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

汇报提纲

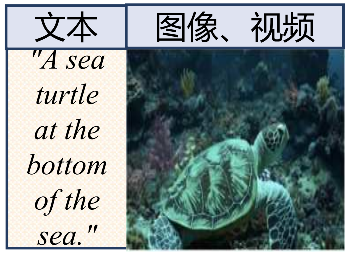
CONTENTS

- **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

深度伪造取证的研究背景

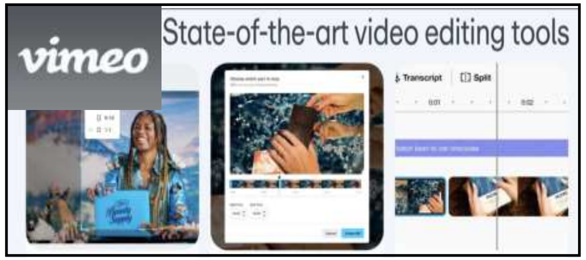
多媒体内容的取证

- 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。

Ps ZAO Midjourney STABLE DIFFUSION



深度伪造取证的研究背景

眼见不再为实，深度伪造真假难辨

■ Real or fake （2019年的伪造水平）

深度伪造取证的研究背景

北京交通大學
BEIJING JIAOTONG UNIVERSITY

深度伪造的现象和危害

2017

深度伪造诞生
Reddit出现大量深度伪造图像。

十一月

2018

伪造奥巴马演讲视频
引起美国国防部高度重视。

四月

2019

“杨幂版”黄蓉
明星换脸进行视频篡改。

三月

扎克伯格假视频
新闻真假难辨。

六月

2020

英国女皇假视频
篡改女皇表情动作及唇部运动。

四月

2021

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

影视明星换脸视频
将“蜘蛛侠”的面部替换至“钢铁侠”身份上。

八月

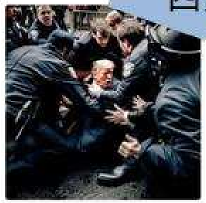
2022

泽连斯基假视频
Deepfake参与战争，总统被伪造投降视频。

三月

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

“拟音”+“拟脸”进行诈骗。

2024

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

深度伪造取证的研究背景

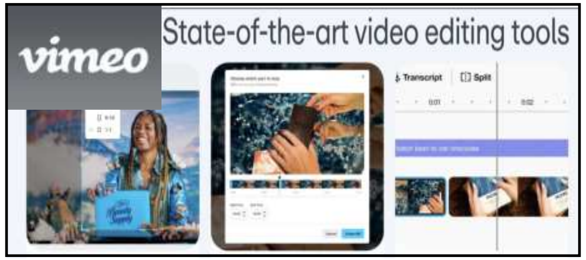
大模型进一步提升了伪造的真实感


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。

深度伪造取证的研究背景

深度伪造流程及研究任务

伪造工具

用户

任务2：
主动防御

伪造发生前

ZAO

STABLE DIFFUSION

Midjourney

OpenAI DALL·E 3

