

# 深度伪造人脸图像及视频的检测方法研究

**倪蓉蓉教授团队**

**汇报人：陈瑜**

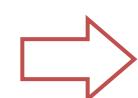
2025-3-21

北京交通大学 信息科学研究所  
“数字媒体信息处理” 科技部重点领域创新团队

<http://mepro.bjtu.edu.cn>



# 深度伪造的取证及主动防御方法



## 深度伪造取证的研究背景



研究现状及主要方法



AI深度伪造的取证方法

- 基于桥接样本对齐的深度伪造人脸图像检测
- 基于非关键音素-视素区域VA相关性学习的Deepfake检测



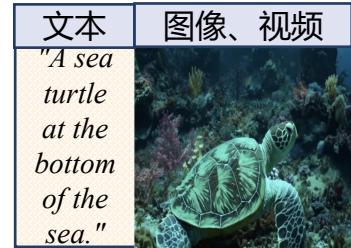
北京交通大学

BEIJING JIAOTONG UNIVERSITY

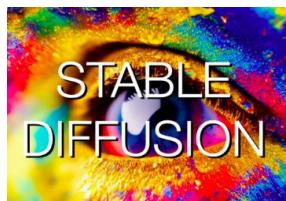


## 多媒体内容的取证

- 大数据时代，多媒体信息丰富、传播速度快
- 图像、视频等视觉媒体的优点
  - 信息量大、感受直观、影响广泛



- 然而，图像修改、视频编辑等工具使用便捷。





## 眼见不再为实，深度伪造真假难辨

□ Real or fake (2019年的伪造水平)





## 深度伪造的现象和危害

**深度伪造诞生**  
Reddit出现大量深度伪造图像。



十一月

**“杨幂版”黄蓉**  
明星换脸进行视频篡改。



三月

**英国女皇假视频**  
篡改女皇表情动作及唇部运动。



四月

**影视明星换脸视频**  
将“蜘蛛侠”的面部替换成“钢铁侠”身份上。



八月

**AIGC的逐渐发展**  
特朗普被捕以及梅西拥抱球迷等照片引起热议。



四月

2023

**伪造趋向“平民化”**  
**爆发式增长**



AI诈骗常用手法  
——通过AI技术换脸骗人  
筛选受骗人群  
个人信息  
视频通话

AI换脸  
个人信息  
视频通话

“拟音” + “拟脸”进行诈骗。

2018

**伪造奥巴马演讲视频**  
引起美国国防部高度重视。



四月

2019

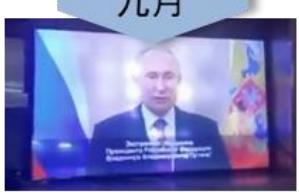
**扎克伯格假视频**  
新闻真假难辨。



六月

2020

**伪造普京讲话视频**  
在俄媒体传播，造成不利的政治影响。



九月

2021

**泽连斯基假视频**  
Deepfake参与战争，总统被伪造投降视频。



三月

**韩国N号房AI换脸**  
Deepfake换脸制造色情视频，并在社交媒体广泛传播。



八月

2024

近日，“N号房”丑闻重现引起韩国社会恐慌。“DeepFake”（深度伪造技术），即利用AI换脸制造色情视频，并通过社交媒体平台传播的性犯罪行为被大量曝光。据外媒报道，最近一周，韩国警方发现了大量与学校、医院、军队相关的涉案社交媒体群组，参与用户多达 22 万人。经初步调查，受害者 80% 为女性。  
资料显示，2021 年至 2023 年，韩国共有 527 名深度伪造淫秽影像受害者报案，未成年人占比达 59.3%，随着技术发展，使用深度伪造技术制造虚假视频更加方便，每款加害者中未成年人占比明显上升。(综合重庆广电第 1 频、中国新闻网)



## 大模型进一步提升了伪造的真实感



街头跳舞图像

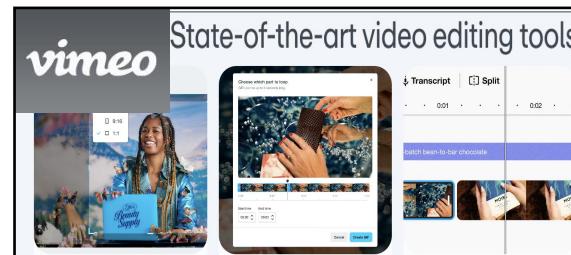
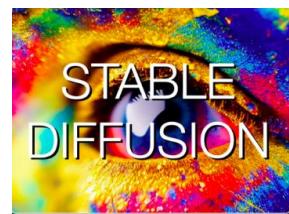


