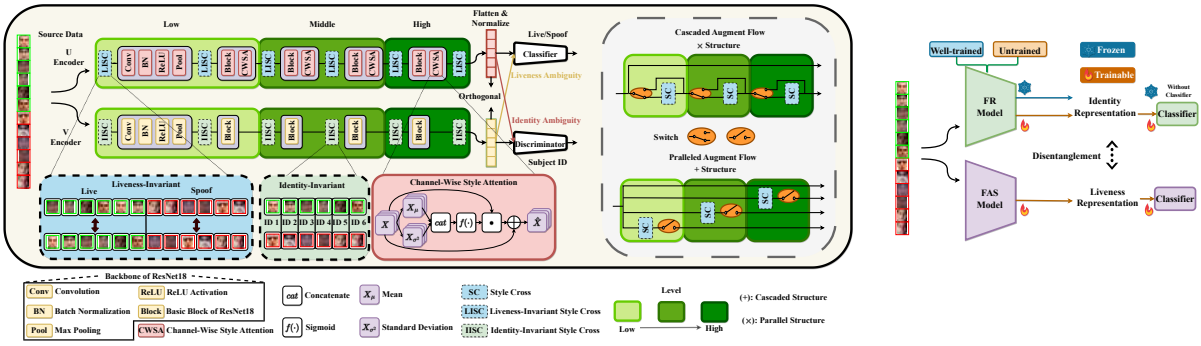

Figure 2: Architecture.

方法：以数据集作为域划分比较粗略，因为同一个数据集中的人脸样本不一致因素很多，如身份、光照、相机分辨率。而我们提出以身份作为更细致的域划分，缩小了域的概念，减小同一个域中不一致因素的影响，同时解耦活性和身份表征(特征的正交化)，学习身份不变的表征。优势是当训练数据集个数少时，甚至只有一个源数据集时，我们的框架也可以工作，而以数据集作为域划分的算法是失效的(*LOO设置，多个数据集作为源域，一个数据集作为目标域)。此外，我们的框架具有很强的可扩展性，可以借用已经提前训练好的人脸识别网络辅助活体检测网络进行表征解耦，Fig 2。

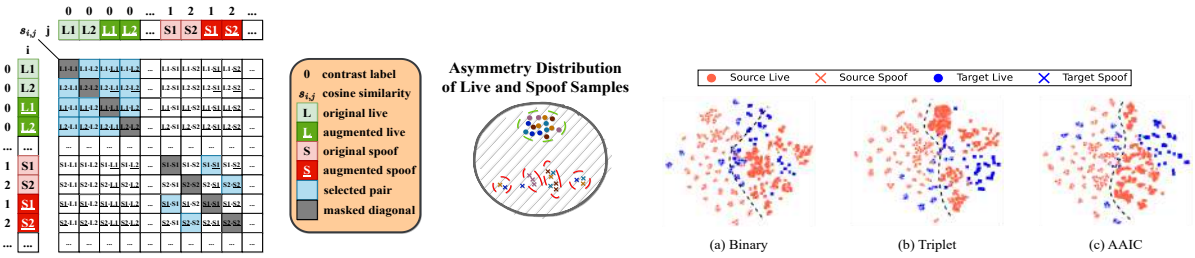

Figure 3: Asymmetric Augmented Instance Contrast.

独特的设计：采用了轻量化(只在训练时有计算开销，推理时失活)关注到类的风格增强组件以扩大训练样本分布(通常所说的风格指均值方差——特征的统计量)，串联结构/并联结构的风格增强结构。并且考虑到真实人脸和欺骗人脸在现实中出现的概率不平衡，对此设计了非对称的对比损失以贴近真实世界中的真实/欺骗样本分布。我们的方法在实际场景下具有一定的实用性。Demo

视频理解/动作识别 (Video Understanding/Action Recognition)