生成式模型

任务1：
真伪取证

伪造已发生

伪造图像

深度伪造的取证及主动防御方法

汇报提纲

CONTENTS

- 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

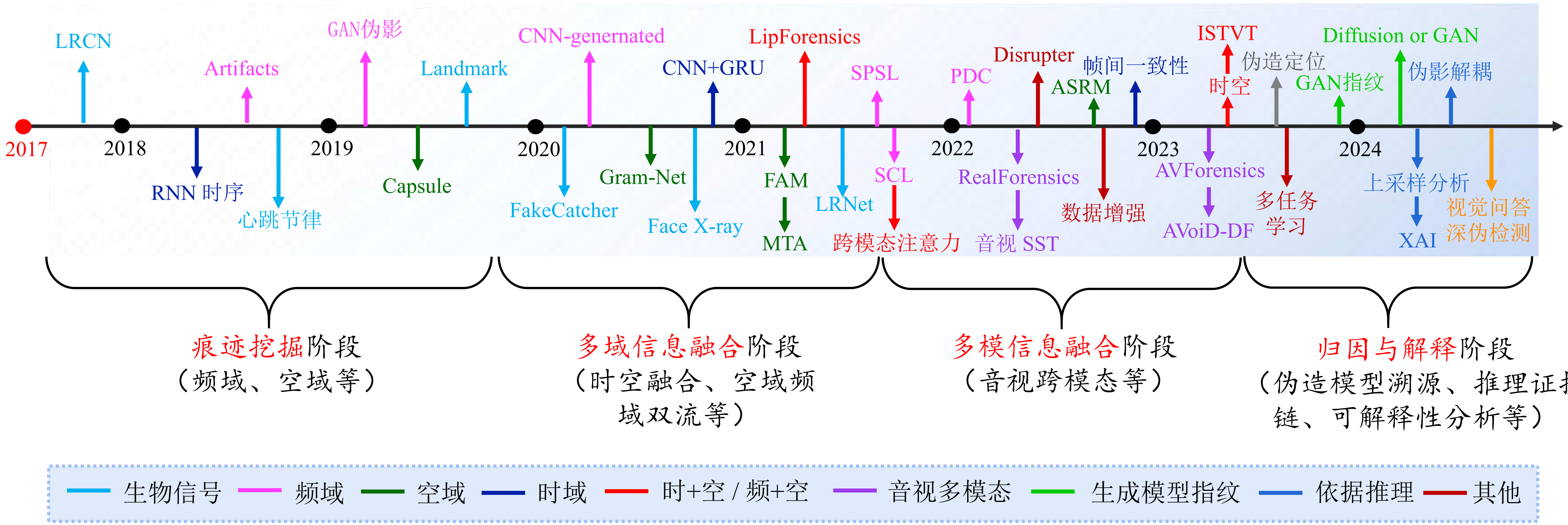
北京交通大学

BEIJING JIAOTONG UNIVERSITY

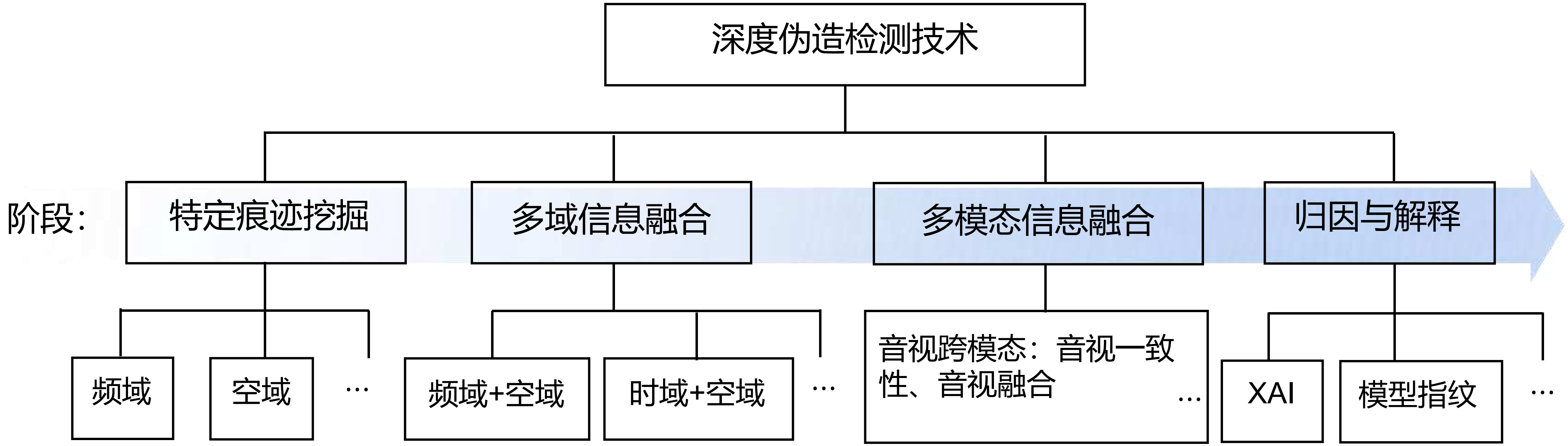
深度伪造检测的发展历程

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



难点问题：

仅能检测特定伪造模式，难泛化。

伪造水平提升，伪造痕迹难挖掘。同时，伪造者开始音+视联合篡改。

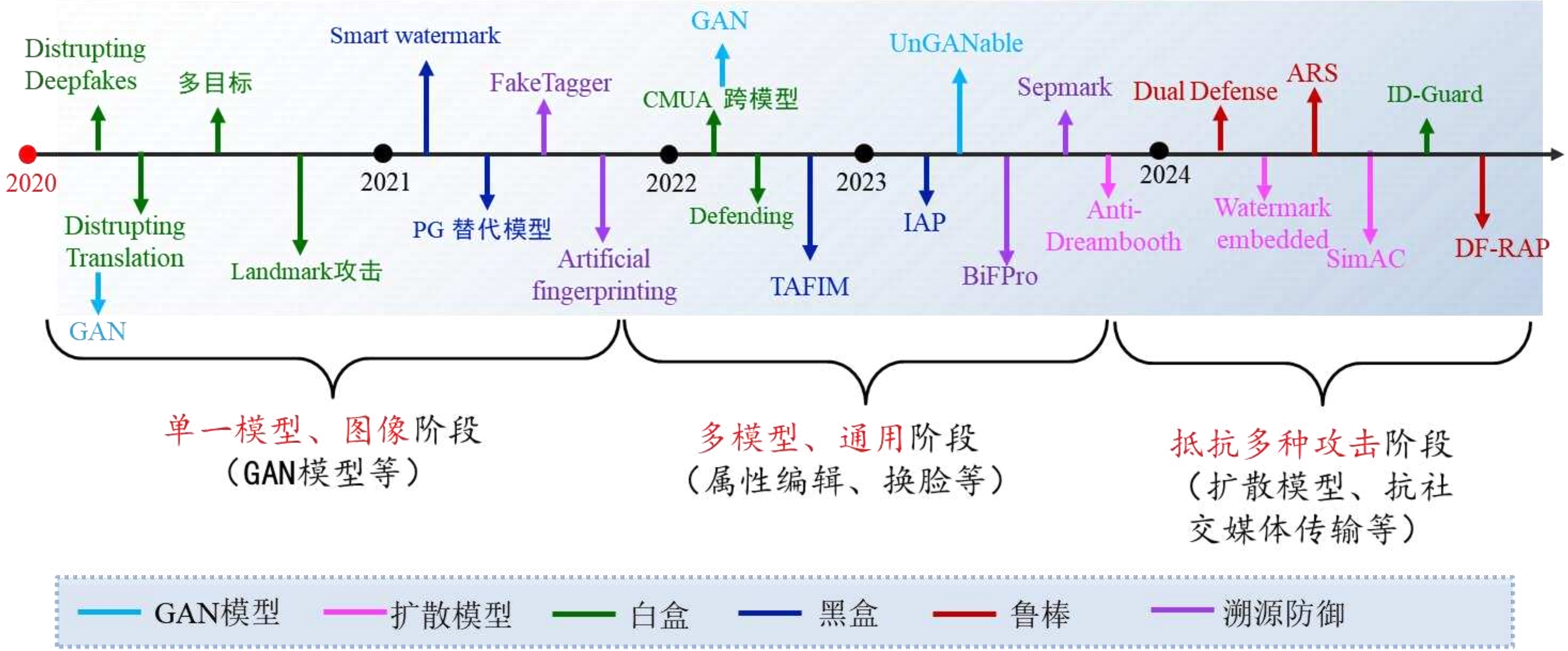
伪造真实度进一步提升，聚焦于有效进行音视多模态信息挖掘。

多种伪造方法难溯源、判别结果证据链难推理、深度模型的可解释性不足。

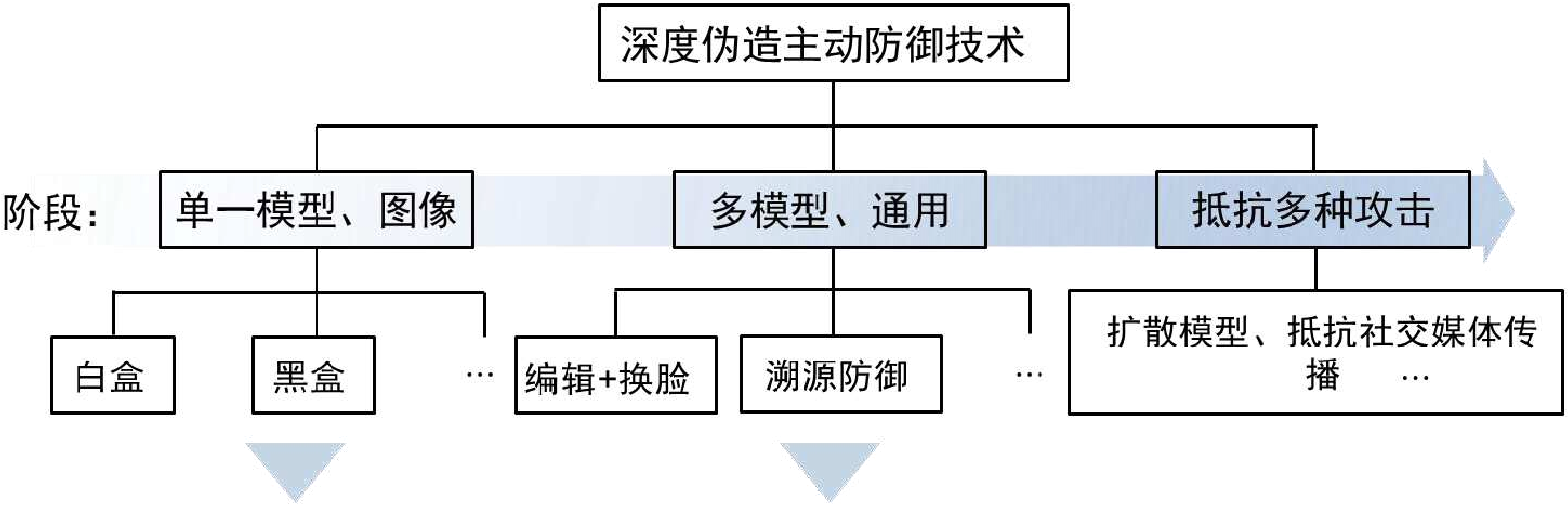
深度伪造主动防御的发展历程

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



难点问题： 仅能针对某一特定生成模型和训练数据，难泛化。

伪造水平提升，多种类型伪造难防御。同时，多种传输以及数据处理攻击难抵抗。

生成式模型进一步发展，聚焦于进行生成式AI模型的鲁棒主动防御。

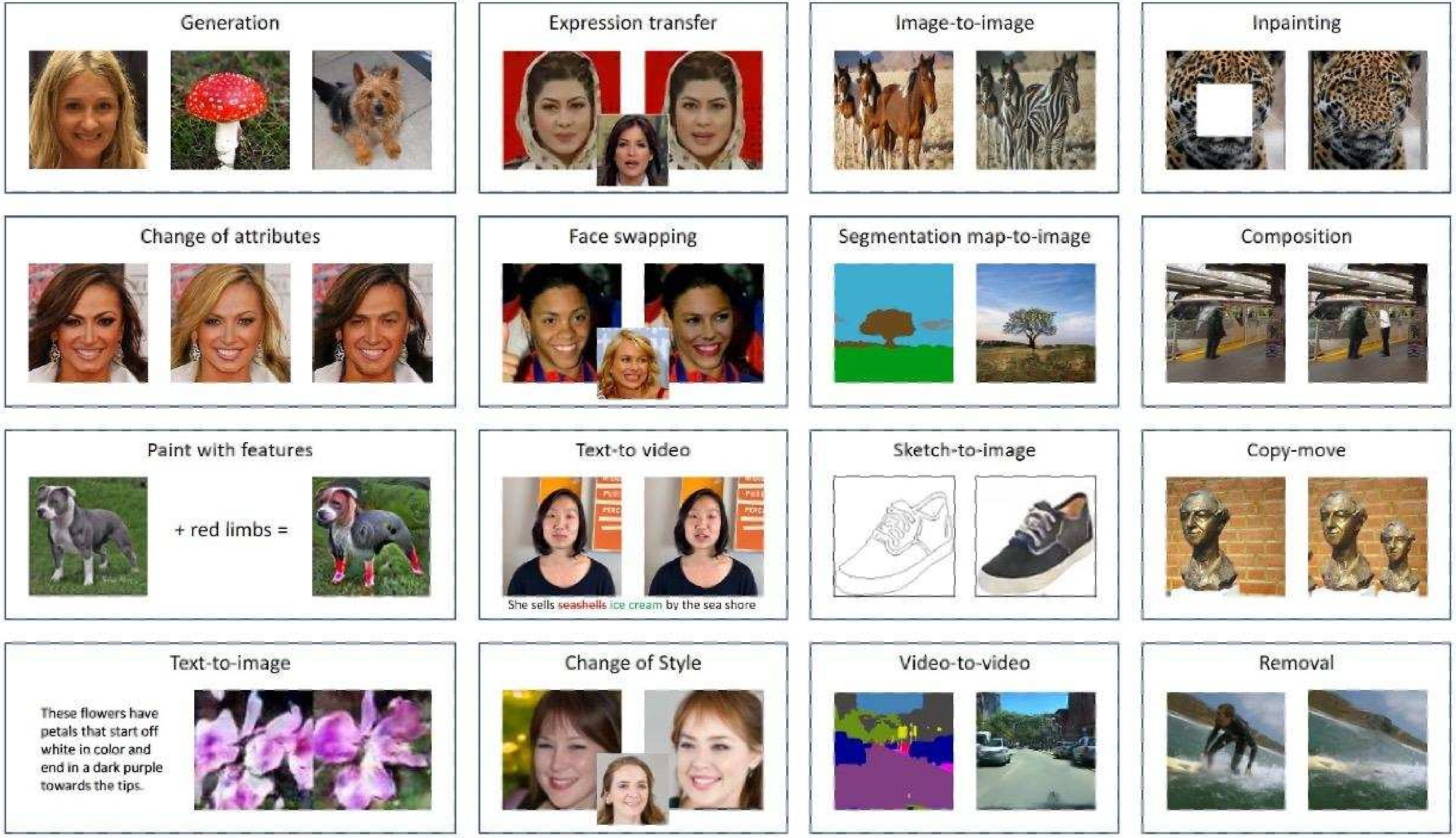
深度伪造的取证及主动防御方法

汇报提纲

CONTENTS

- 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

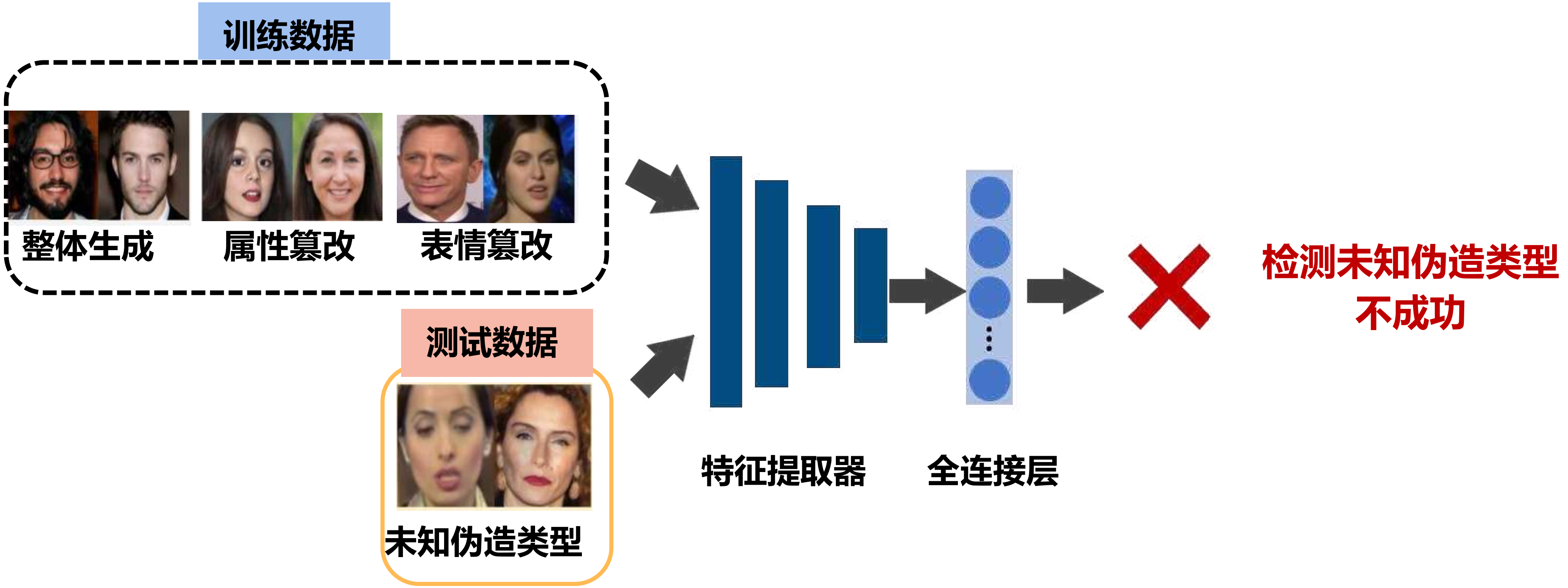
AI生成（合成/伪造）手段越来越多



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

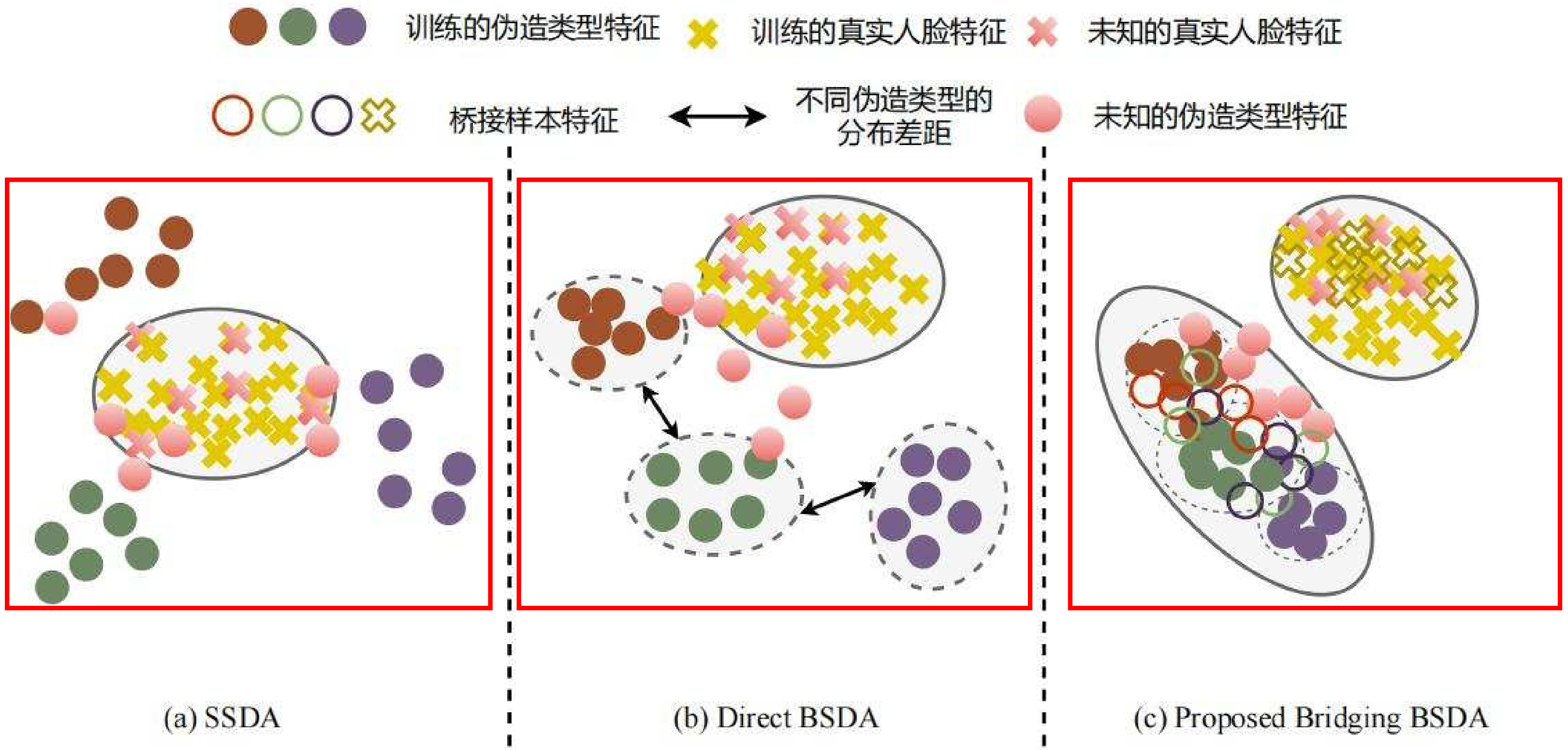
- 面向未知伪造类型，检测方法泛化性能进一步下降



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

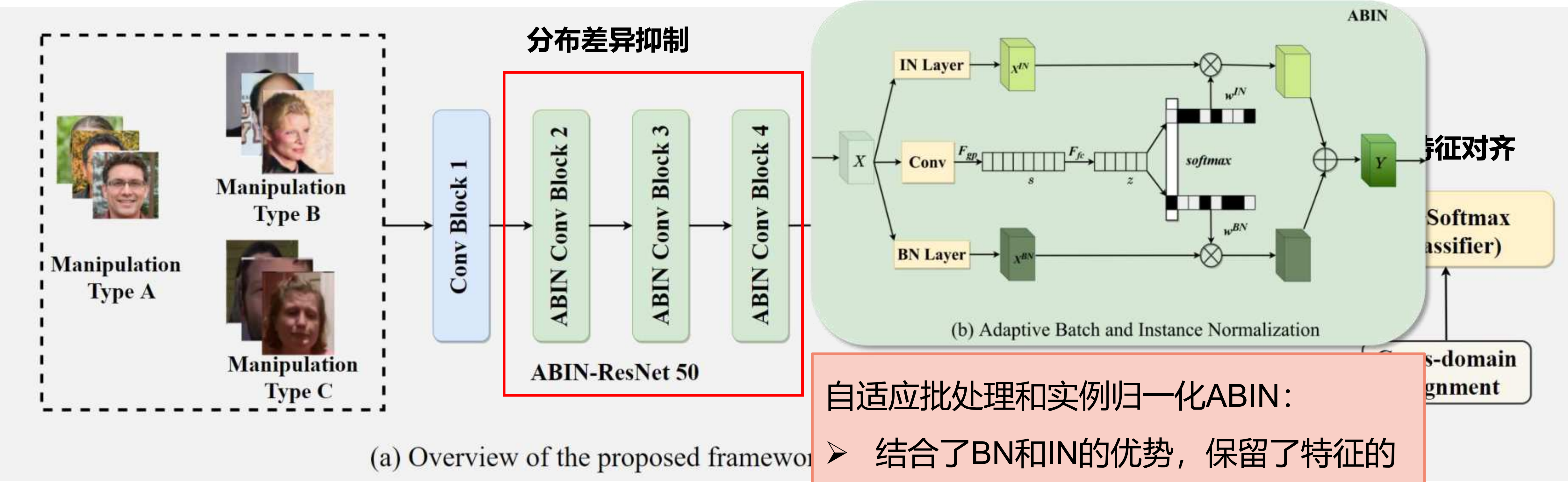
- 现有检测方法存在的问题：不同伪造类型之间存在**显著的分布差异**，现有方法难以对齐伪造特征表示