海龟的海底世界



森林高燃骑行

公开模型使伪造“零门槛”，伪造内容规模迅速扩大。



因此，为了确保信息的真实性和安全性，深度伪造检测具有重要意义。



## 深度伪造流程及研究任务



# 深度伪造的取证及主动防御方法



- 深度伪造取证的研究背景
- 研究现状及主要方法
- AI深度伪造的取证方法

- 基于桥接样本对齐的深度伪造人脸图像检测
- 基于非关键音素-视素区域VA相关性学习的Deepfake检测



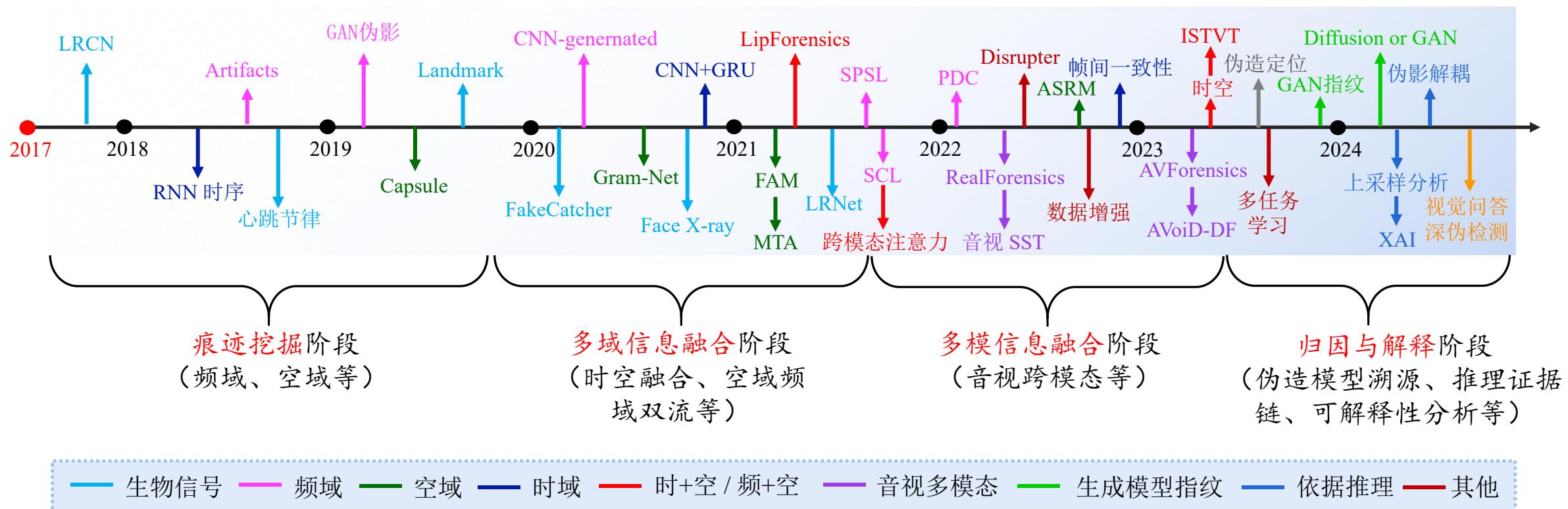
北京交通大学  
BEIJING JIAOTONG UNIVERSITY



## 深度伪造检测的发展历程

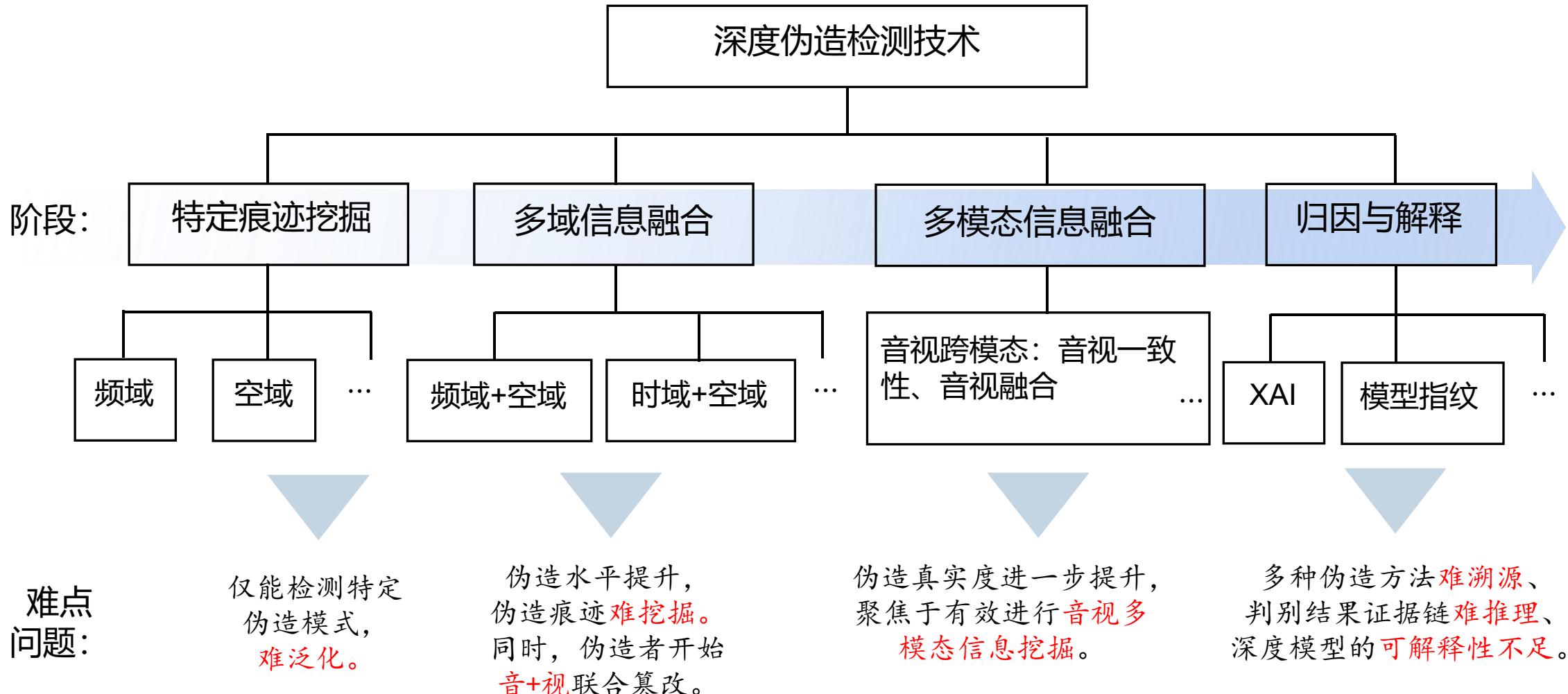
### 时间线

- 深度伪造出现：2017年，名为“deepfakes”的深度伪造图像发布。
- 伪造检测发展：痕迹挖掘 → 多域信息融合 → 多模信息融合。





## 现有方法分类及难点

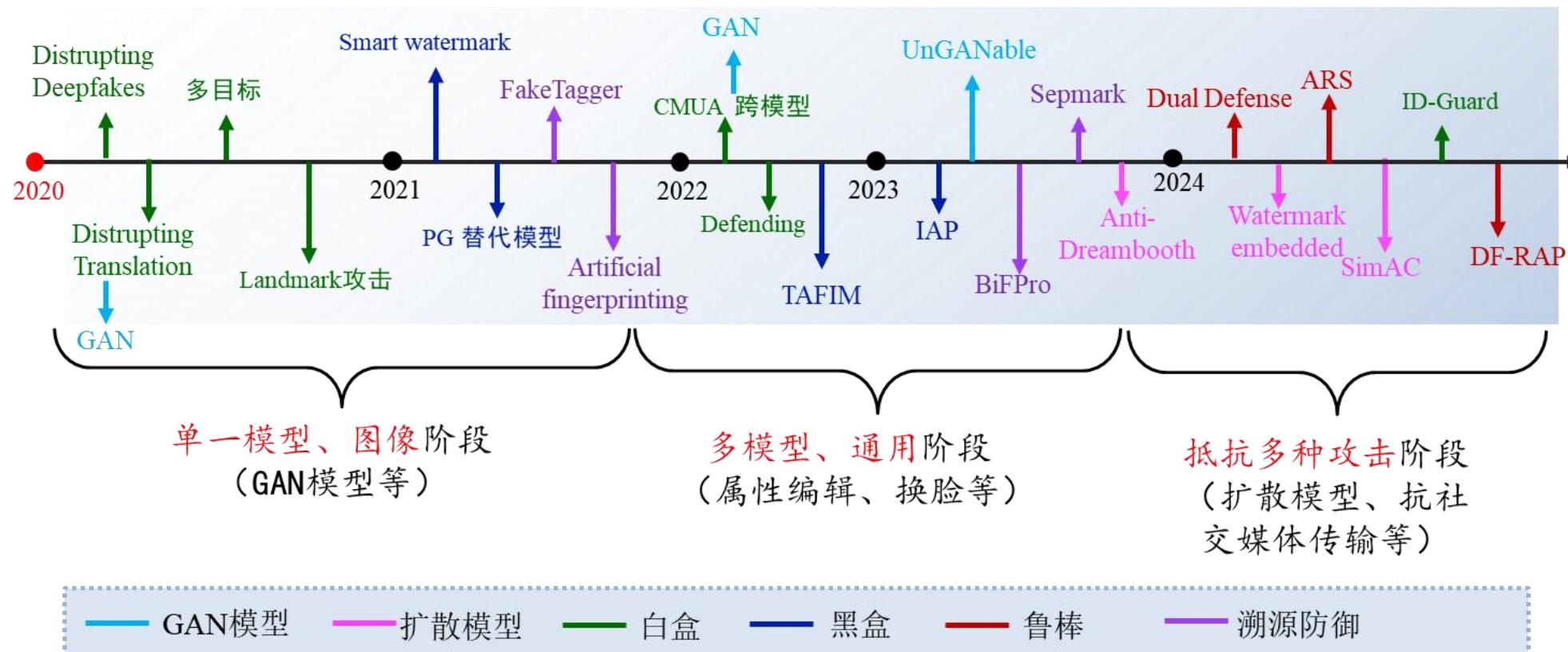




## 深度伪造主动防御的发展历程

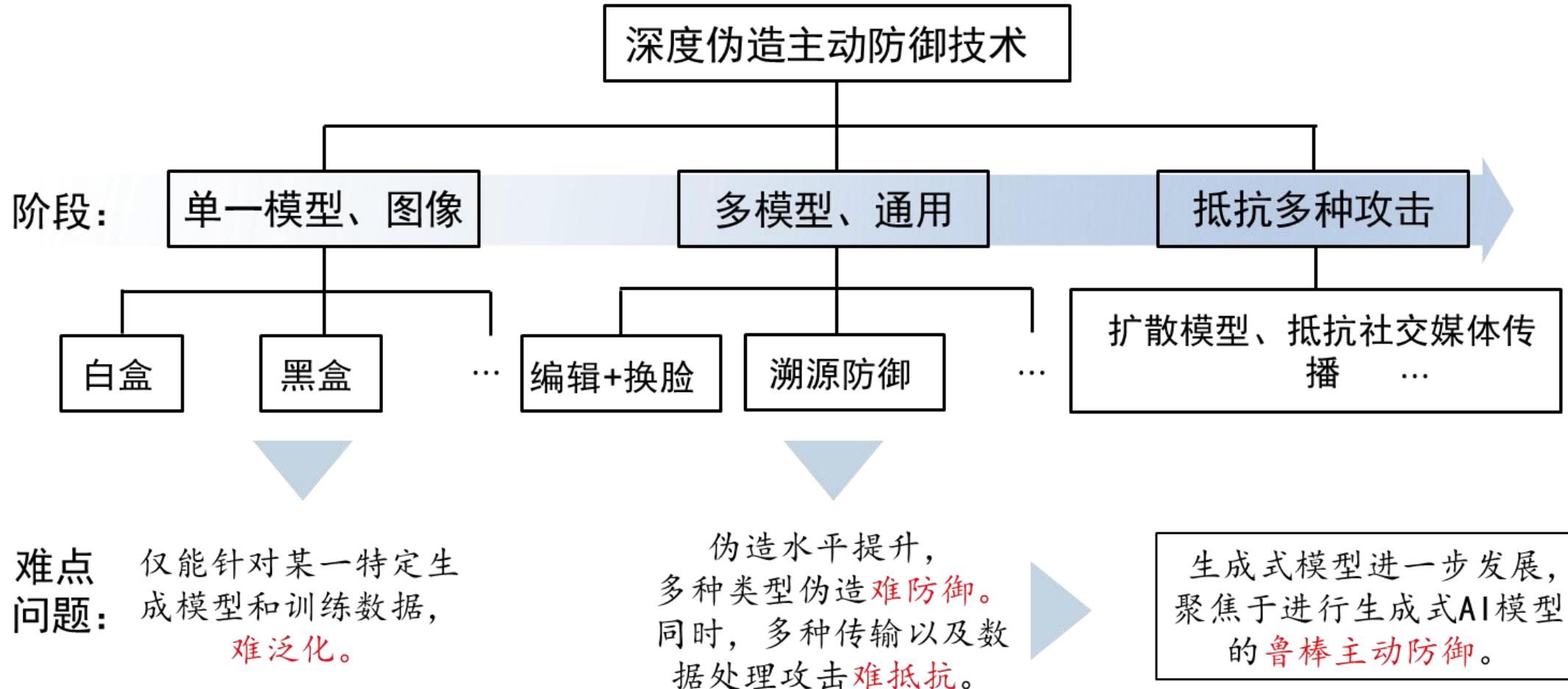
### 时间线

- 主动防御一词出现：2020年，研究者提出“Disrupting deepfakes”一词，意指中断伪造。
- 主动防御发展：单一模型、图像→多模型、通用→抵抗多种攻击。





## 现有方法分类及难点



# 深度伪造的取证及主动防御方法



- 深度伪造取证的研究背景
- 研究现状及主要方法
- **深度伪造的取证方法**

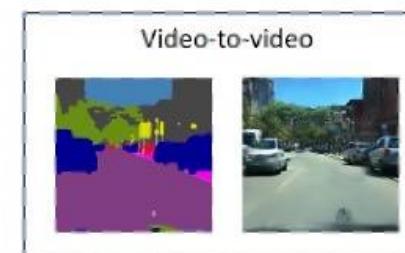
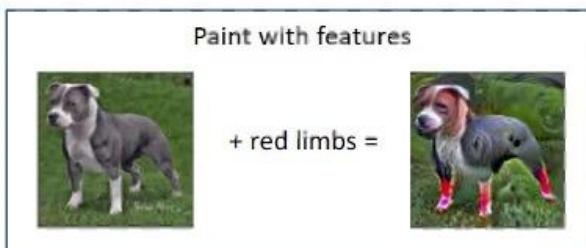
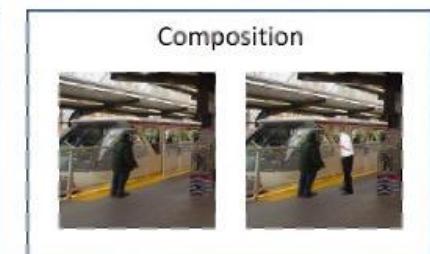
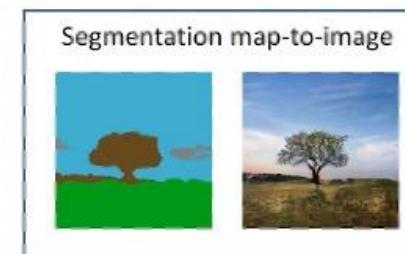
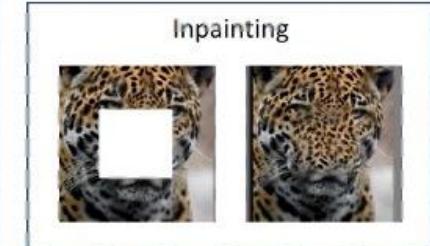
- 基于桥接样本对齐的深度伪造人脸图像检测
- 基于非关键音素-视素区域VA相关性学习的Deepfake检测