(ICLR25 UnderReview) 会议论文(一作)：Jingyi Yang, Zitong Yu, Xiuming Ni, Jia He, Hui Li. Kronecker Mask and Interpretive Prompts are Language-Action Video Learners 2023 – 2024

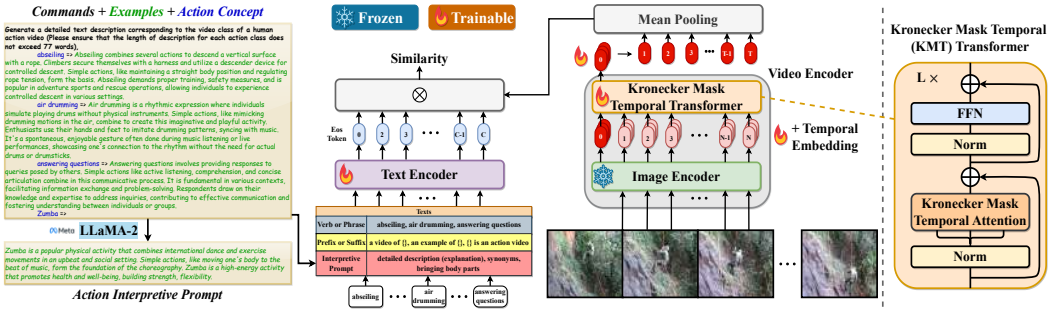

Figure 4: Framework of our video learners.

研究动机：大规模Visual-Language Models (VLMs)展现出极强的泛化性能和zero-shot能力，但将VLMs推广到视频理解/动作识别领域时，如何将VLMs中静态视觉物体与具象名词之间的对齐调整为动态动作过程和抽象动词之间的对齐有待解决。同时，我们侧重于探究时空建模的可解释性，给模型引入时空结构归纳偏差。

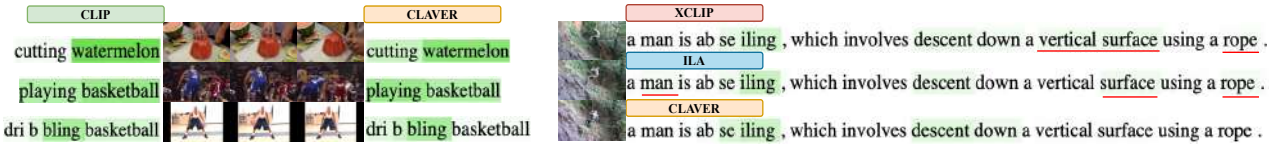

Figure 5: Word Importance of CLIP and video learners.

方法：模型架构如Fig 4所示，我们将联合注意力分解为空间和时间维度的注意力，设计了一种新颖的Kronecker Mask时间注意力，相比于之前的方法，增加了时间维度的感受野，削弱时空同质化的影响，以一种新视角揭示空间注意力和时间注意力的区别和内在关联，即大多数时空注意力都可以视为Kronecker Mask注意力的特例，如Fig 6。并设计了Kronecker Mask Causal Temporal Attention以缓解自注意力矩阵的低秩问题。

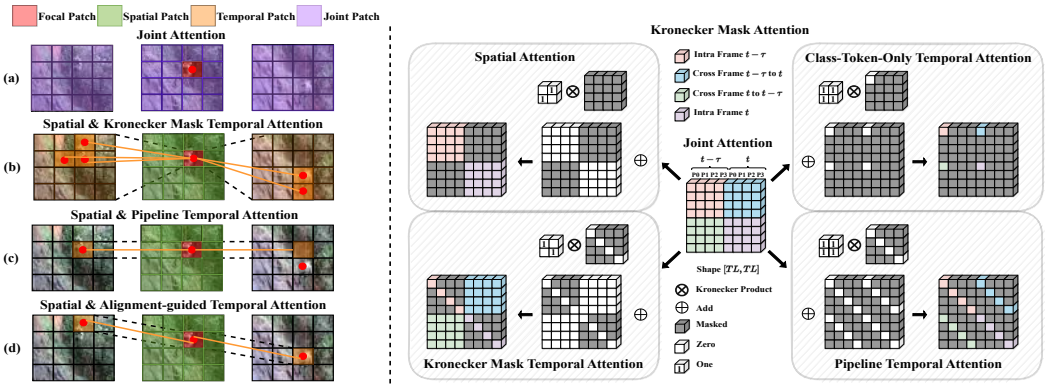

Figure 6: Kronecker Mask Attention.