- **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

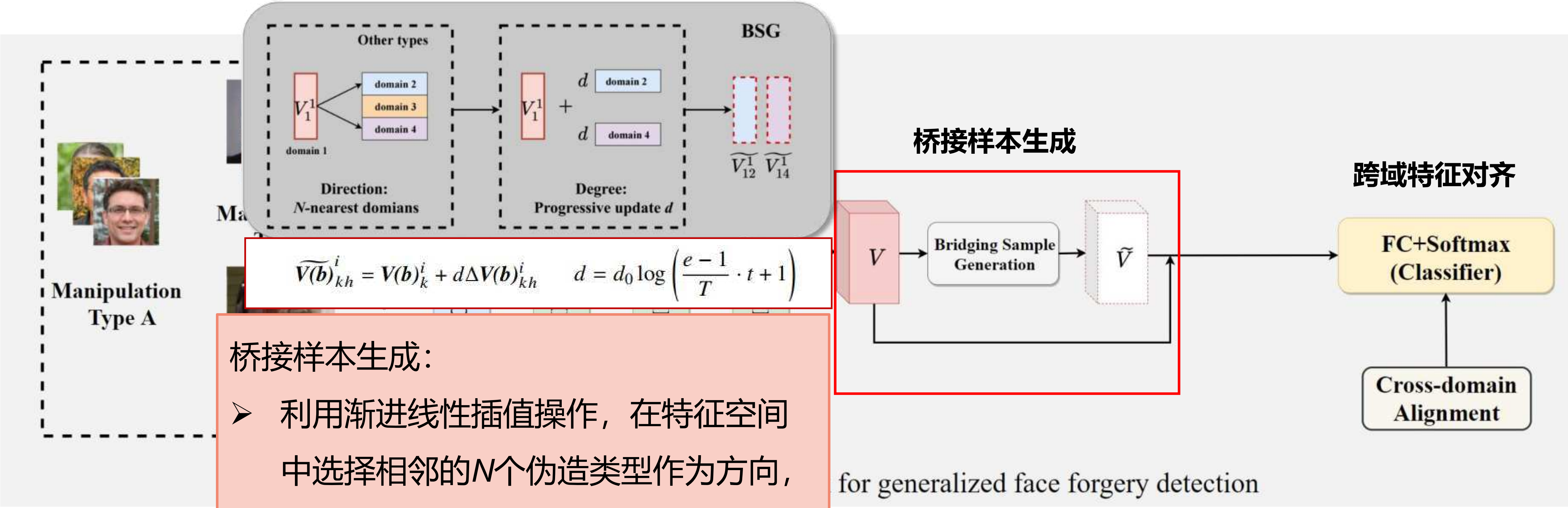


自适应批处理和实例归一化ABIN：

- 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



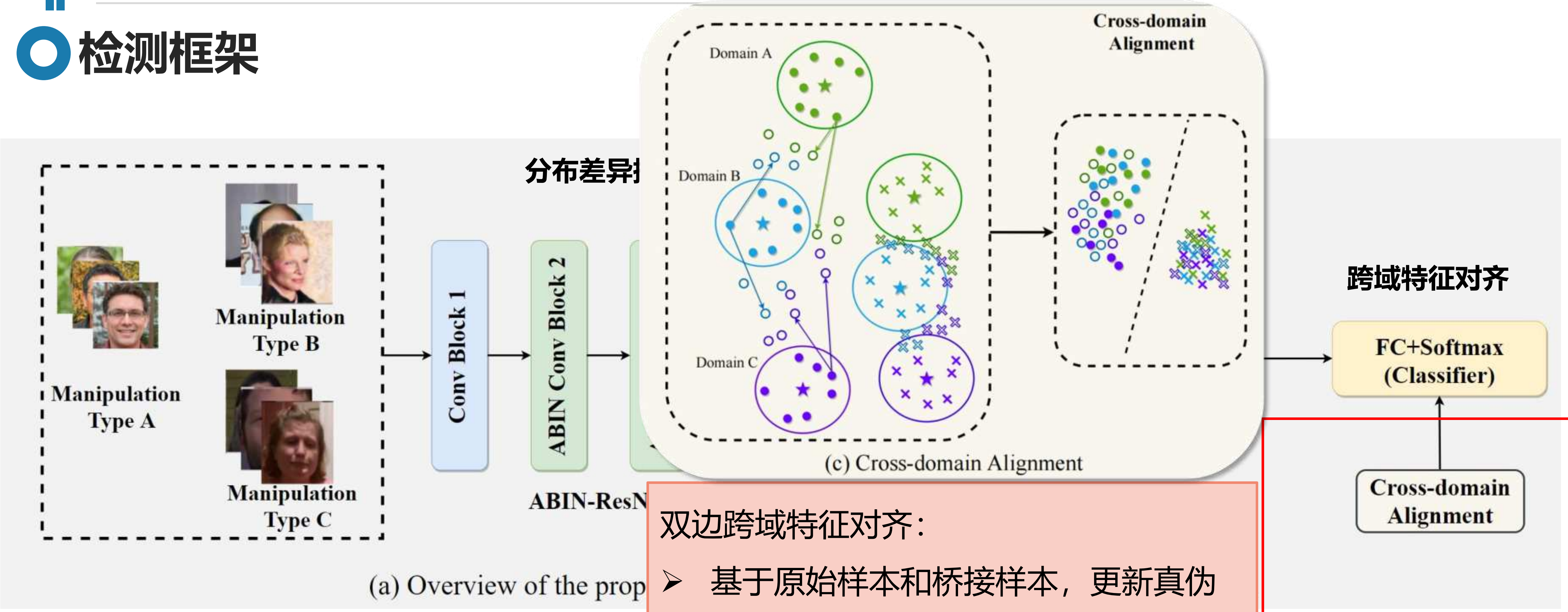
$$\widetilde{V(b)}_{kh}^{i} = V(b)_{k}^{i} + d\Delta V(b)_{kh}^{i} \qquad d = d_0 \log\left(\frac{e-1}{T}\cdot t + 1\right)$$

桥接样本生成：

- 利用渐进线性插值操作，在特征空间中选择相邻的N个伪造类型作为方向，随着训练逐步更新生成位置d，基于原始样本来生成桥接样本分布在各伪造类型之间，填补分布间隙

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



(a) Overview of the prop

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[4]	公开	5000	FFHQ	5000	1024 × 1024
	StyleGAN2[5]		5000	FFHQ	5000	
	StyleSwin[7]		5000	CelebA-HQ	5000	
FEM (Facial Expression Manipulation)	Face2Face[26]	FF++ (c23, c40)	10000	FF++	10000	256 × 256
	FReeNet[9]	个人生成	10000	RaFD	10000	
	DGN[11]		10000	VoxCeleb1	10000	
FAM (Facial Attribute Manipulation)	AttGAN[14]	个人生成	20000	LFW	20000	128 × 128
	StarGAN-V2[16]		10000	FFHQ	10000	256 × 256
	InterFaceGAN[18]		5000	CelebA-HQ	5000	1024 × 1024
FIM (Face Identity Manipulation)	FaceSwap[19]	FF++ (c23, c40)	10000	FF++	10000	256 × 256
	DeepFakes[109]	DFDC	10000	DFDC	10000	
	ADDNet[122]	WildDeepfake	10000	WildDeepfake	10000	

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 StyleSwin FReeNet DGN AttGAN InterFaceGAN DeepFakes ADDNet	StyleGAN Face2Face StarGAN-V2 FaceSwap	N3	StyleGAN StyleGAN2 Face2Face DGN AttGAN StarGAN-V2 FaceSwap ADDNet	StyleSwin FReeNet InterFaceGAN DeepFakes
N2	StyleGAN StyleSwin Face2Face FReeNet InterFaceGAN StarGAN-V2 FaceSwap DeepFakes	StyleGAN2 DGN AttGAN ADDNet	N4	StyleGAN StyleGAN2 Face2Face FReeNet AttGAN StarGAN-V2 FaceSwap DeepFakes	StyleSwin DGN InterFaceGAN ADDNet

方法	N1		N2		N3		N4		平均值	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MaDD[52]	91.36	92.17	91.47	93.26	89.79	91.15	89.63	91.95	90.56	92.13
FrePGAN[60]	92.27	93.59	92.93	93.57	90.09	91.91	91.21	92.36	91.63	92.86
PCL[65]	94.27	96.17	95.06	96.41	92.11	94.54	93.09	95.07	93.63	95.55
PEL[61]	94.18	95.83	94.78	95.12	91.97	94.06	91.77	93.19	93.18	94.55
DABN[118]	92.36	94.17	93.47	94.26	91.79	92.15	91.63	92.95	92.31	93.38
LTW[71]	91.67	92.39	91.74	93.56	90.81	92.37	90.67	92.12	91.22	92.61
FDFL[58]	93.18	95.83	94.78	96.12	92.07	94.57	91.97	95.19	93.00	95.43
DAM[119]	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	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MaDD [52]	80.39	82.74	81.34	83.73	82.07	84.29	79.21	82.09	80.75	83.21
FrePGAN [60]	81.72	83.04	82.05	84.34	83.19	84.64	80.79	83.89	81.94	83.98
PCL [65]	83.62	86.15	84.81	87.35	84.94	86.51	82.81	84.81	84.05	86.21
PEL [61]	82.48	85.08	83.67	86.72	83.89	85.06	81.12	83.86	82.79	85.18
DABN [118]	81.26	84.83	82.57	84.15	81.21	83.73	78.05	80.61	81.77	84.08
LTW [71]	81.07	83.01	81.74	83.87	82.16	84.89	80.84	82.39	81.45	83.54
FDFL [58]	82.56	85.07	83.09	86.91	84.77	86.52	82.23	83.24	83.41	85.44
DAM [119]	83.78	86.79	84.69	87.33	85.47	86.76	83.96	85.57	84.48	86.61
本章方法	**86.97**	**89.43**	**87.89**	**89.64**	**87.81**	**89.82**	**84.97**	**86.67**	**86.91**	**88.89**

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	86.97	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	87.89	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	87.81	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	84.97	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
✓	-	-	-	-	-	**70.09**
-	✓	-	-	-	-	**72.02**
-	-	✓	-	-	-	**77.87**
-	-	✓	✓	-	-	**77.92**
-	-	✓	-	✓	-	**81.79**
-	-	✓	-	✓	✓	**84.97**

N		Protocol N4 ACC	Protocol N4 AUC
Nearest	1	92.86	95.83
	2	94.64	97.78
	3	**94.67**	**97.81**
Random	1	89.36	91.69
	2	91.24	93.71
	3	93.19	95.37

(a) 桥接方向 N 参数设置研究

d_0	FIM ACC	FIM AUC
0.1	82.72	84.79
0.5	83.18	85.31
1.0	**84.97**	**86.67**
2.0	84.51	86.19

(b) 桥接程度初始值 d_0 参数设置研究

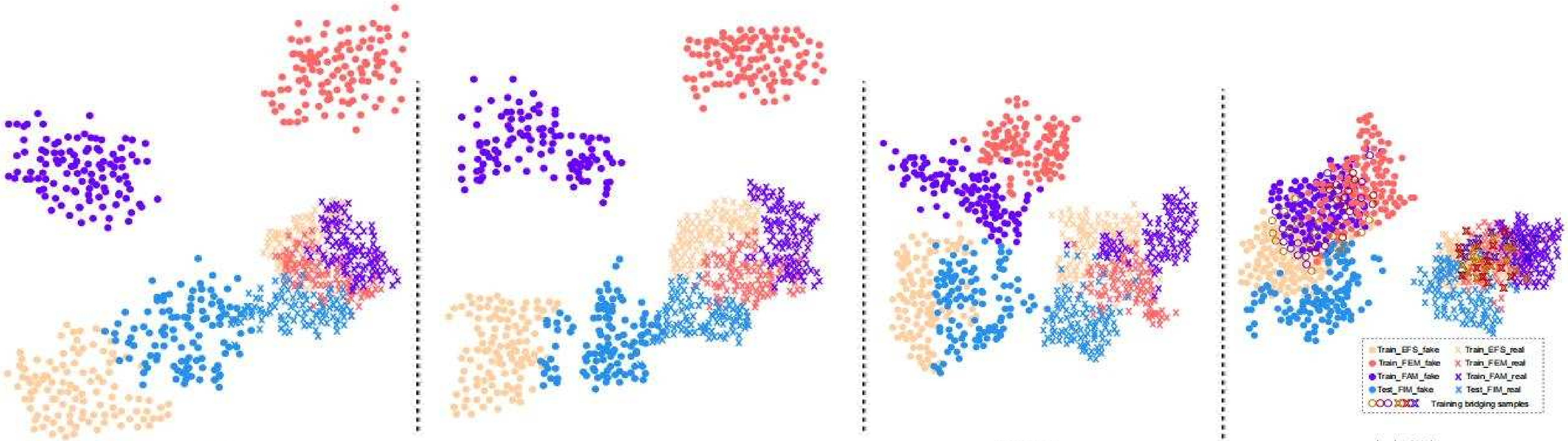
Residual Stage (RS)	FIM ACC	FIM AUC
RS-1-2	78.33	80.12
RS-3-4	81.67	83.96
RS-2-3-4	**84.97**	**86.67**
RS-1-2-3-4	83.06	85.21

(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



(a) LTW (b) FDFL (c) DAM (d) 本章方法

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, Rongorng 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