北京交通大学  
BEIJING JIAOTONG UNIVERSITY



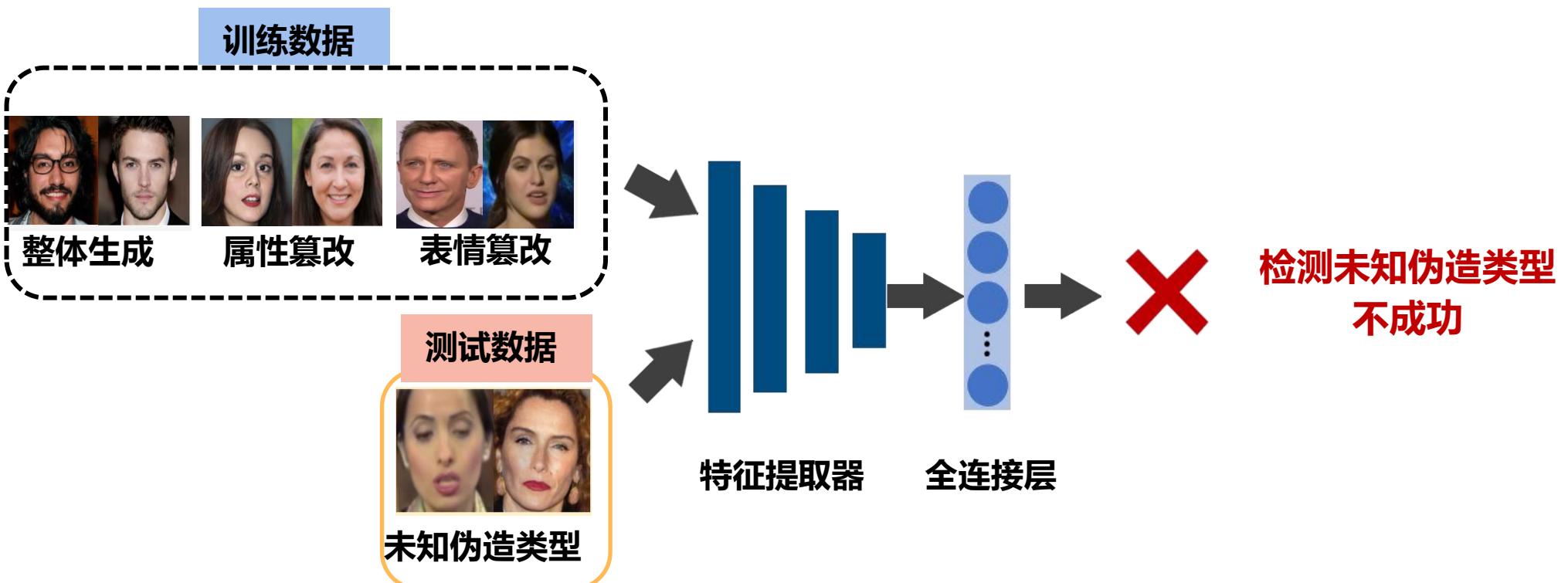
## AI生成（合成/伪造）手段越来越多



# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 研究动机

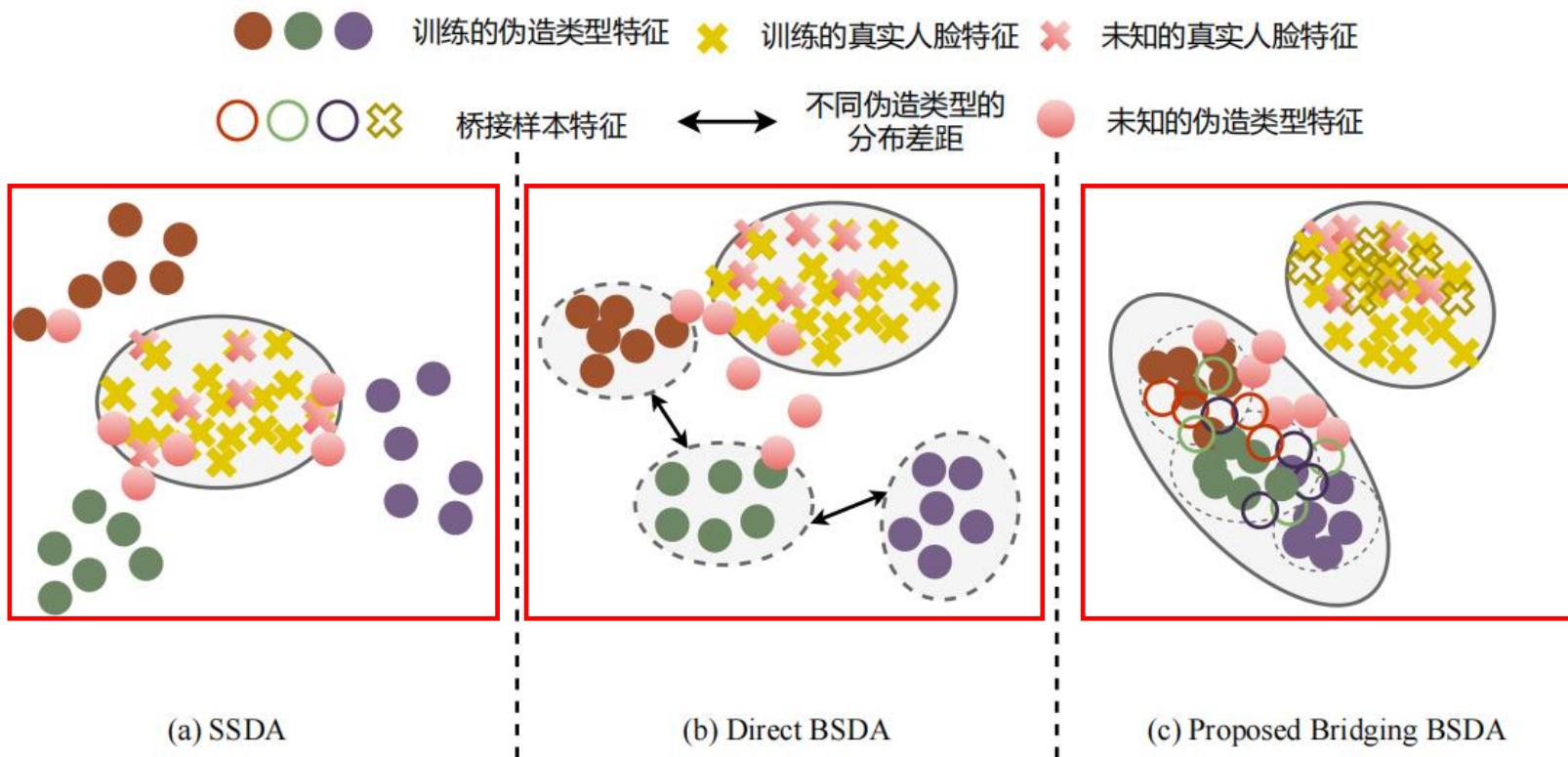
- 面向未知伪造类型，检测方法泛化性能进一步下降



# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 研究动机

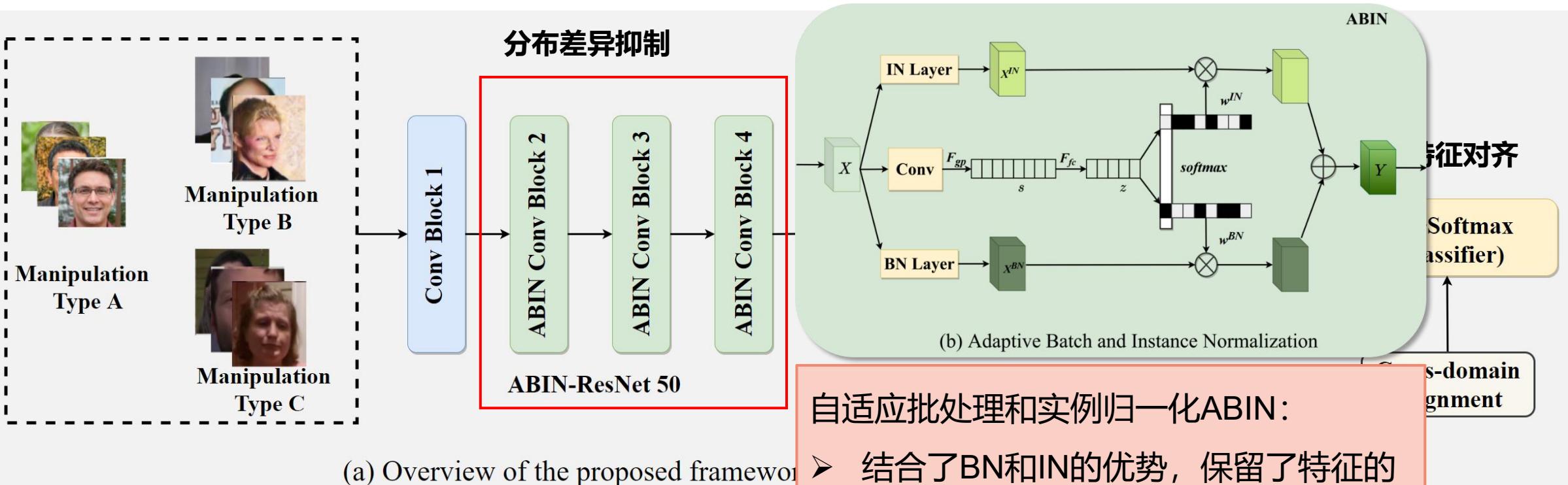
- 现有检测方法存在的问题：不同伪造类型之间存在**显著的分布差异**，现有方法难以对齐伪造特征表示



- **单边跨域对齐：**只对齐真实样本，难以得到强大的伪造特征表示
- **双边跨域对齐：**直接对齐真实和伪造样本，但伪造特征分布差异过大，对齐性能有限
- **桥接双边跨域对齐：**生成桥接样本来填补特征分布差异，对齐真实和伪造样本，提升对齐性能

# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

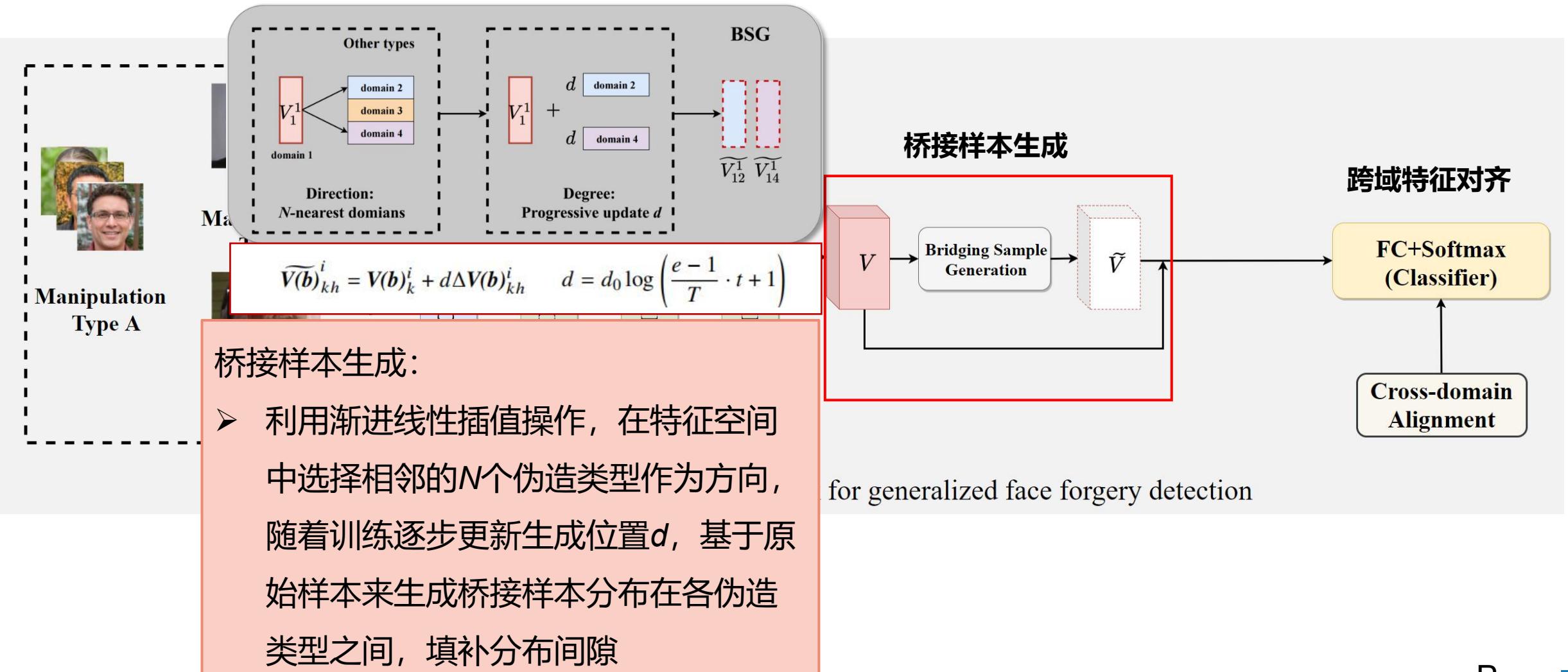
## 检测框架



自适应批处理和实例归一化ABIN：  
➤ 结合了BN和IN的优势，保留了特征的  
判别能力，同时也能初步抑制不同伪  
造类型间的特征差异

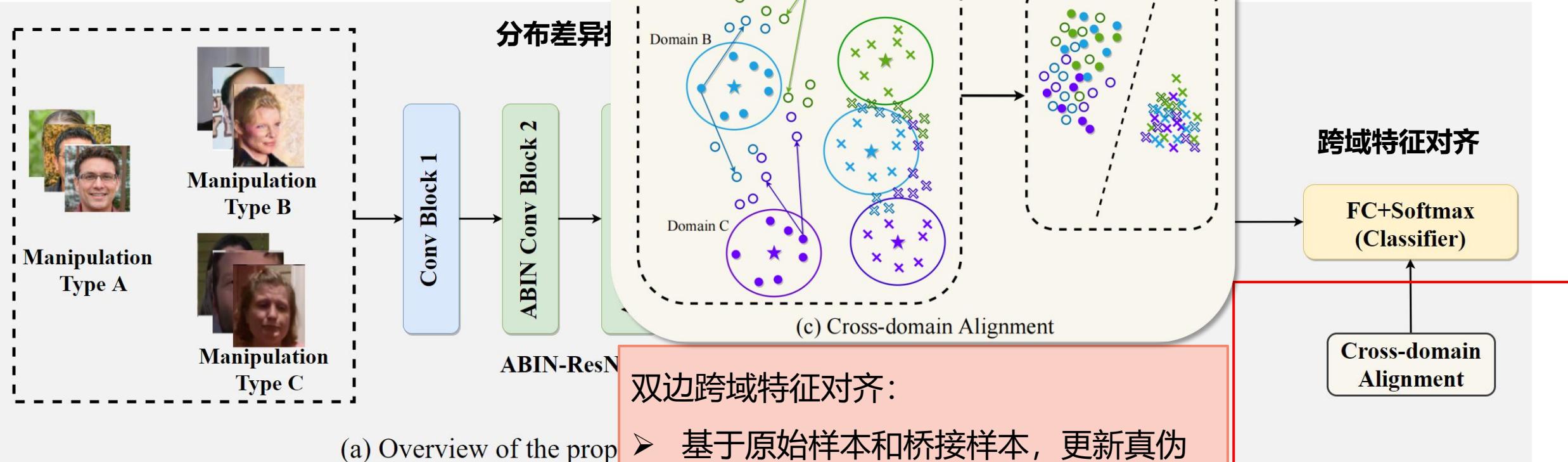
# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 检测框架



# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 检测框架



双边跨域特征对齐：

- 基于原始样本和桥接样本，更新真伪样本的全局域质心，通过缩小全局域质心与样本的距离，调整数据分布，得到通用的特征表征

# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 实验数据集

数据集	伪造技术	伪造样本收集方式	伪造样本数量	真实样本收集方式	真实样本数量	图像大小
EFS (Entire Face Synthesis)	StyleGAN <sup>[4]</sup>	公开	5000	FFHQ	5000	1024 × 1024
	StyleGAN2 <sup>[5]</sup>		5000	FFHQ	5000	
	StyleSwin <sup>[7]</sup>		5000	CelebA-HQ	5000	
FEM (Facial Expression Manipulation)	Face2Face <sup>[26]</sup>	FF++ (c23, c40)	10000	FF++	10000	256 × 256
	FReeNet <sup>[9]</sup>	个人生成	10000	RaFD	10000	
	DGN <sup>[11]</sup>		10000	VoxCeleb1	10000	
FAM (Facial Attribute Manipulation)	AttGAN <sup>[14]</sup>	个人生成	20000	LFW	20000	128 × 128
	StarGAN-V2 <sup>[16]</sup>		10000	FFHQ	10000	256 × 256
	InterFaceGAN <sup>[18]</sup>		5000	CelebA-HQ	5000	1024 × 1024
FIM (Face Identity Manipulation)	FaceSwap <sup>[19]</sup>	FF++ (c23, c40)	10000	FF++	10000	256 × 256
	DeepFakes <sup>[109]</sup>	DFDC	10000	DFDC	10000	
	ADDNet <sup>[122]</sup>	WildDeepfake	10000	WildDeepfake	10000	

# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 实验结果

### 未知伪造技术实验方案及结果

协议	训练技术	测试技术	协议	训练技术	测试技术
N1	StyleGAN2	StyleGAN	N3	StyleGAN	StyleSwin
	StyleSwin			StyleGAN2	
	FReeNet			Face2Face	
	DGN			DGN	
	AttGAN			AttGAN	
	InterFaceGAN			StarGAN-V2	
	DeepFakes			FaceSwap	
	ADDNet			ADDNet	
N2	StyleGAN	StyleGAN2	N4	StyleGAN	StyleSwin
	StyleSwin			StyleGAN2	
	Face2Face			Face2Face	
	FReeNet			FReeNet	
	InterFaceGAN			AttGAN	
	StarGAN-V2			StarGAN-V2	
	FaceSwap			FaceSwap	
	DeepFakes			DeepFakes	

方法	N1		N2		N3		N4		平均值	
	ACC	AUC								
MaDD <sup>[52]</sup>	91.36	92.17	91.47	93.26	89.79	91.15	89.63	91.95	90.56	92.13
FrePGAN <sup>[60]</sup>	92.27	93.59	92.93	93.57	90.09	91.91	91.21	92.36	91.63	92.86
PCL <sup>[65]</sup>	94.27	96.17	95.06	96.41	92.11	94.54	93.09	95.07	93.63	95.55
PEL <sup>[61]</sup>	94.18	95.83	94.78	95.12	91.97	94.06	91.77	93.19	93.18	94.55
DABN <sup>[118]</sup>	92.36	94.17	93.47	94.26	91.79	92.15	91.63	92.95	92.31	93.38
LTW <sup>[71]</sup>	91.67	92.39	91.74	93.56	90.81	92.37	90.67	92.12	91.22	92.61
FDFL <sup>[58]</sup>	93.18	95.83	94.78	96.12	92.07	94.57	91.97	95.19	93.00	95.43
DAM <sup>[119]</sup>	94.41	96.54	95.79	97.91	92.77	95.82	93.53	95.48	94.13	96.44
本章方法	95.86	97.89	96.57	97.81	94.13	96.91	94.64	97.78	95.30	97.60

# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 实验结果

### 未知伪造类型检测实验结果

方法	训练于其它三种伪造类型									
	EFS		FEM		FAM		FIM		平均值	
	ACC	AUC								
MaDD <sup>[52]</sup>	80.39	82.74	81.34	83.73	82.07	84.29	79.21	82.09	80.75	83.21
FrePGAN <sup>[60]</sup>	81.72	83.04	82.05	84.34	83.19	84.64	80.79	83.89	81.94	83.98
PCL <sup>[65]</sup>	83.62	86.15	84.81	87.35	84.94	86.51	82.81	84.81	84.05	86.21
PEL <sup>[61]</sup>	82.48	85.08	83.67	86.72	83.89	85.06	81.12	83.86	82.79	85.18
DABN <sup>[118]</sup>	81.26	84.83	82.57	84.15	81.21	83.73	78.05	80.61	81.77	84.08
LTW <sup>[71]</sup>	81.07	83.01	81.74	83.87	82.16	84.89	80.84	82.39	81.45	83.54
FDFL <sup>[58]</sup>	82.56	85.07	83.09	86.91	84.77	86.52	82.23	83.24	83.41	85.44
DAM <sup>[119]</sup>	83.78	86.79	84.69	87.33	85.47	86.76	83.96	85.57	84.48	86.61
本章方法	<b>86.97</b>	<b>89.43</b>	<b>87.89</b>	<b>89.64</b>	<b>87.81</b>	<b>89.82</b>	<b>84.97</b>	<b>86.67</b>	<b>86.91</b>	<b>88.89</b>

### 鲁棒实验

测试	原始准确率	JPEG 压缩			调整大小			中值滤波			高斯噪声		
		90	85	80	512	224	64	3×3	5×5	7×7	0.4	0.7	1.0
EFS	<b>86.97</b>	85.19	83.61	81.27	84.82	85.07	81.26	84.63	85.37	83.38	85.12	84.81	81.36
FEM	<b>87.89</b>	85.71	83.31	82.16	83.84	84.76	80.74	86.61	85.49	83.83	85.16	84.67	83.29
FAM	<b>87.81</b>	86.51	85.83	82.93	84.67	86.27	82.36	87.14	86.38	85.77	86.59	85.79	84.52
FIM	<b>84.97</b>	82.78	81.01	80.06	80.95	82.87	79.88	82.81	81.21	80.62	81.87	80.73	79.86

# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 实验结果

### 消融实验

BN	IN	ABIN	SSDA	BSDA	BSG	FIM
✓	-	-	-	-	-	<b>70.09</b>
-	✓	-	-	-	-	<b>72.02</b>
-	-	✓	-	-	-	<b>77.87</b>
-	-	✓	✓	-	-	<b>77.92</b>
-	-	✓	-	✓	-	<b>81.79</b>
-	-	✓	-	✓	✓	<b>84.97</b>