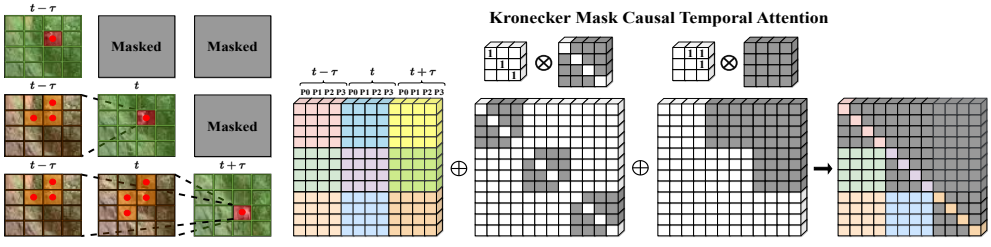

Figure 7: Kronecker Mask Causal Temporal Attention.

借助LLMs实现对动作概念或短语进行自动化解释性扩写(包含对动作的步骤分解，复杂动作由哪些简单动作组成，动作涉及到身体哪些部位的运动)，增加训练文本内容的多样性，同时增加了推理阶段使用文本的灵活性(单词-短语-句子级别)，有助于模型捕捉视频中的动作内容并重点根据动词关注动作发生区域，而不是根据名词关注静态物体。

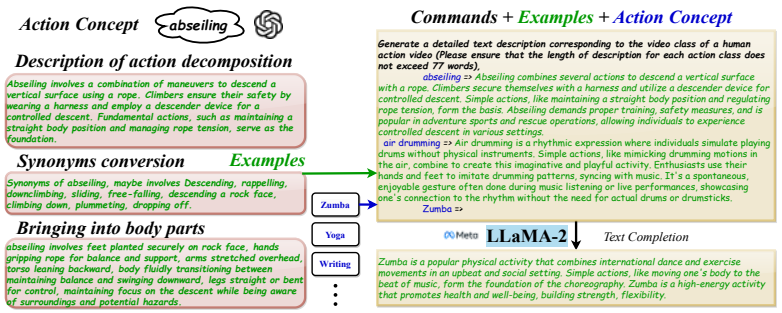

Figure 8: Interpretive Prompt.