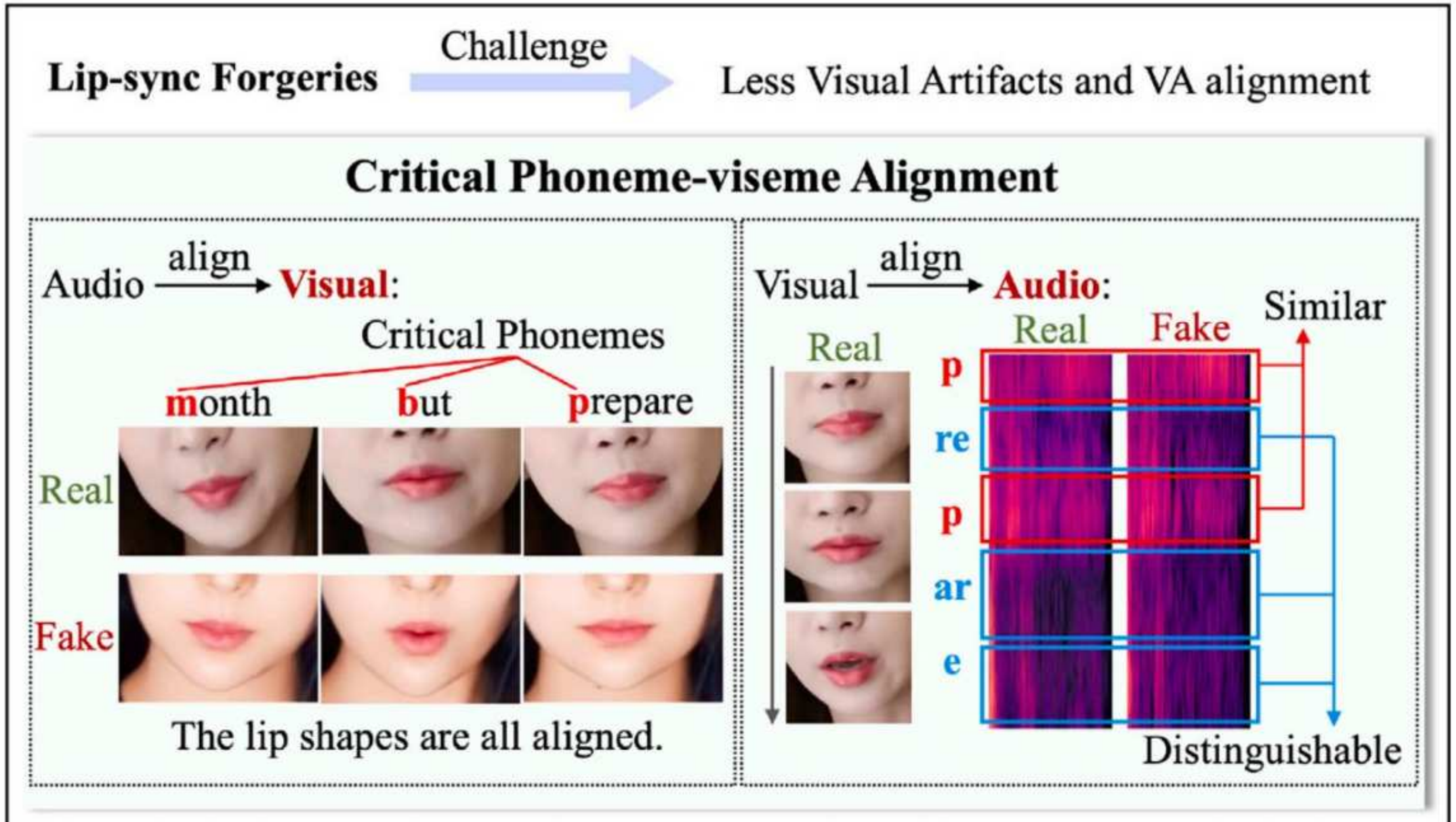
汇报提纲

CONTENTS

- 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**

研究动机

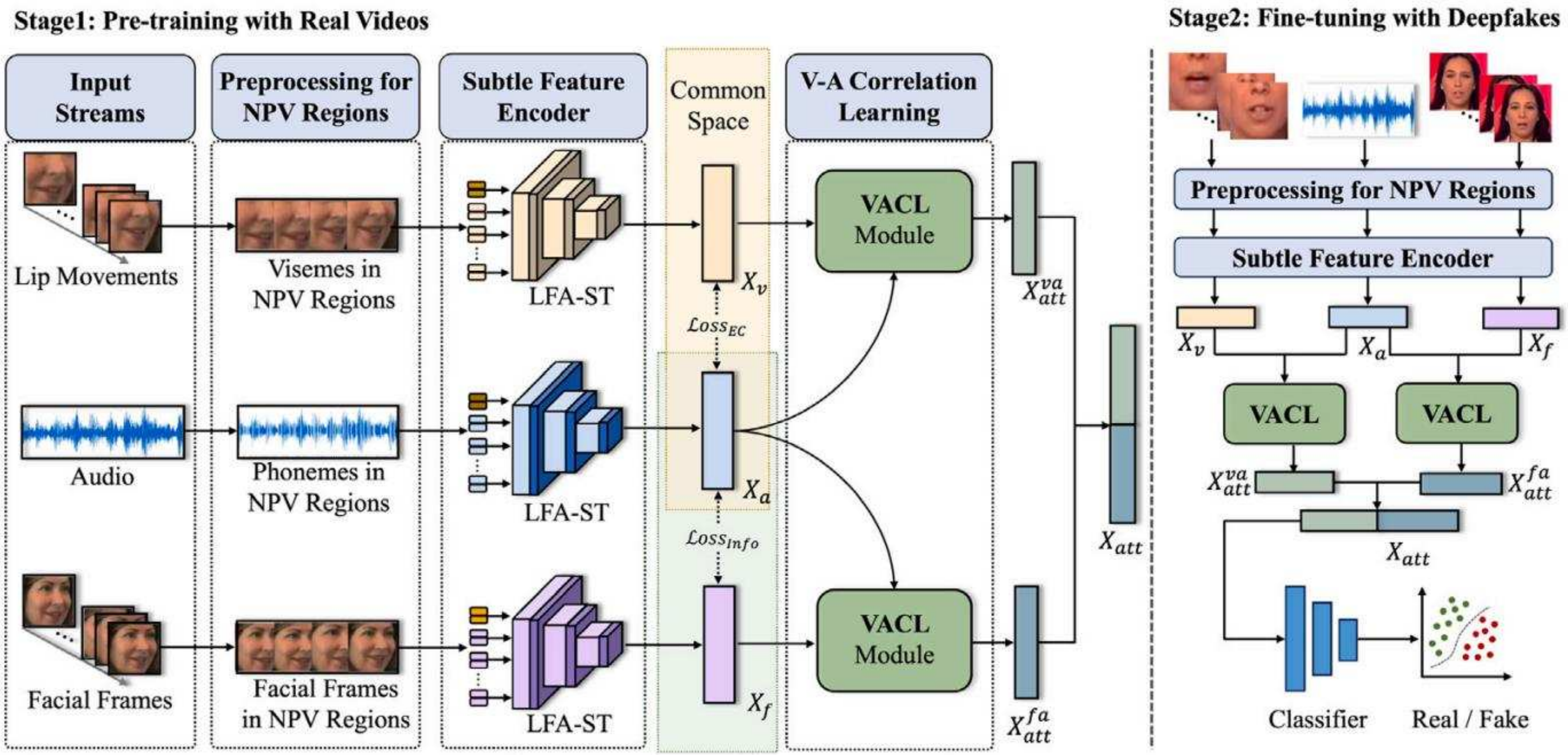
- 非关键区域的音素-视素不一致性，是一个通用伪造检测线索。



Realistic visual-audio coordination deepfake video instance.

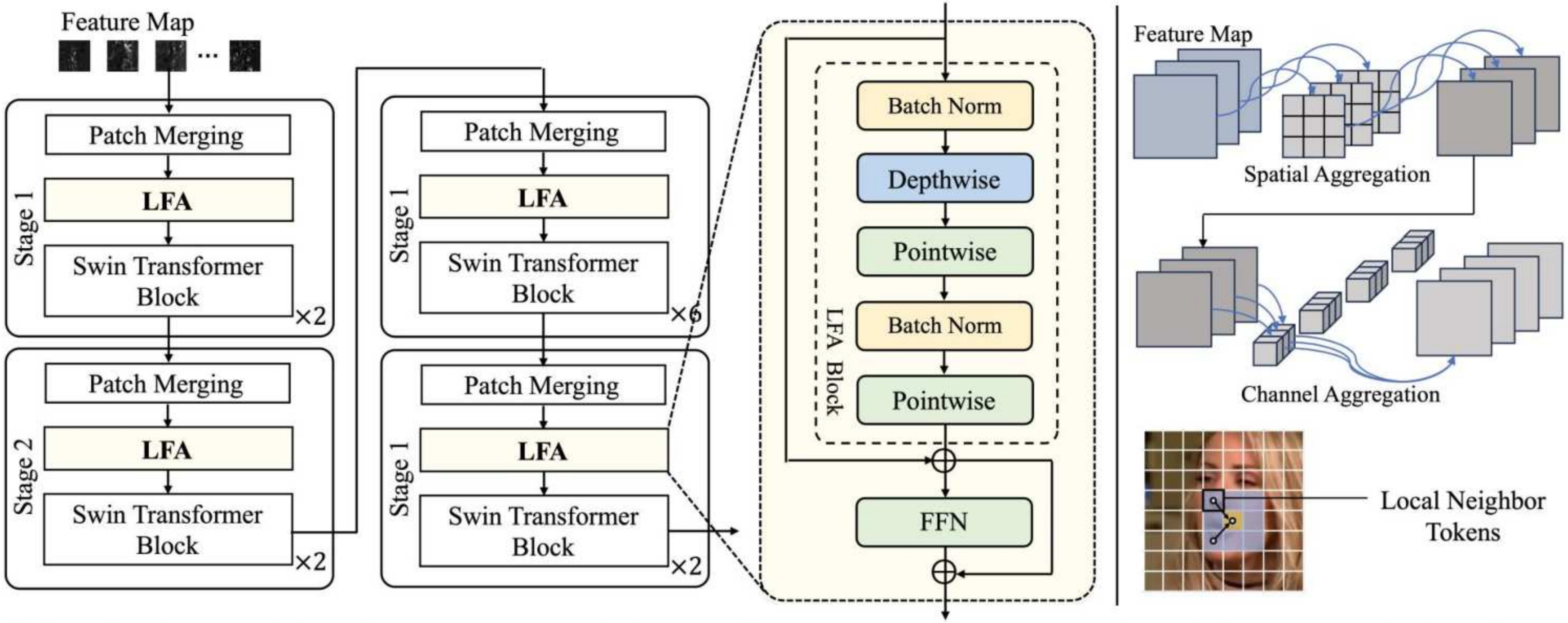
方法框架

- 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。



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。

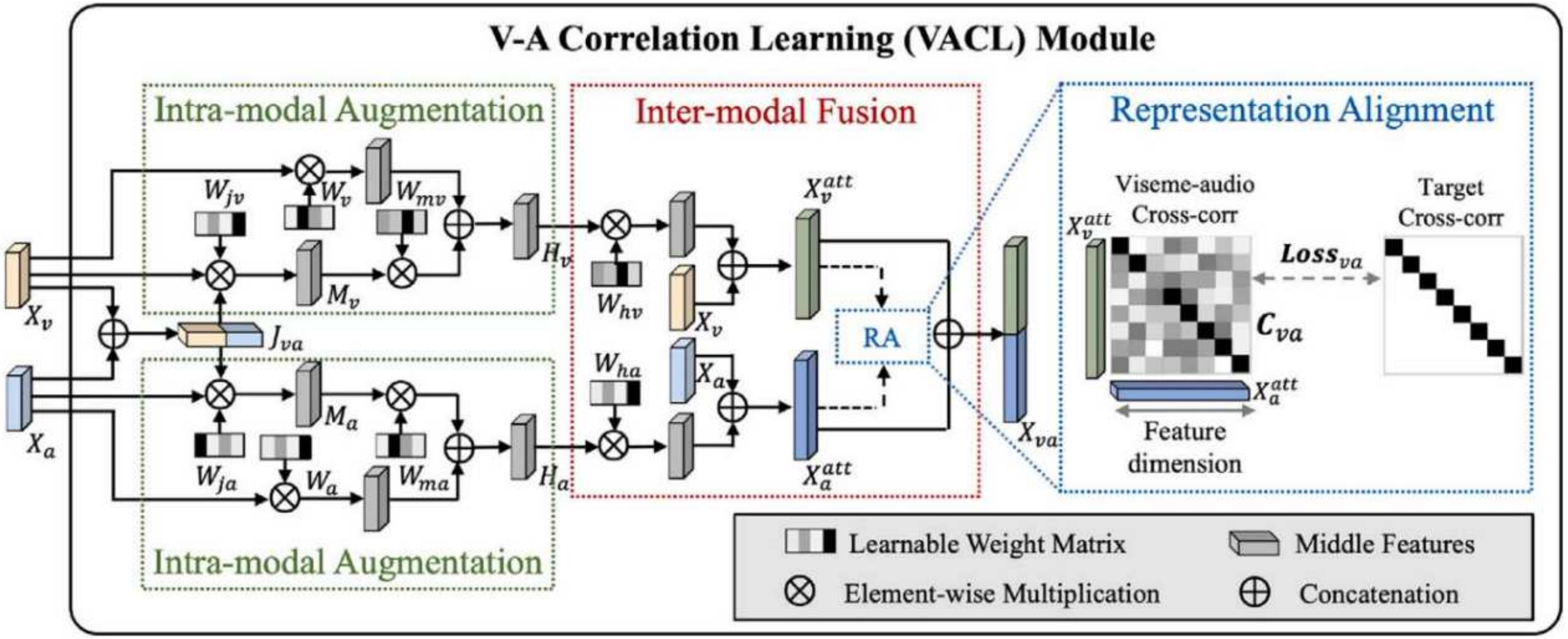


(a) Swin Transformer with **Local Feature Aggregation** (LFA-ST) (b) **Local Feature Aggregation** (LFA) (c) Aggregate Local Information