N	Protocol N4			$d_0$	FIM		Residual Stage (RS)	FIM	
	ACC	AUC			ACC	AUC		ACC	AUC
Nearest	1	92.86	95.83	0.1	82.72	84.79	RS-1-2	78.33	80.12
	2	<u>94.64</u>	<u>97.78</u>	0.5	83.18	85.31		81.67	83.96
	3	<b>94.67</b>	<b>97.81</b>	1.0	<b>84.97</b>	<b>86.67</b>		<b>84.97</b>	<b>86.67</b>
Random	1	89.36	91.69	2.0	84.51	86.19	RS-1-2-3-4	83.06	85.21
	2	91.24	93.71	3	93.19	95.37			

(a) 桥接方向 N 参数设置研究

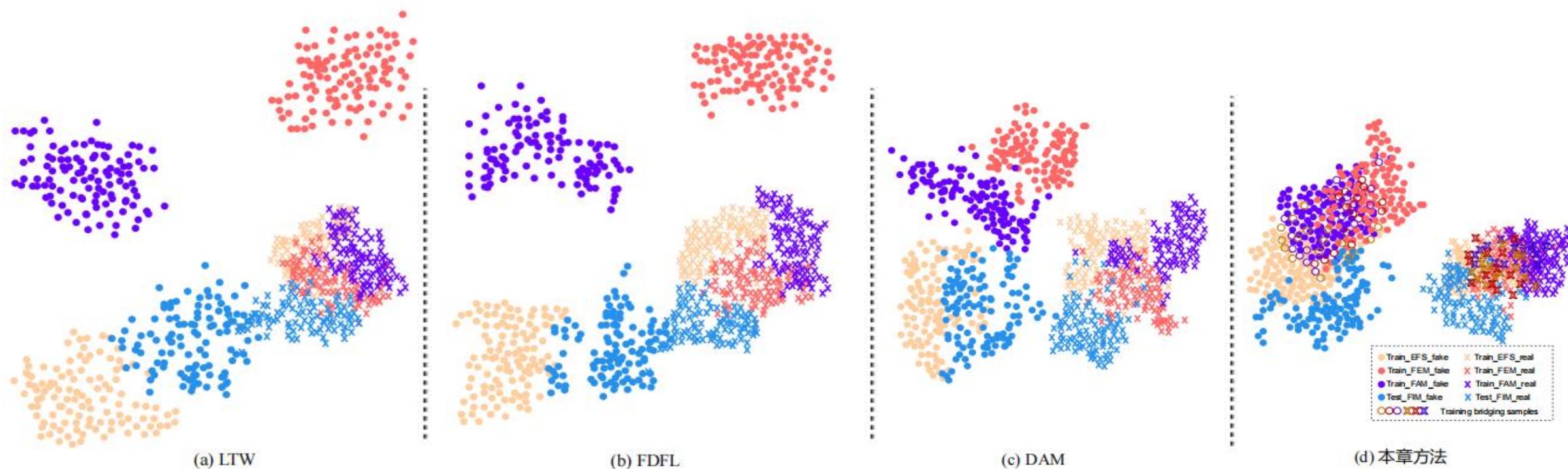
(b) 桥接程度初始值  $d_0$  参数设置研究

(c) ABIN 嵌入位置研究

# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

## 实验结果

### 特征分布可视化



# 工作1：基于桥接样本对齐的深度伪造人脸图像检测

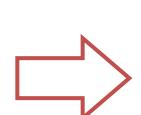
## ○ 总结

- 提出面向未知空域伪造人脸类型的检测新思路
- 采用了抑制、桥接和对齐三个关键步骤以减小不同伪造类型间特征分布的巨大差异，突破了未知伪造类型图像的检测难题

## ○ 对应成果

- Yang Yu, Rongrong Ni, Siyuan Yang, Yao Zhao, Alex C. Kot. Narrowing Domain Gaps with Bridging Samples for Generalized Face Forgery Detection. *IEEE Transactions on Multimedia (TMM)*. 2023.