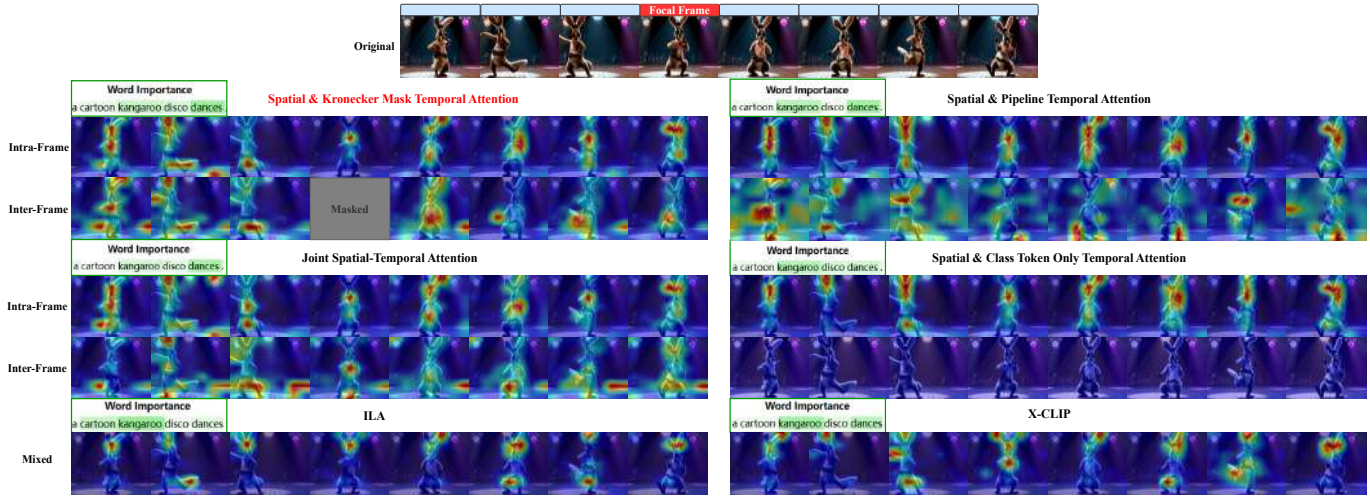

Figure 9: Spatiotemporal attention map.

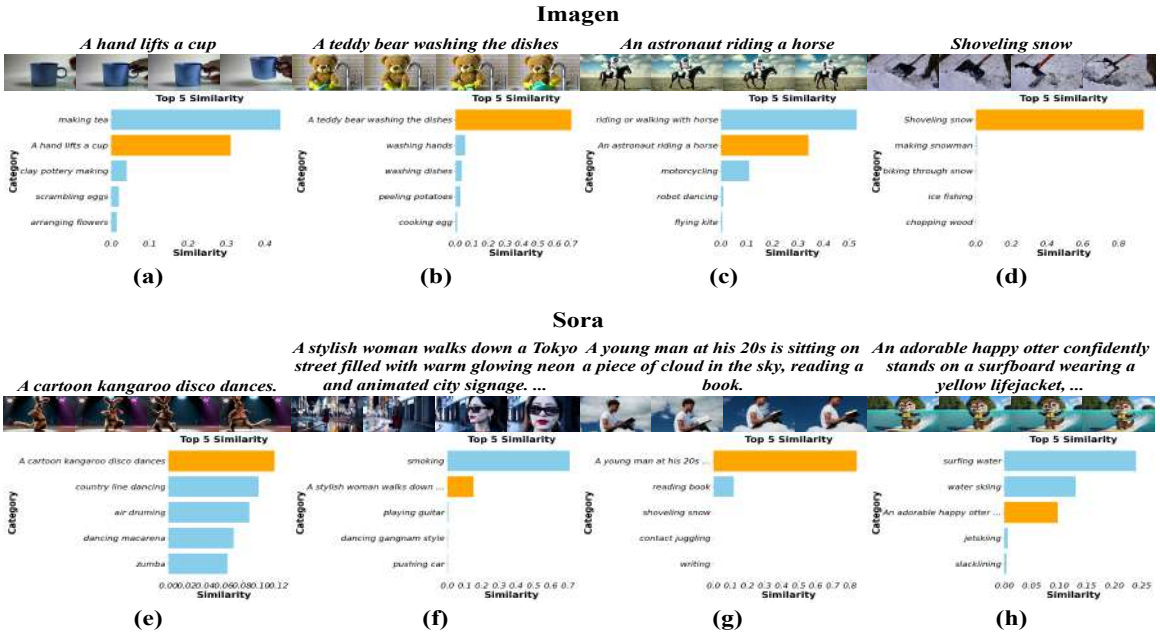

Figure 10: Synthetic video testing.