聚合了邻近Token信号的局部信息

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。



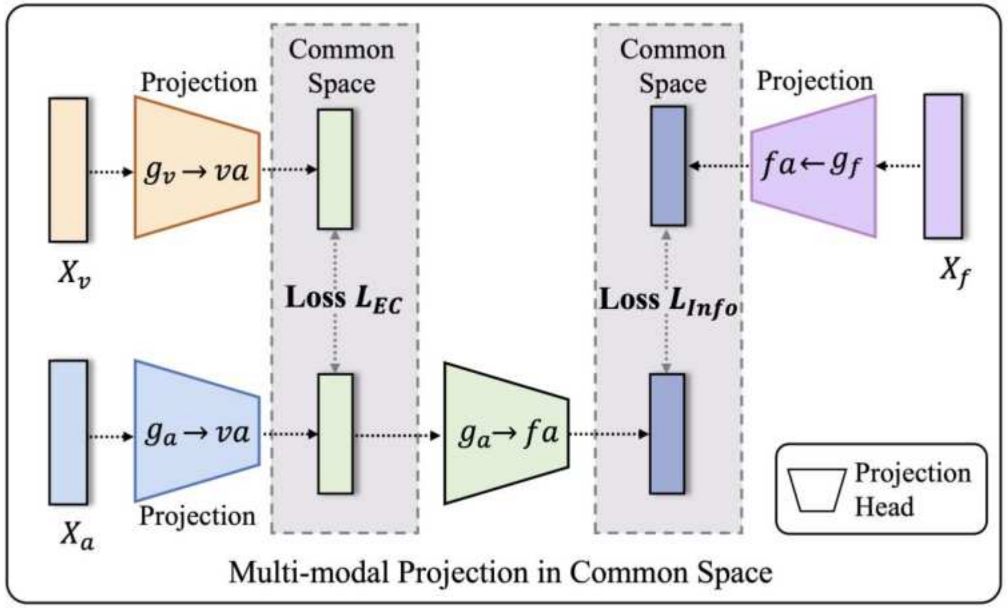
相关性损失L_{cor}：

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

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

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

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：

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

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

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	**99.7**	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	**99.9**	99.2	93.5	95.7	96.2
SST [54]	V-A	Self-Sup	**99.2**	**98.5**	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	**99.2**	**98.5**	**96.3**	**99.9**	**99.7**	**96.4**	**98.9**	**99.0**

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]	**98.9**	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	**99.7**	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
NPVForensics	98.5	**99.8**	**97.8**	**94.7**	**96.2**	**94.5**	**98.5**	99.4	**95.8**	**98.2**	**98.6**	**98.0**

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	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	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**

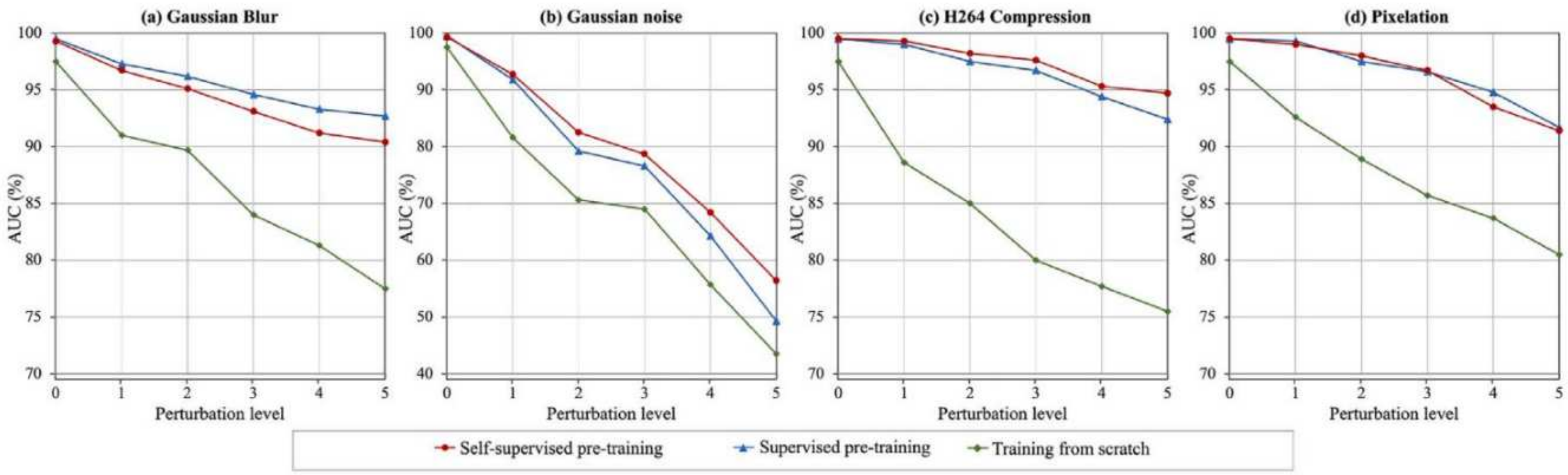
工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

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	**99.1**	49.8	63.8	88.6	55.2	75.7
LipForensics [19]	**99.9**	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
NPVForensics	99.7	97.9	**79.7**	**94.3**	**96.2**	**97.6**	**94.2**



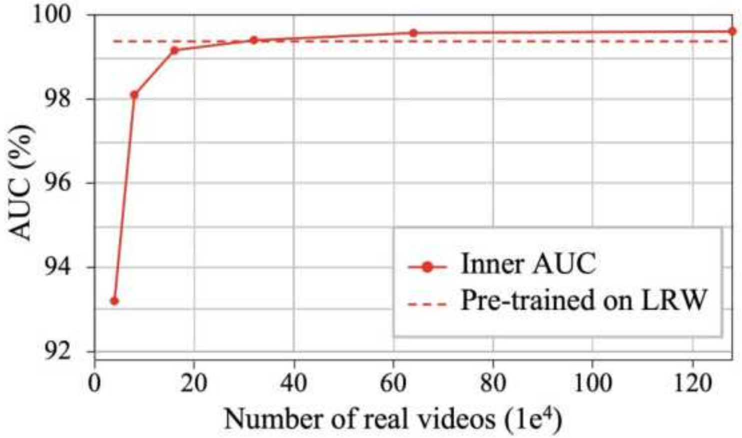
工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

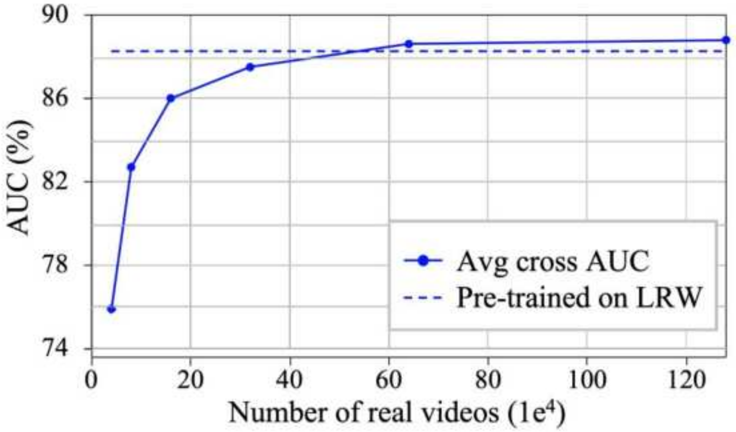
实验结果

（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
NPVForensics (Ours)	**96.7**	**97.8**	**96.3**	**96.9**