# 深度伪造的取证及主动防御方法



- 深度伪造取证的研究背景
- 研究现状及主要方法
- **AI深度伪造的取证方法**

- 基于桥接样本对齐的深度伪造人脸图像检测
- **基于非关键音素-视素区域VA相关性学习的  
Deepfake检测**

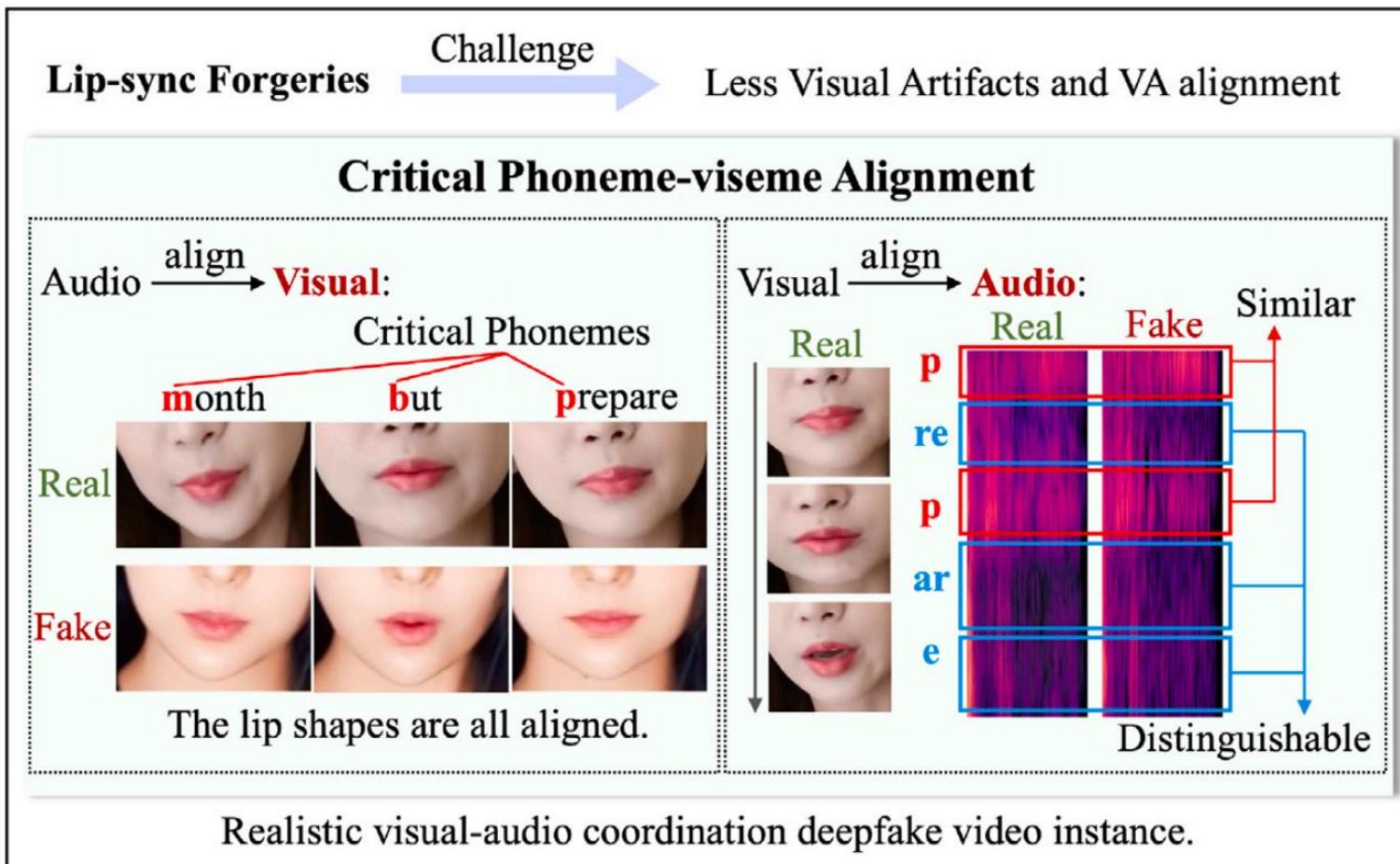


北京交通大学  
BEIJING JIAOTONG UNIVERSITY

# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## 研究动机

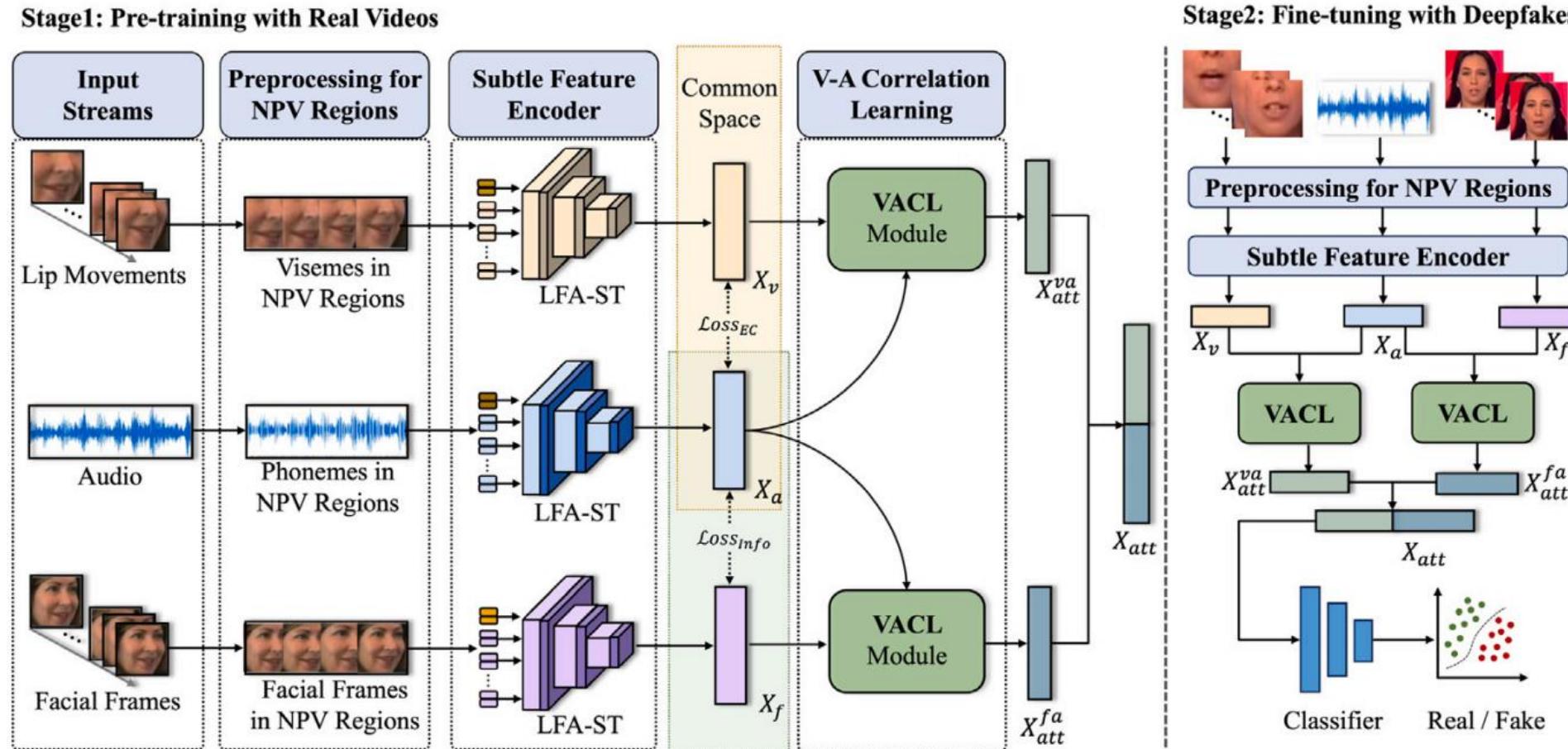
- 非关键区域的音素-视素不一致性，是一个通用伪造检测线索。



# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## 方法框架

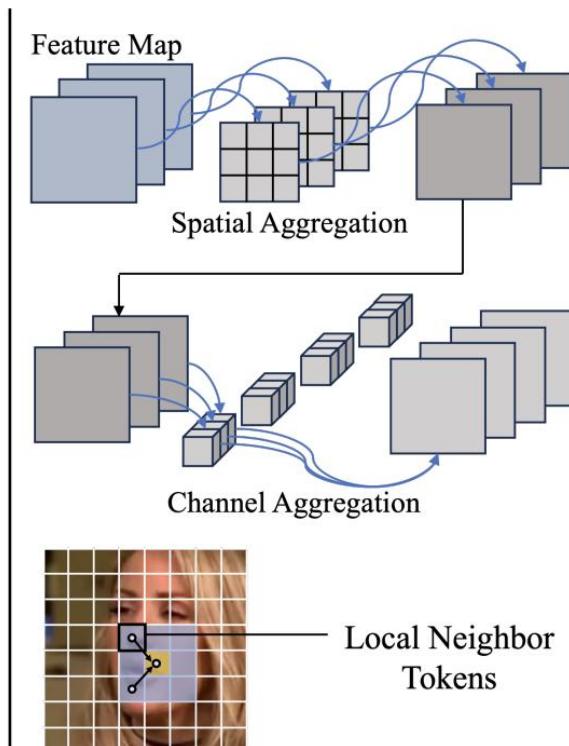
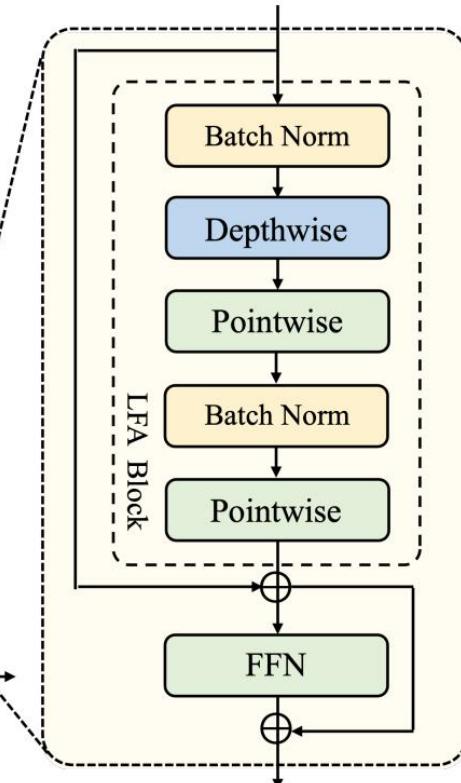
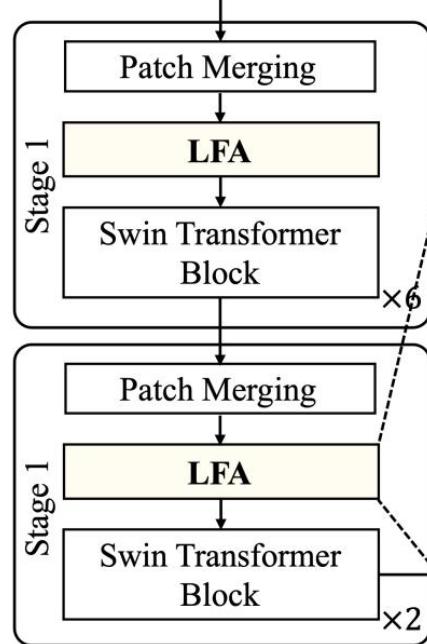
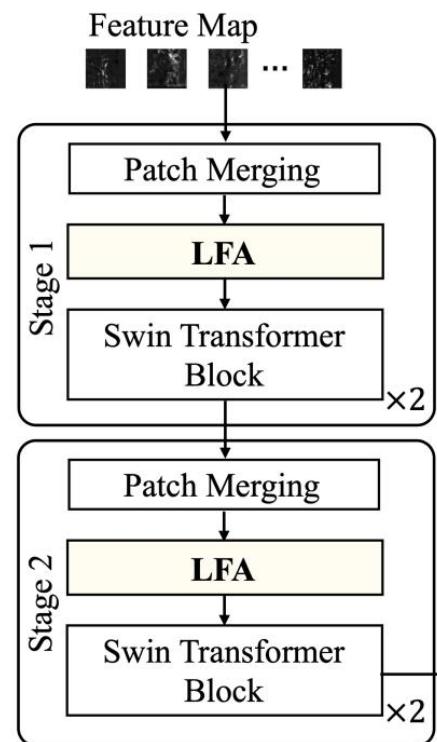
- 学习真实视频中的VA一致性，以帮助Deepfake检测阶段捕捉伪造视频中相对细微的VA不一致性。  
**细微特征提取、VA相关性学习有助于Deepfake检测。**



# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## 方法框架

- 提取细微特征角度：设计集成**局部聚合模块**的Swin Transformer编码器。



(a) Swin Transformer with **Local Feature Aggregation** (LFA-ST)

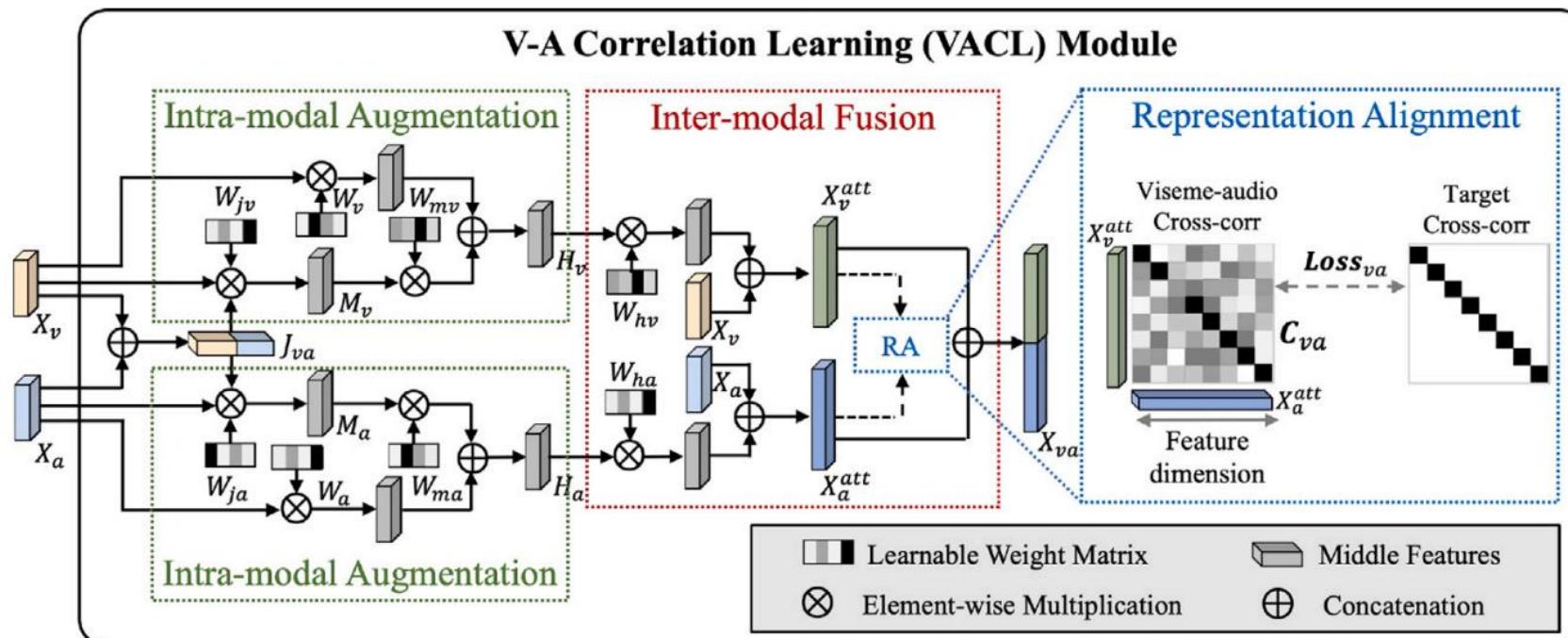
(b) **Local Feature Aggregation** (LFA)