时空同质化是指，当随机打乱token的排列顺序时，模型视觉表征与对应的文本表征的相似度以及单词重要性并

没有很大变化，这似乎是不太符合逻辑的。就是说，当破坏时空数据的时空结构时，那么对应的视觉语义也应受到明显的影响，否则说明模型似乎并没有学习到时空结构关系，可学习的位置编码可能只是性能导向的，因为他们的优化就是有数据驱动的梯度决定的。我们设计的KMT和KMCT能够缓解时空同质化。

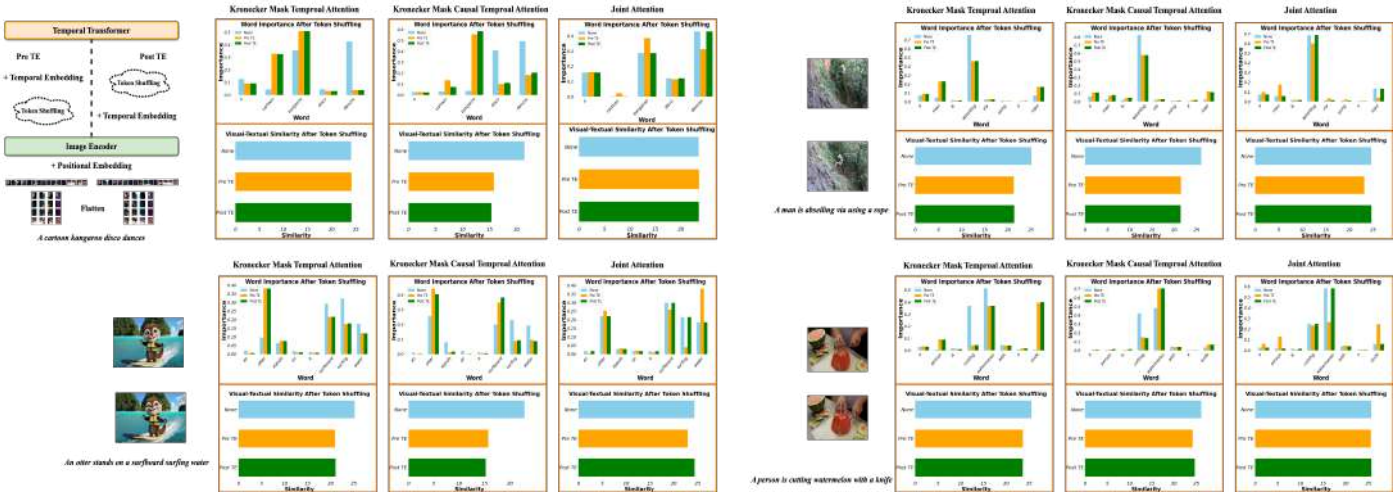

Figure 11: Spatiotemporal homogenization study 1.