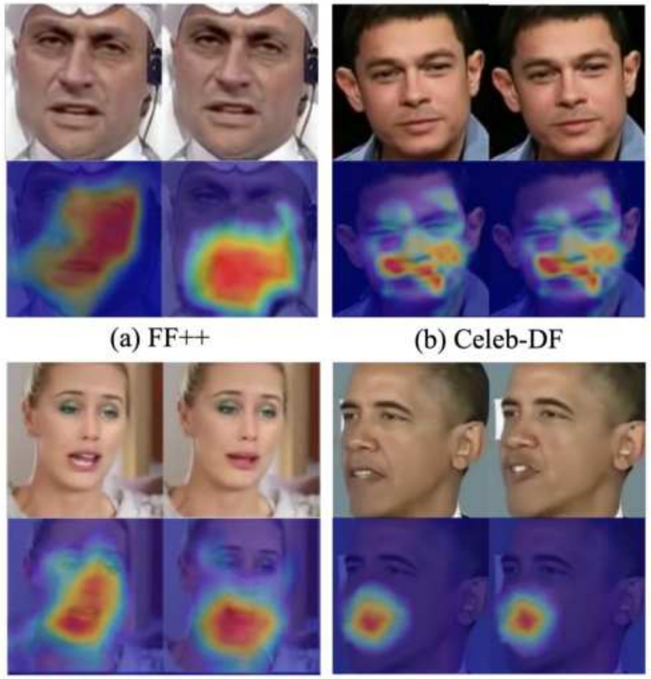




工作2：基于非关键音素-视素区域VA相关性学习的Deepfake检测

实验结果

（六）可视化



(a) FF++ (b) Celeb-DF (c) FakeAVCeleb (d) A2V

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！

Towards General AIGC image Detection in Open-World Scenarios

面向开放世界的深度伪造图像通用检测研究

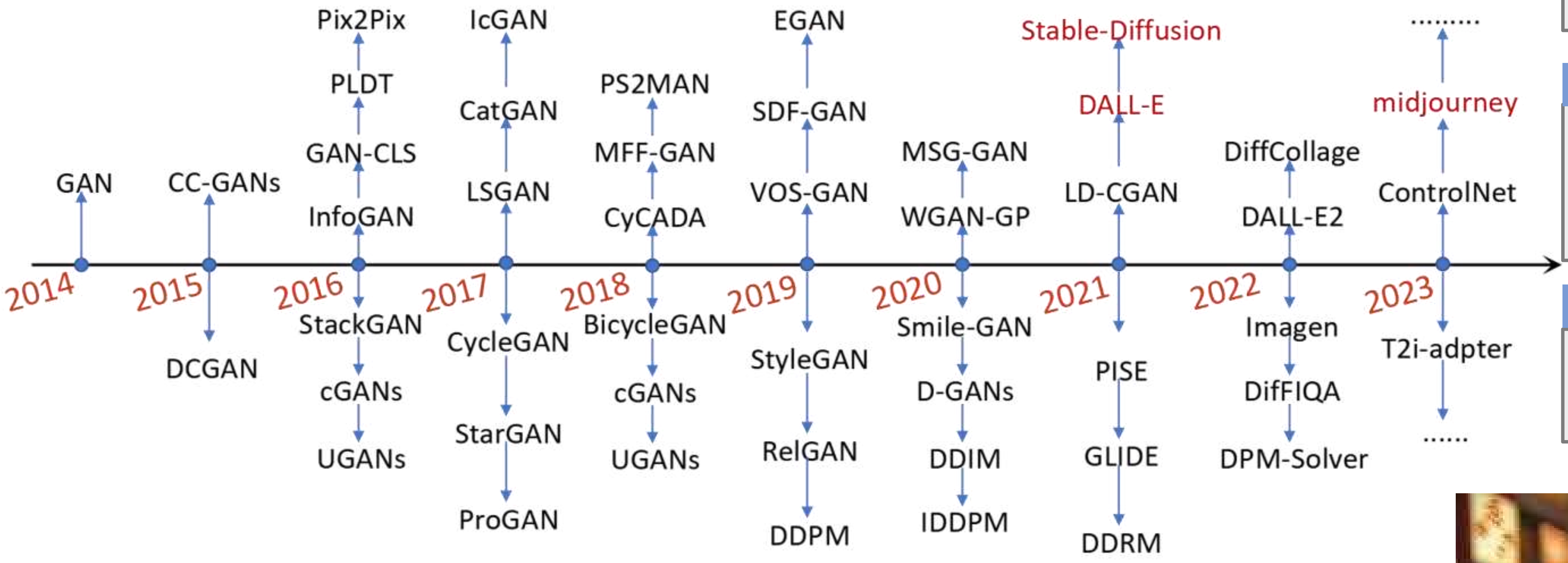
报 告 人：谭创创

汇报题纲

CONTENTS

1 深伪检测任务

AIGC发展迅猛，引领创作热潮



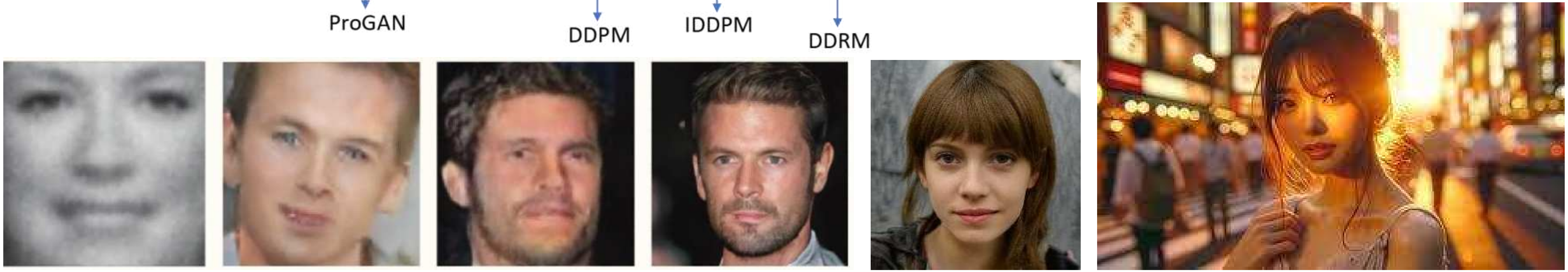
社交媒体

文学创作

医疗健康

金 融

教 育



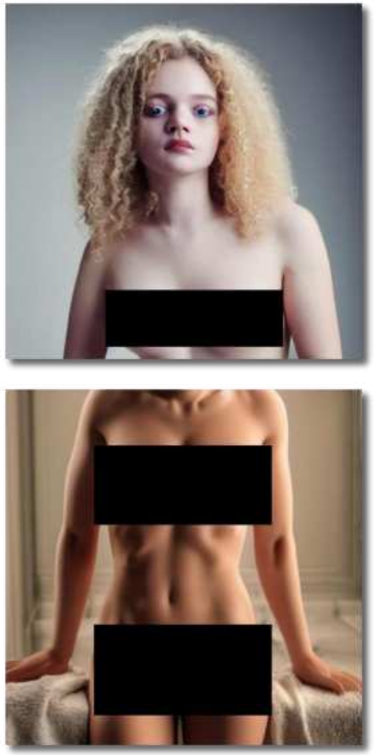
1 深伪检测任务

生成潜力被释放，同时伴随危害

眼见不再为实

种族偏见

色情



政治谣言

信息欺诈

◆ 深度伪造检测任务旨在帮助用户更好地抵御深度伪造信息带来的风险。

伪造检测

用户 → 被动防御

防范风险 →

真 / 假

为什么假

对抗生成流程,扩散流程
NeRF渲染流程,自回归流程

生成方法多种多样

生成内容复杂多变

剖析伪造机理

解译内容逻辑

准确性

泛化性

可解释性

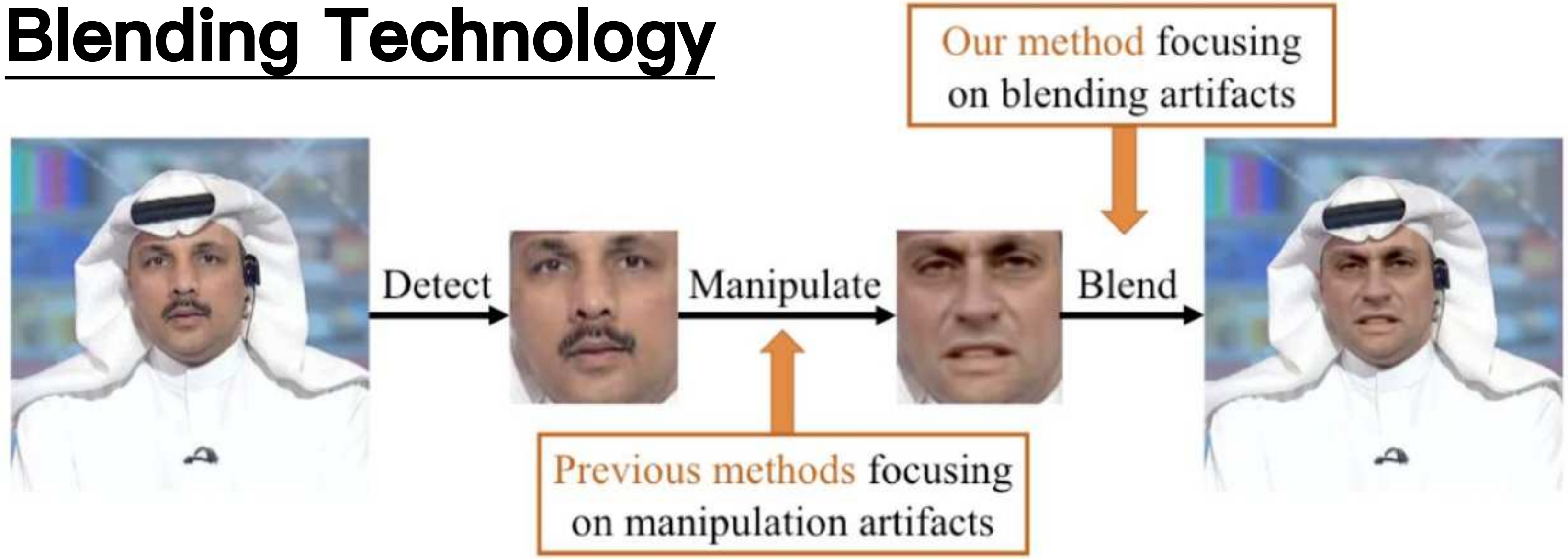
2 深伪检测研究现状

伪造方法 → 伪造数据 → 传播途径 →

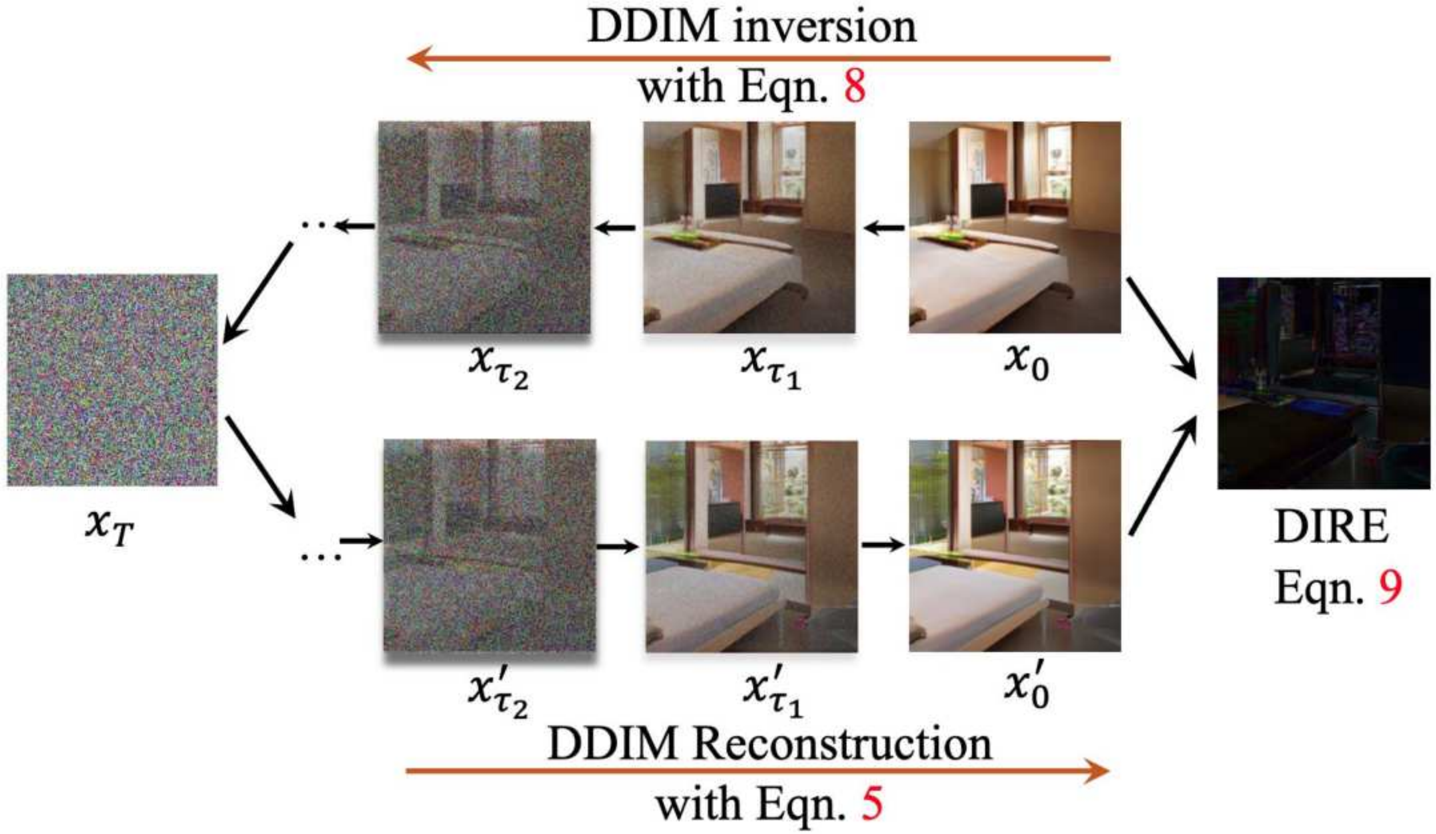
鉴伪模型

伪造方法 → 伪造数据 → 传播途径 → 鉴伪模型

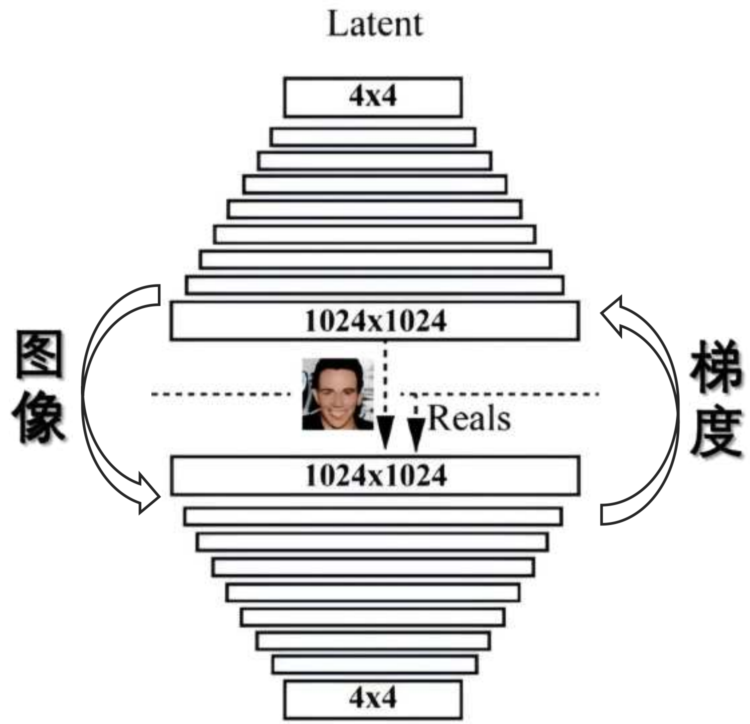
Blending Technology



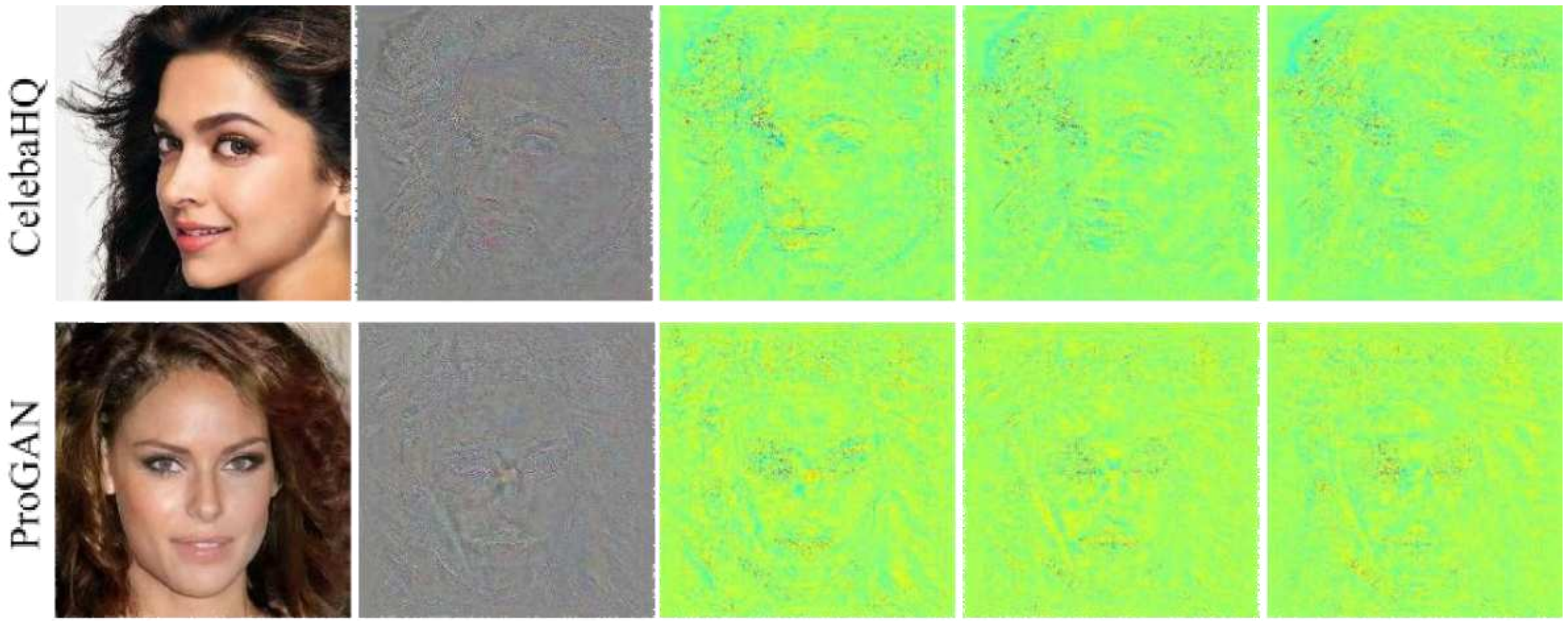
Face X-ray for More General Face Forgery Detection， CVPR2020



DIRE for Diffusion-Generated Image Detection，ICCV2023

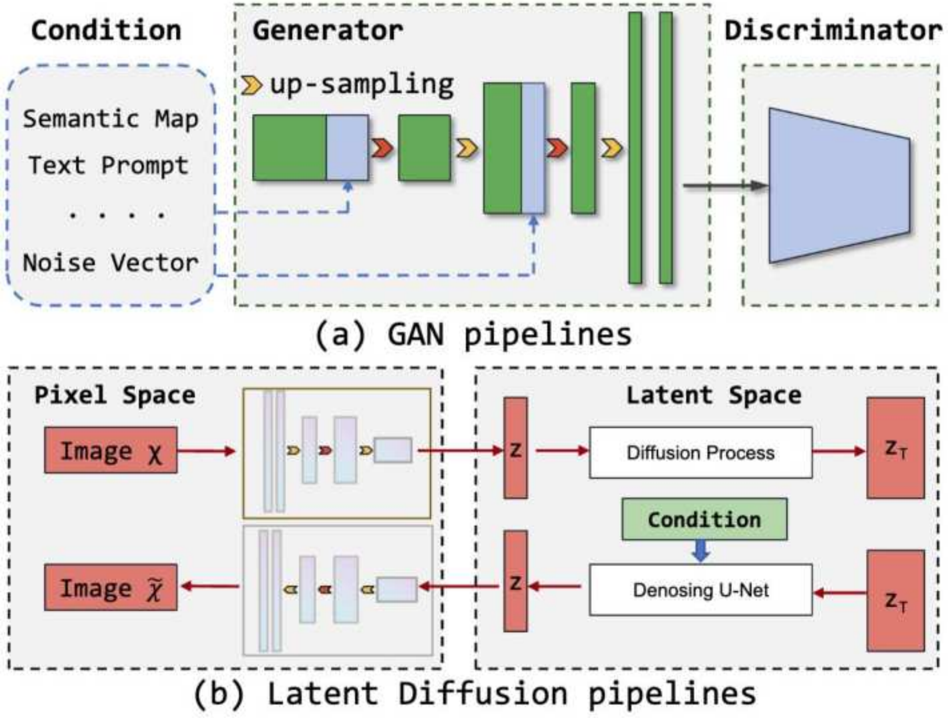


对抗生成网络结构

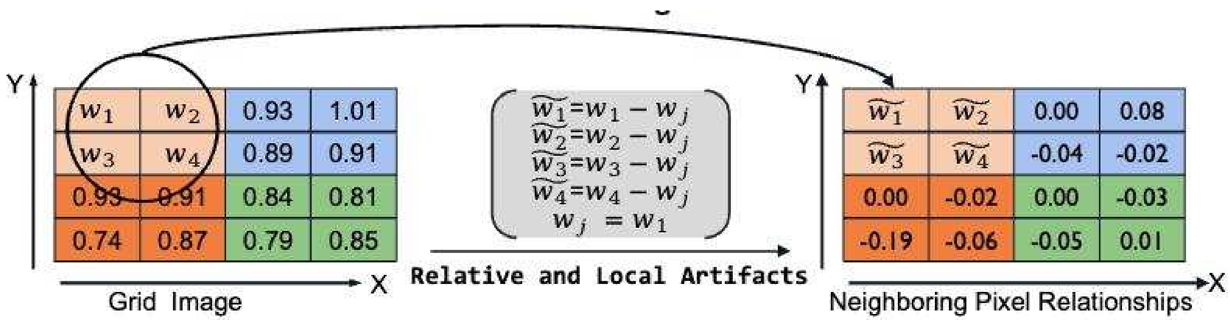


Learning on gradients: Generalized artifacts representation for gan-generated images detection, CVPR2023