聚合了邻近Token信号的局部信息

# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## 方法框架

- 挖掘VA相关性角度：设计VA相关性学习模块，进行跨模态特征融合和表征对齐，从而缩小模态差距，更好地探索VA相关性。



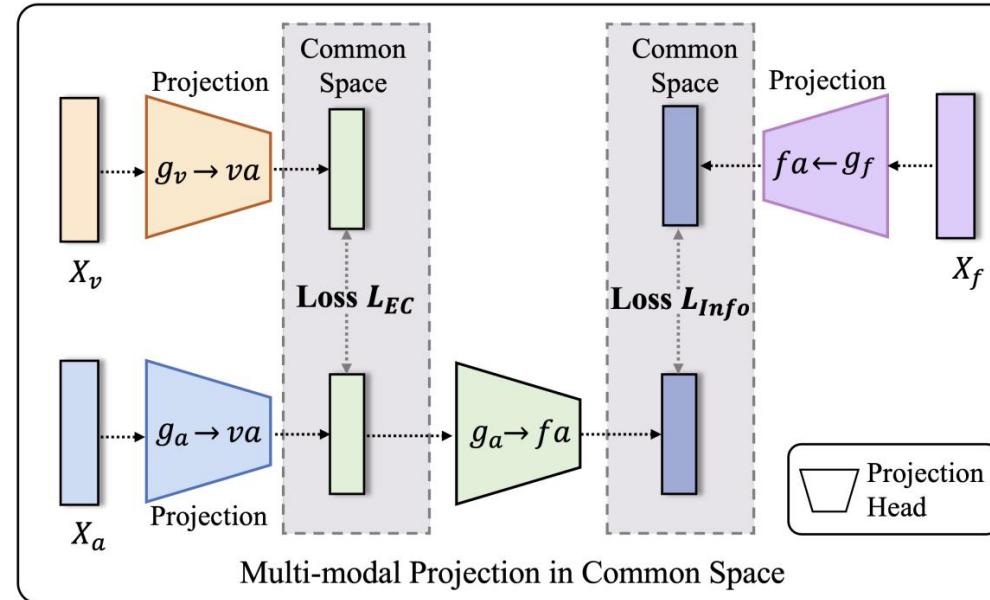
$$\text{相关性损失 } L_{cor}: \quad Loss_{va} \triangleq \sum_i (1 - C_{va}^{ii})^2 + \lambda \sum_i \sum_{j \neq i} (C_{va}^{ij})^2,$$

$$Loss_{fa'} \triangleq \sum_i (1 - C_{fa'}^{ii})^2 + \lambda \sum_i \sum_{j \neq i} (C_{fa'}^{ij})^2.$$

$$Loss_{cor} = Loss_{va} + Loss_{fa'}.$$

## 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

### 两阶段模式的约束



(1) 预训练阶段:

$$L_{pre} = \lambda L_{EC} + \beta L_{Info} + L_{cor}.$$

$$\lambda = \beta = 0.1.$$

(2) 微调阶段:

$$L = \lambda L_{EC} + \beta L_{Info} + \gamma L_{cor} + \omega L_{ce}.$$

$$\lambda = \beta = 0.1 \quad \omega = 0.4.$$

微调视频为fake时,  $\gamma = 0$ ;

微调视频为real时,  $\gamma = 1$ .

## 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

### ○ 实验数据集

**Table 2**

Details of the datasets used in our experiments.

Datasets	Fake/Real scale	Subclasses
FF++[42]	4000/1000	RVRA, FVFA
FSh [44]	Based on FF++[42]	RVRA, FVFA
Celeb-DF [45]	5639/590	RVRA, FVRA
DFo [46]	50 000/10 000	RVRA, FVRA
DFDC [47]	6000/32 200	RVRA, FVRA
FakeAVCeleb [48]	391/16,869	RVRA, RVFA, FVRA, FVFA
A2V [49]	14 h Obama's videos	RVRA, FVRA (Only lip motions forged)
T2V [28]	428/100	RVRA, FVRA (Only lip motions forged)

# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## ○ 实验结果

### (一) 库内评估

Table 3

Intra-dataset evaluation (%). The best results are shown in bold.

Method	Modality	Pre-train	ACC in FF++			AUC in FF++			FakeAVCeleb	
			Raw	HQ	LQ	Raw	HQ	LQ	ACC	AUC
Patch-based [50]	V	-	98.9	92.6	79.1	99.8	97.1	78.3	80.2	83.1
CNN-aug [51]	V	-	98.7	96.9	81.6	99.8	99.1	86.9	82.6	83.7
Xception [42]	V	-	98.9	97.0	89.1	99.8	99.3	91.4	86.9	88.1
Two-branch [14]	V	-	-	-	-	-	99.1	91.0	80.5	81.8
Face X-ray [52]	V	Sup-BI	99.1	78.4	34.2	99.8	97.6	77.3	86.2	87.2
LipForensics [19]	V	Sup-LRW	98.9	98.0	94.2	99.4	<b>99.7</b>	96.1	92.1	93.4
Self-LipForensics [19]	V	Sup-LRW	98.6	96.5	88.4	99.8	99.3	94.8	92.4	94.1
RECCE [53]	V	-	99.1	97.1	91.0	99.8	99.3	95.0	88.9	89.7
CDIN [22]	V	-	96.7	95.3	90.8	97.0	96.5	92.7	84.6	85.0
Method [34]	V	-	98.9	97.3	88.6	99.2	97.8	92.9	87.6	88.4
HPE-based [35]	V	-	99.3	98.2	96.8	99.8	98.7	94.7	83.9	84.6
Joint A-V [25]	V-A	-	98.6	98.0	95.8	99.3	99.0	91.4	98.7	97.1
AVoID-DF [24]	V-A	-	99.0	98.2	93.9	<b>99.9</b>	99.2	93.5	95.7	96.2
SST [54]	V-A	Self-Sup	<b>99.2</b>	<b>98.5</b>	93.5	99.7	99.6	95.7	95.4	97.0
VFD [23]	V-A	Self-Sup	98.3	94.2	90.1	99.4	96.5	89.6	91.5	92.4
AVForensics [36]	V-A	Self-Sup	98.5	96.8	91.4	98.8	97.1	92.9	93.5	94.4
NPVForensics	V-A	Self-Sup	<b>99.2</b>	<b>98.5</b>	<b>96.3</b>	<b>99.9</b>	<b>99.7</b>	<b>96.4</b>	<b>98.9</b>	<b>99.0</b>

# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## ○ 实验结果

### (二) 跨操纵方法评估

Table 4

Cross-manipulation evaluation. Video-level ACC, AUC, and F1 score (%) when testing on each forgery type of FF++ HQ after training on the remaining three. The best results are shown in bold.

Method	DF			FS			F2F			NT		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Patch-based [50]	91.6	94.0	89.8	57.8	60.5	57.7	86.0	87.3	85.8	82.4	84.8	82.3
CNN-aug [51]	85.9	87.5	85.1	56.0	56.3	55.8	78.9	80.1	78.9	66.7	67.8	66.0
Xception [42]	93.5	93.9	92.6	51.2	51.2	50.3	84.9	86.8	83.7	76.2	79.7	75.9
Face X-ray [52]	97.4	99.7	96.8	87.5	90.1	87.1	96.6	99.2	93.4	94.2	98.1	92.5
Self-LipForensics [19]	92.7	97.8	88.4	85.9	90.5	83.1	95.2	98.0	93.3	93.5	96.9	90.8
RECCE [53]	96.6	98.2	96.4	88.0	88.9	86.4	89.5	92.1	89.3	87.1	88.5	86.0
CDIN [22]	84.1	84.5	83.3	83.8	84.7	82.5	85.4	86.6	85.0	84.8	86.3	84.2
Method [34]	<b>98.9</b>	99.3	94.6	93.9	96.0	92.2	95.7	97.1	95.7	86.1	90.1	86.1
HPE-based [35]	86.4	87.2	84.5	85.5	87.2	83.8	86.7	87.0	84.1	85.7	87.6	82.0
Emotions do not lie [55]	92.3	94.5	91.1	87.6	89.3	86.1	85.3	86.9	84.8	90.4	93.7	90.1
Joint A-V [25]	92.8	97.7	91.8	86.9	90.5	83.7	95.5	<b>99.7</b>	94.8	94.6	97.3	93.9
SST [54]	93.2	94.5	93.1	90.3	91.9	90.0	96.8	98.3	94.9	95.4	96.4	95.1
VFD [23]	91.6	96.2	88.2	81.1	86.3	79.4	85.0	89.6	84.6	90.3	94.2	88.7
AVoID-DF [24]	96.9	97.3	96.5	93.8	94.7	93.0	94.2	94.5	93.7	95.0	95.1	94.6
AVForensics [36]	98.1	98.9	97.5	94.2	95.7	93.8	95.4	96.9	95.1	97.5	98.1	97.4
<b>NPVForensics</b>	98.5	<b>99.8</b>	97.8	94.7	96.2	94.5	<b>98.5</b>	99.4	<b>95.8</b>	<b>98.2</b>	<b>98.6</b>	<b>98.0</b>

# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## ○ 实验结果

### (三) 跨数据集评估

Table 5

Generalizability across datasets. Video-level ACC, AUC, and F1 score (%) tests on Celeb-DF, DFDC, FSh HQ, DFo, A2V, T2V-L, and T2V-S. The best results are shown in bold (Train on FF++).

Method	Celeb-DF	DFDC	FSh	DFo	A2V	T2V-L	T2V-S
	ACC AUC F1						
Xception [42]	69.8 73.7 66.5	65.6 70.9 67.3	68.2 72.0 66.7	79.4 84.5 77.0	77.7 81.6 74.8	59.3 63.9 56.2	70.4 75.3 67.5
Face X-ray [52]	77.5 79.5 76.2	62.9 65.5 62.0	88.6 92.8 84.1	83.5 86.8 82.0	72.8 77.2 69.3	58.4 60.3 57.5	77.0 81.4 75.2
CNN-GRU [56]	68.5 69.8 66.1	67.6 68.9 65.9	78.3 80.8 76.2	72.7 74.1 70.5	78.7 80.1 75.8	64.5 66.3 62.8	80.1 82.2 77.6
LipForensics [19]	80.3 82.4 78.5	71.8 73.5 70.4	93.9 95.9 92.2	95.2 96.6 94.1	82.9 84.4 81.0	72.8 74.2 71.4	84.7 86.6 82.5
Self-LipForensics [19]	81.5 83.6 79.3	71.9 73.6 70.2	94.1 96.1 92.4	95.3 97.0 94.3	83.1 84.6 81.3	73.0 74.3 71.7	85.0 87.1 82.8
RECCE [53]	76.9 78.5 75.4	67.4 69.1 65.9	83.4 84.9 81.7	87.0 88.7 85.1	67.5 69.1 65.7	64.7 66.3 62.6	76.4 78.2 74.3
CDIN [22]	86.3 87.5 85.1	82.5 84.7 81.3	75.4 76.8 73.7	89.1 90.3 87.9	81.7 82.8 80.5	77.2 78.5 76.0	69.8 71.4 69.3
Method [34]	85.9 87.2 84.3	84.8 86.0 83.5	80.9 82.1 80.0	89.7 90.9 88.5	82.3 83.4 81.1	77.8 79.1 76.6	71.4 73.0 68.9
HPE-based [35]	85.0 86.3 83.7	82.9 84.3 81.6	77.0 78.2 75.4	85.8 87.0 84.6	76.4 77.5 75.2	71.9 73.2 70.7	64.5 66.1 63.0
Emotions do not lie [55]	80.5 82.1 79.1	82.8 84.4 81.5	94.5 96.0 92.9	94.8 96.3 93.2	-	-	-
Joint A-V [25]	81.8 83.9 79.7	74.9 76.5 73.4	95.0 96.9 93.6	94.0 95.8 92.4	78.9 80.5 77.2	68.3 69.8 66.7	78.6 80.2 77.0
SST [54]	82.6 84.2 80.9	73.0 74.5 71.7	96.1 97.8 94.4	95.5 97.3 94.1	88.1 89.6 86.2	75.9 77.4 74.5	85.0 86.7 83.2
VFD [23]	79.4 80.7 75.8	82.8 85.1 75.2	83.0 85.9 76.4	80.5 84.3 75.9	66.3 67.8 62.9	57.6 60.9 58.1	64.1 65.4 62.5
PVM [27]	84.4 85.7 83.1	84.8 86.2 83.4	87.8 89.0 86.1	89.9 91.2 88.4	93.1 94.6 91.8	76.6 79.7 78.2	71.3 74.1 71.5
AVoID-DF [24]	84.7 86.2 83.0	80.8 82.3 79.4	94.2 95.7 92.5	94.9 96.5 93.1	83.4 85.1 81.7	77.0 78.3 75.2	83.1 84.4 81.2
AVForensics [36]	86.6 87.9 85.3	83.4 84.7 81.7	96.5 97.9 94.8	96.7 98.1 95.1	86.4 87.9 84.6	76.5 77.7 74.8	80.7 82.3 78.9
NPVForensics	86.2 88.4 85.4	86.1 86.2 85.8	96.9 98.2 95.7	96.4 98.6 95.8	95.3 97.1 94.7	80.4 83.3 80.0	89.3 91.9 87.8

# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

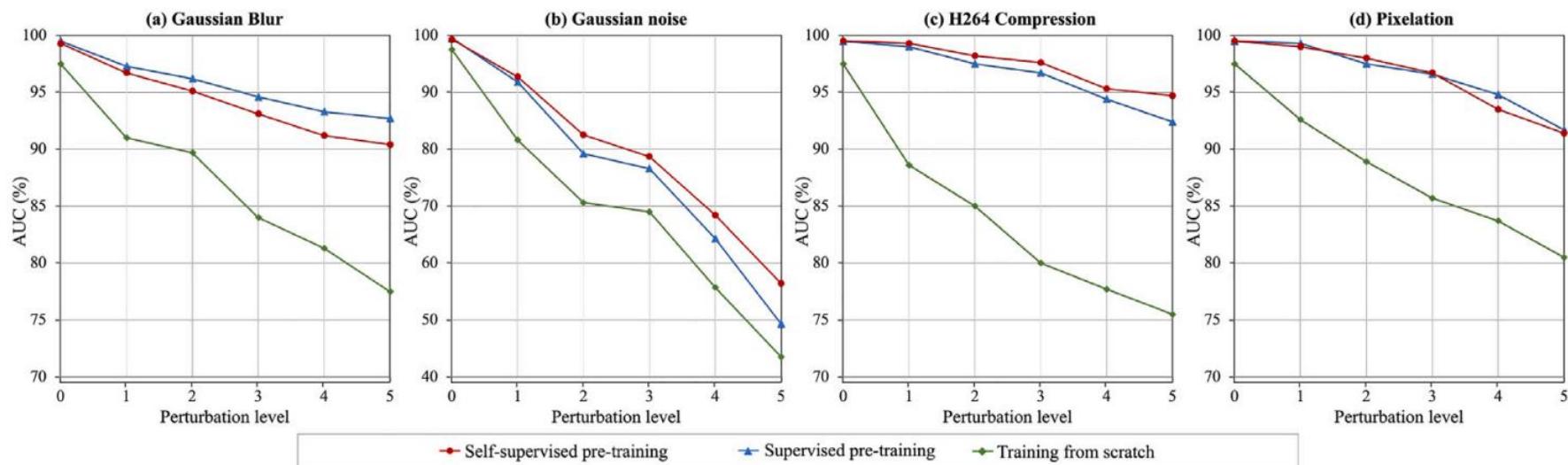
## 实验结果

### (四) 鲁棒性实验

Table 6

Robustness to common degradations. The models are fine-tuning with FakeAVCeleb datasets. AUC (%) scores at five intensity levels are averaged for each degradation type. The best results are shown in bold.

Method	Saturation	Block	Noise	Blur	Pixel	Compress	Avg
Patch-based [50]	84.3	98.8	50.0	54.4	56.7	53.4	66.3
CNN-aug [51]	99.3	95.2	54.7	76.5	91.2	72.5	81.6
Xception [42]	99.3	98.7	53.8	60.8	74.2	62.1	74.8
Face X-ray [52]	97.6	<b>99.1</b>	49.8	63.8	88.6	55.2	75.7
LipForensics [19]	<b>99.9</b>	87.4	73.8	93.1	95.6	94.5	90.7
Self-LipForensics [19]	99.8	87.1	73.7	94.2	93.7	93.2	90.3
SST [54]	98.3	89.3	64.8	94.0	89.2	86.7	87.1
VFD [23]	90.3	87.3	65.1	85.3	84.9	89.5	83.7
AVoID-DF [24]	97.9	92.4	66.6	92.9	90.4	90.5	88.5
AVForensics [36]	99.1	95.1	70.8	91.2	92.8	91.7	90.1
<b>NPVForensics</b>	99.7	97.9	<b>79.7</b>	<b>94.3</b>	<b>96.2</b>	<b>97.6</b>	<b>94.2</b>



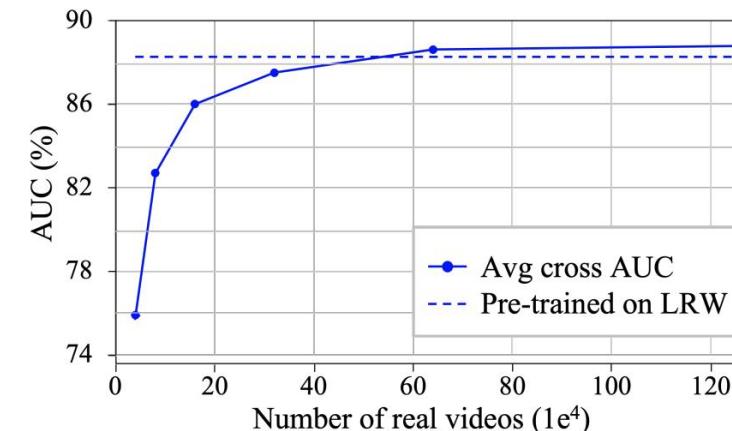
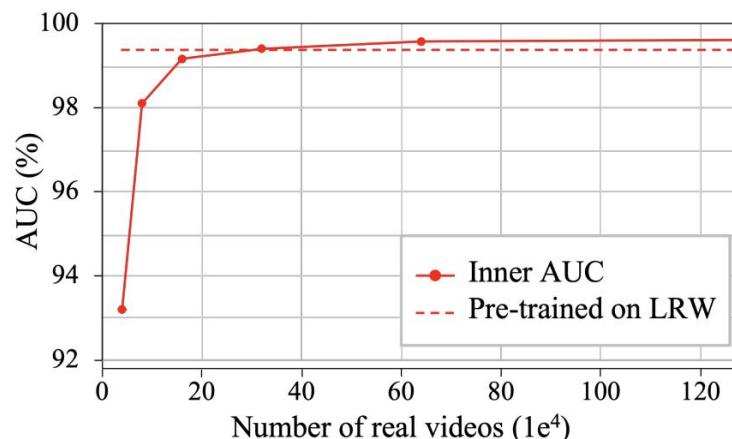
# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## 实验结果

### (五) 消融实验

Table 7 Framework ablation. Accuracy scores (%) of FakeAVCeleb after training on the FF++ dataset. The best results are shown in bold.

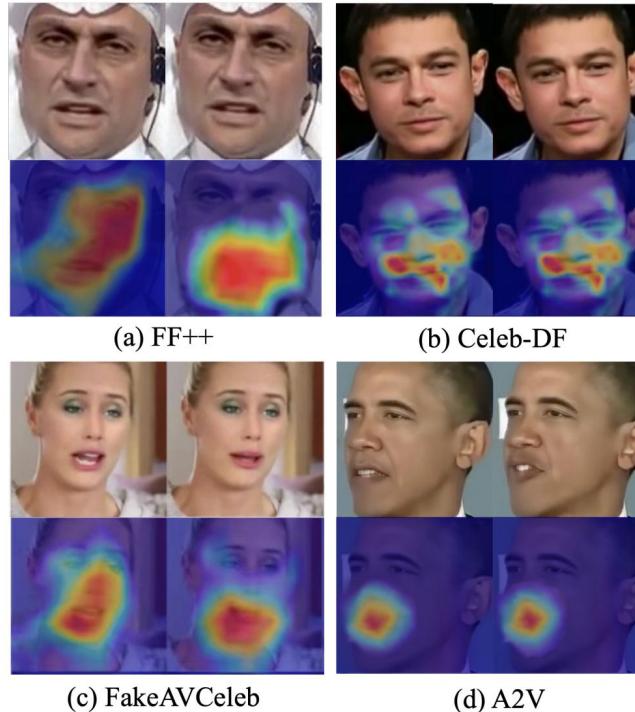
Method	FakeAVCeleb (audio-visual)			Avg
	Real-Fake	Fake-Fake	Fake-Real	
w/o viseme stream	85.3	88.1	87.9	87.1
w/o facial stream	93.5	95.7	95.9	95.0
w/o audio stream	89.9	93.3	93.6	92.3
w/o LFA, only ST	94.8	95.3	93.9	94.7
only ViT as backbone	93.7	94.2	92.1	93.3
LFA-ViT	94.4	95.8	94.0	94.7
w/o VACL	83.4	87.3	84.8	85.2
w/o cross-attention	90.7	91.8	93.3	91.9
w/o representation alignment	93.5	94.6	94.7	94.3
<b>NPVForensics (Ours)</b>	<b>96.7</b>	<b>97.8</b>	<b>96.3</b>	<b>96.9</b>



# 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

## 实验结果

### (六) 可视化



本实例是来自FakeAVCeleb的伪造视频，它的面部身份信息没有被篡改，**只有唇部运动被篡改以配合伪造的音频。这个例子在方法SST、VFD、Avoid-DF中未能被检测出来。**

我们的方法有效的检测出视频真伪，并可视化出视频中的伪造区域。

## 工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

### ○ 总结

- 提出一种新的检测视角，即通过学习非关键音素-视素区域的 VA 相关性来检测具有逼真 VA 操纵效果的 Deepfake 视频。
- 对公共数据集进行综合评估，结果证明该框架的优越性和鉴别能力。特别地，该框架在使用关键音素-视素校准的最新 Deepfake 数据集上表现出色。

### ○ 对应成果

- Yu Chen, Yang Yu, Rongrong Ni, Haoliang Li, Wei Wang, Yao Zhao. NPVForensics: Learning VA Correlations in Non-critical Phoneme-Viseme Regions for Deepfake Detection. **Image and Vision Computing**, 2025: 105461.



# 深度伪造人脸图像及视频的检测方法研究

谢谢各位！敬请指正！