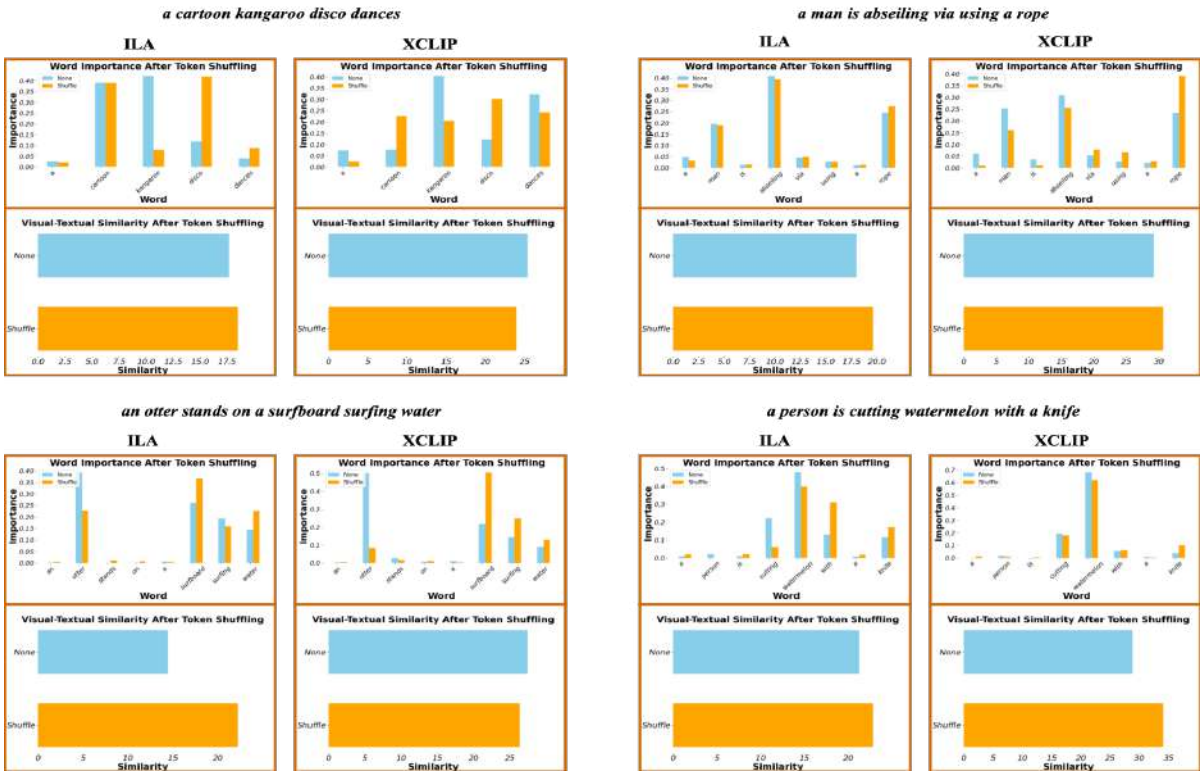

Figure 12: Spatiotemporal homogenization study 2.

基于视频的人脸防欺骗检测 (Video-Based Face Anti-Spoofing)

(AAAI25 UnderReview) 会议论文(一作)：Jingyi Yang, Zitong Yu, Xiuming Ni, Jia He, Hui Li. G^2V^2former: Graph Guided Video Vision Transformer for Face Anti-Spoofing 2023 – 2024

研究动机：传统的FAS以提取光度特征为主，结合静态(光度信息)和动态(运动信息)角度挖掘一段欺骗人脸视频欺骗痕迹的探索尚且不足，人脸地标(landmarks)可以辅助捕捉面部运动信息。眼睛周围，嘴巴，眉毛部位蕴含丰富的表情信息、运动信息。

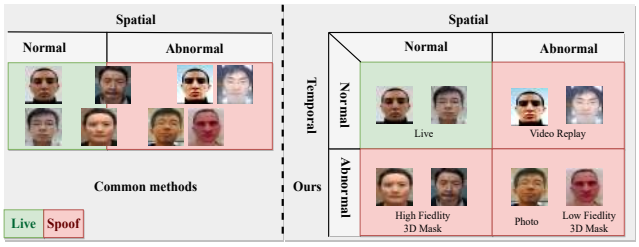

Figure 13: Combining Spatiotemporal perspectives.

解决方法：采用TimeSformer作为视觉流的骨干网络(将Pipeline temporal attention替换为Kronecker mask temporal attention)，采用装配了时间注意力的Graphormer抽取人脸地标中的运动信息。利用时空图中的时间注意力引导视觉流的时间注意力(graph-guided vision temporal attention)，Fig 14。我们在TimeSformer和Graphormer中使用的是一种新颖的Kronecker时间注意力来进行时空建模。

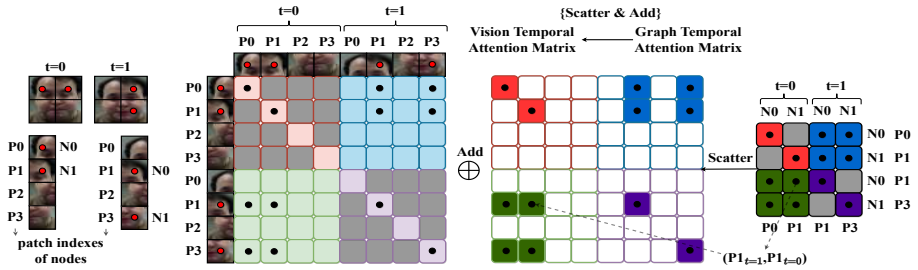

Figure 14: Graph Guided Vision Temporal Attention.

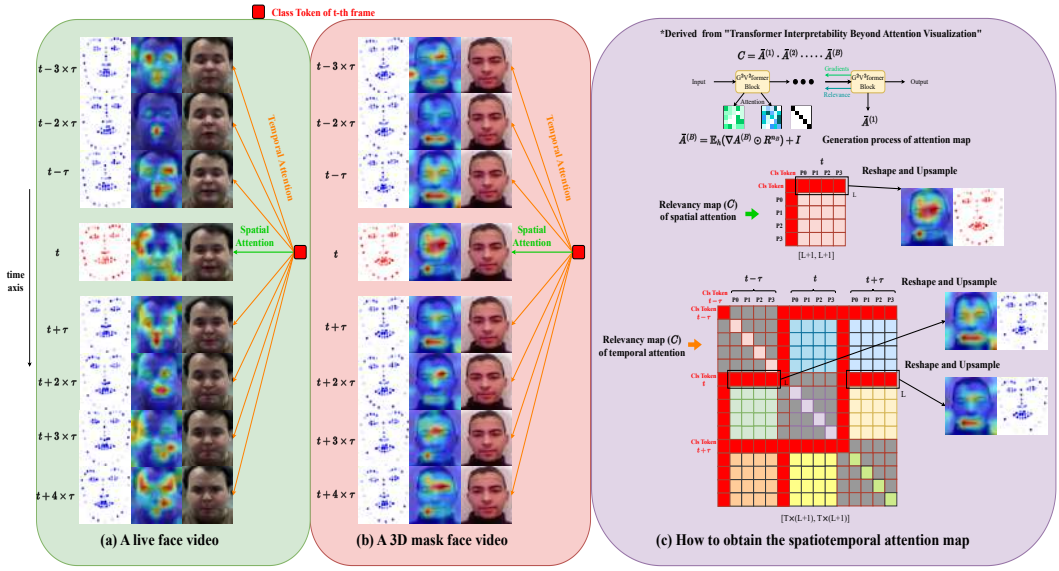

Figure 15: Spatiotemporal attention map.

面向地铁等刷掌应用场景的无接触式手掌识别(企业，安徽清新互联tsinglink)

2023 – 2024

工作内容：为地铁场景无接触刷掌系统提供了一套解决方案。

方法：手掌识别一般分为两个阶段: 1) 手掌检测+分割，2) 手掌身份识别。现有的一些基于传统方法的手掌/掌纹识别研究有一定的局限性，他们需要手掌尽可能平放并保持展平状态，摆放方向固定，否则会在第1)步失败从而影响识别性能。他们大多依赖学术界所使用的、经过较为严格采样和对齐方法的手掌数据集，采样的数据质量较高。但实际场景中手掌的摆放、朝向，伸展程度比较随意。考虑到传统手掌检测、分割方法难以应对实际掌纹采样环境，利用Segement Anything Model (SAM)可以对收集到的实际场景中(pose, illumination noisy)的手掌进行标注，我们分别尝试了不同的标注区域，包括掌纹部分、手掌+手指部分以测试手掌的哪些区域更有利于身份识别。然后利用标注好的数据微调一个YOLO模型，用YOLO实现手掌感兴趣部位的检测。然后将检测到的感兴趣部位送入一个手掌识别网络进行训练(1不进行任何旋转处理，直接输入一个神经网络模型，2将手掌的方向旋转至同一方向/对齐，再由传统方法处理，这种方法过于繁琐且准确性不高)。

(a) Acadamic Datasets

(b) Real World Samples

Figure 16: Data Gap.