伪造方法 → 伪造数据 → 传播途径 → 鉴伪模型



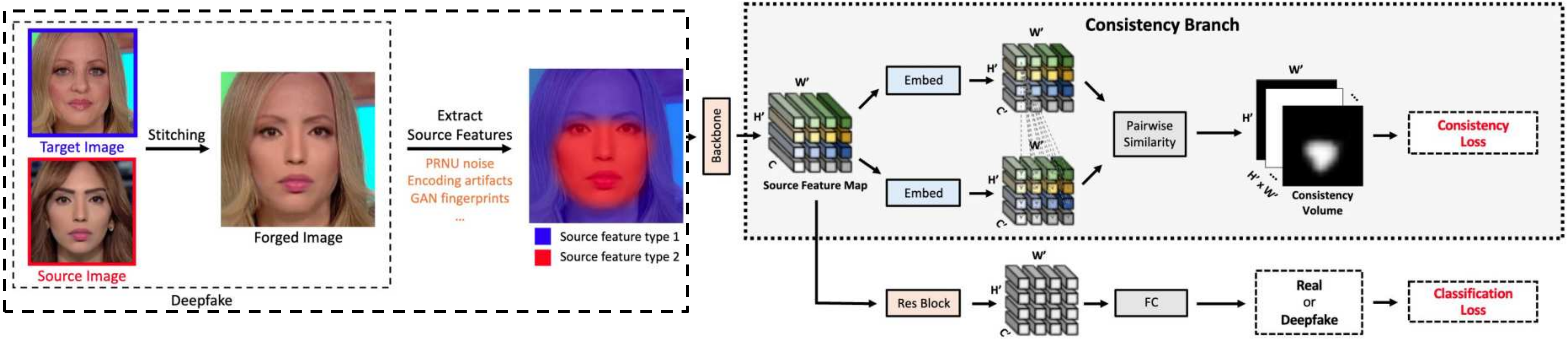
Neighboring Pixel Relationships.



Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection, CVPR2024

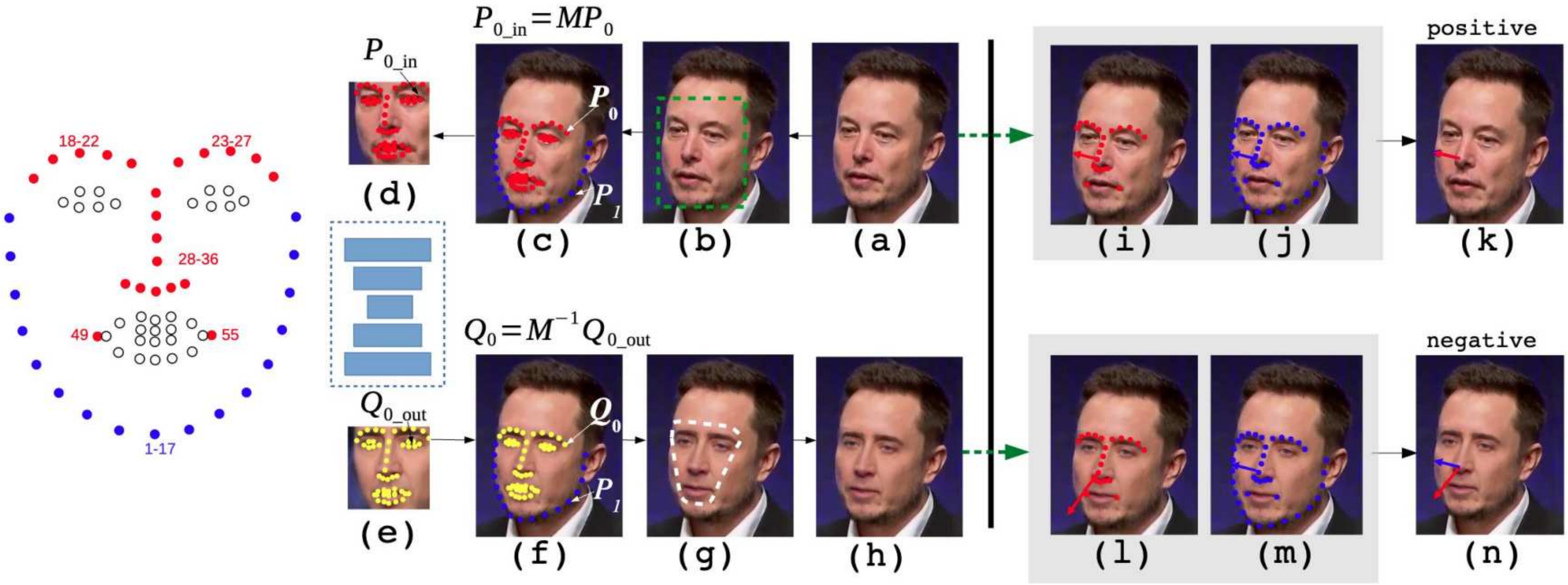
2 深伪检测研究现状

Self-Consistency



Learning Self-Consistency for Deepfake Detection，ICCV2021

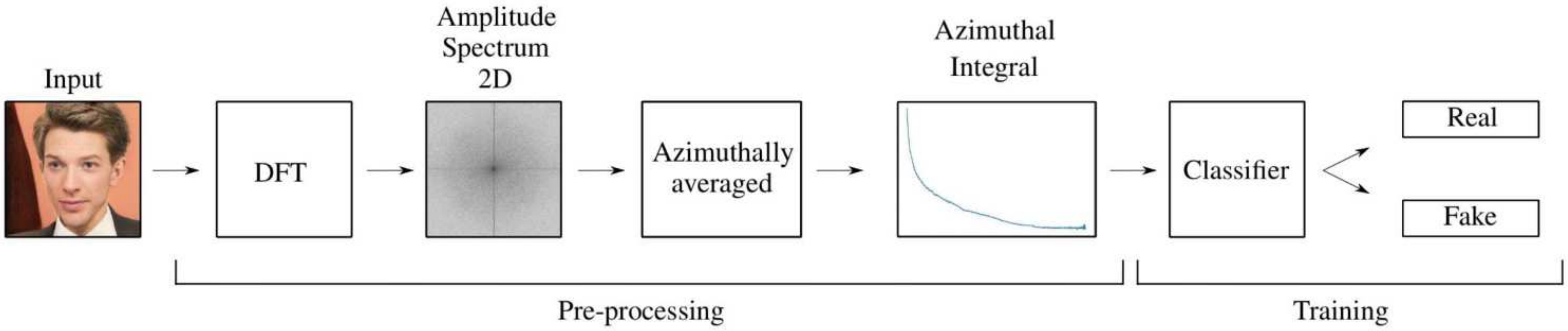
伪造方法 → 伪造数据 → 传播途径 → 鉴伪模型



Exposing deep fakes using inconsistent head poses，ICASSP2019

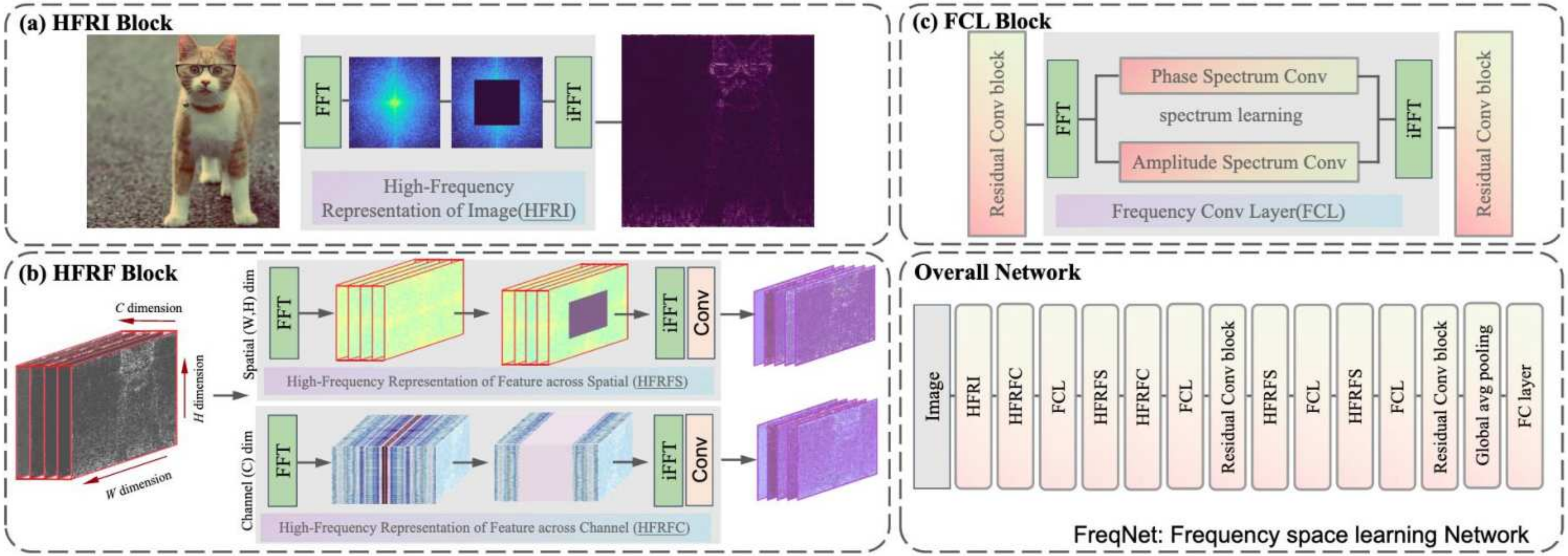
Frequency–based

FFHQ Spectrum　FFHQ　StyleGAN　StyleGAN Spectrum



Leveraging frequency analysis for deep fake image recognition，ICML2020

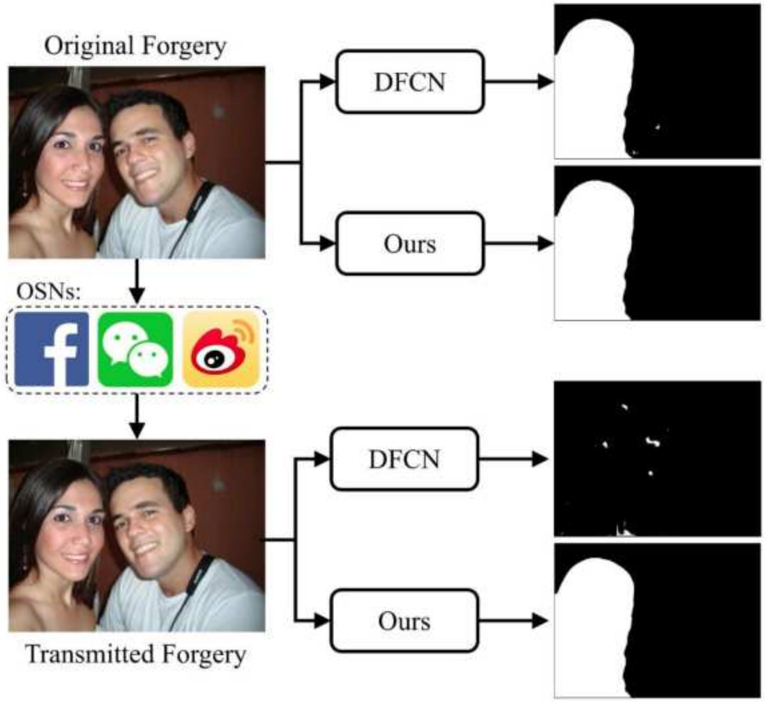
Frequency–based

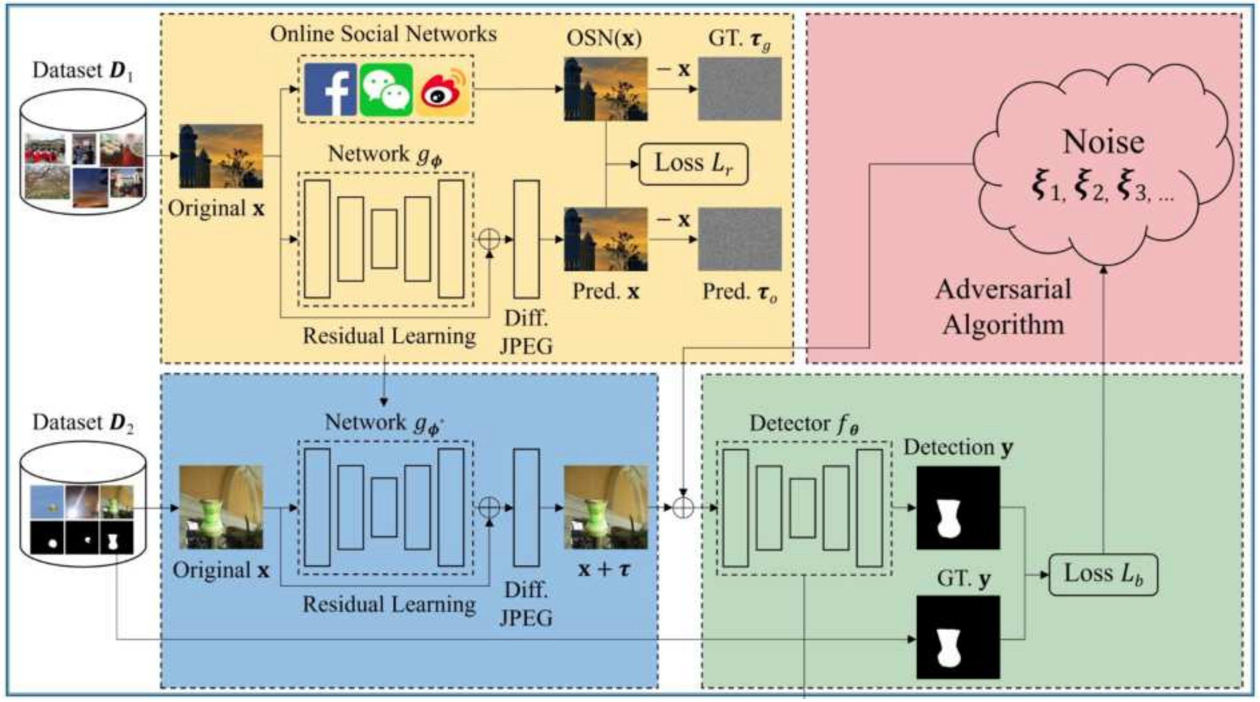


Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning， AAAI2024

2 深伪检测研究现状

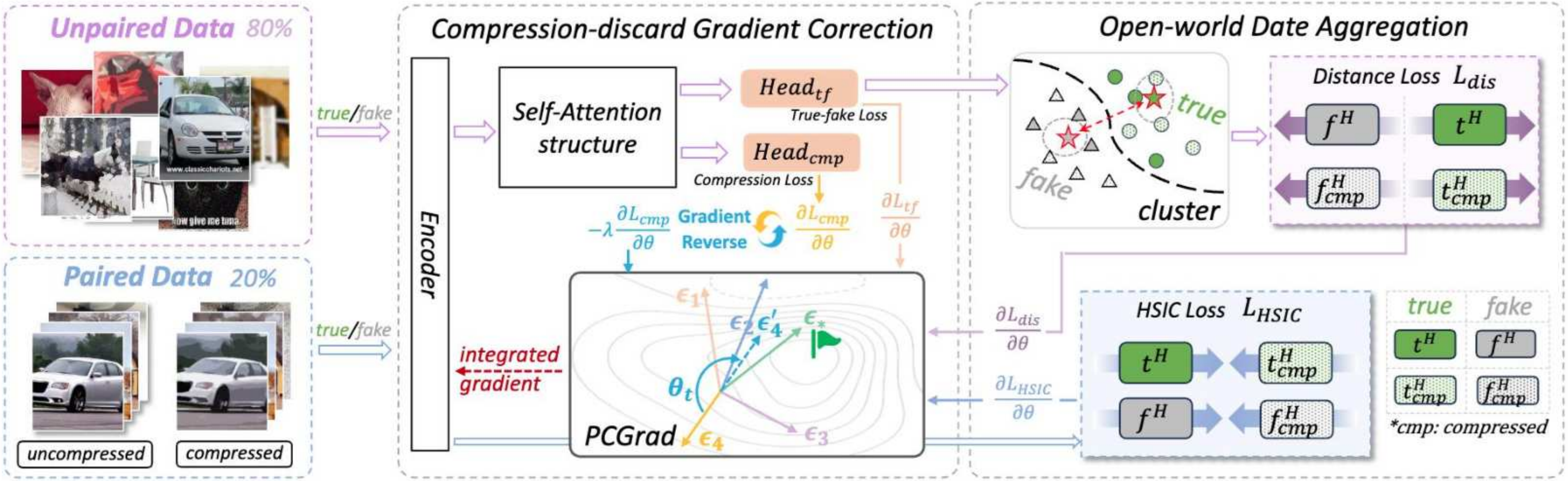
伪造方法 → 伪造数据 → 传播途径 → 鉴伪模型



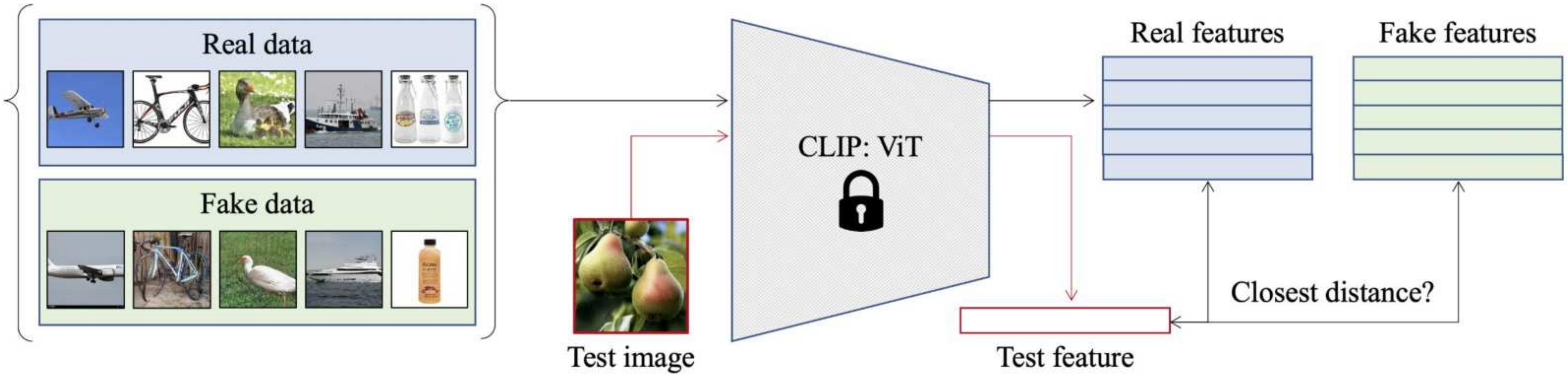


Robust Image Forgery Detection over Online Social Network Shared Images, CVPR2022

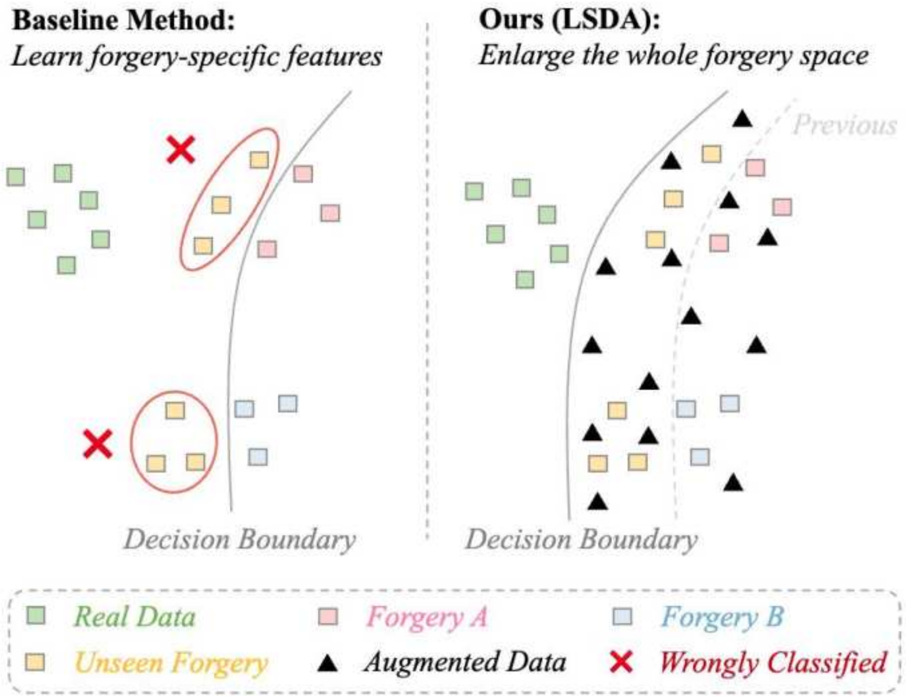
2 深伪检测研究现状

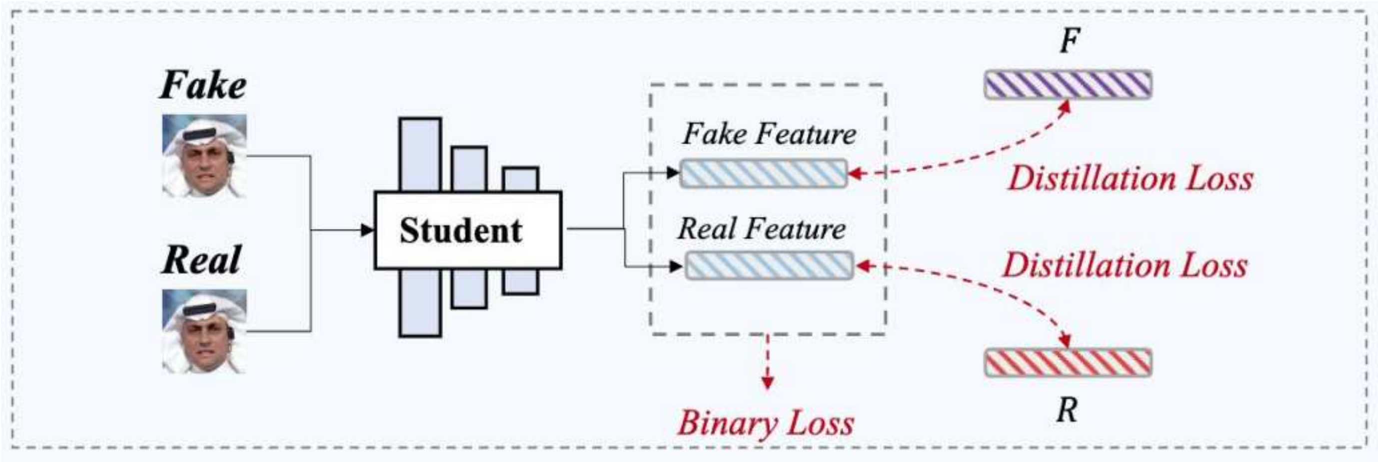


ODDN: Addressing Unpaired Data Challenges in Open-World Deepfake Detection on Online Social Networks, AAAI 2025



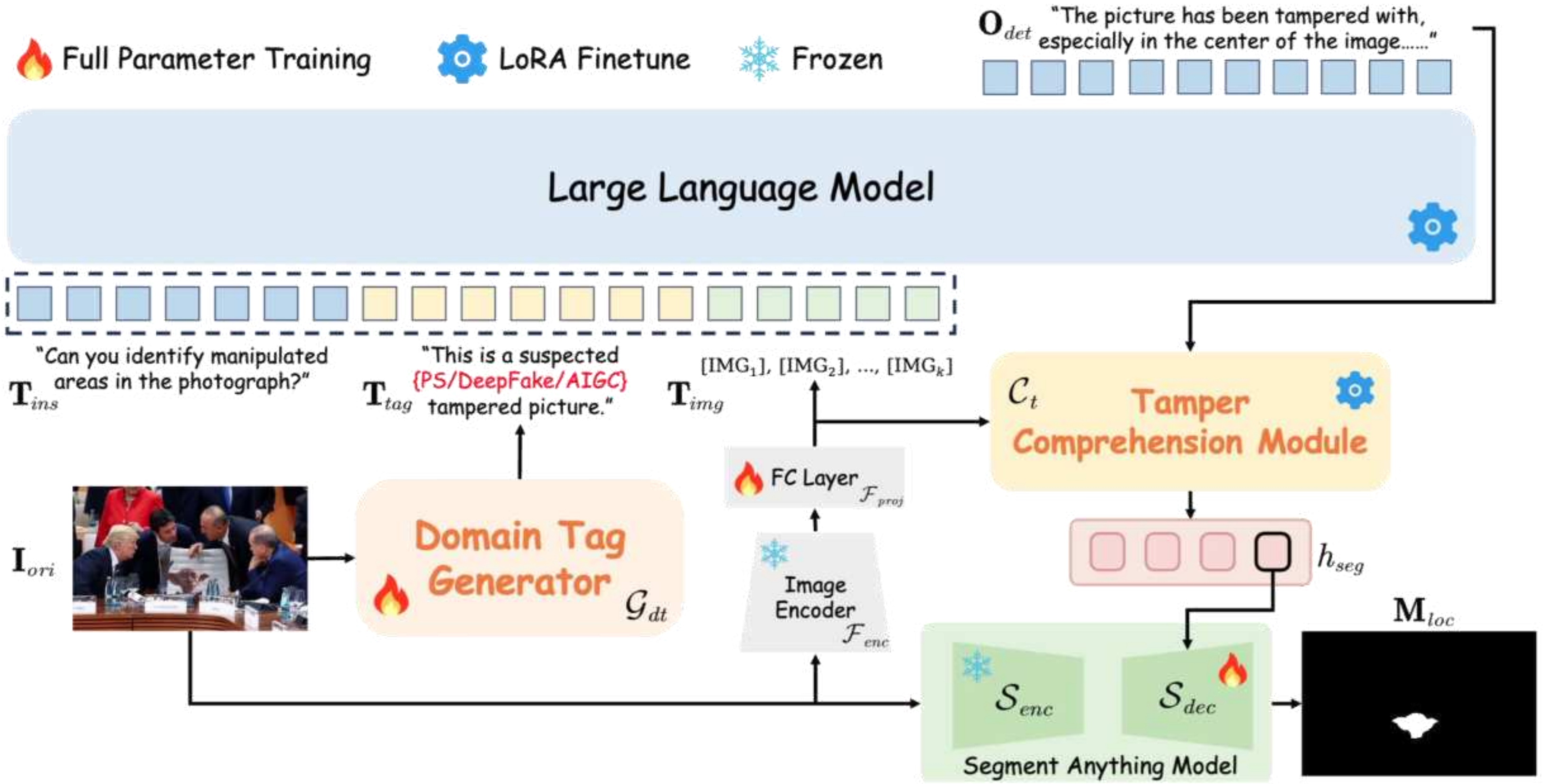
Towards Universal Fake Image Detectors that Generalize Across Generative Models, CVPR 2023





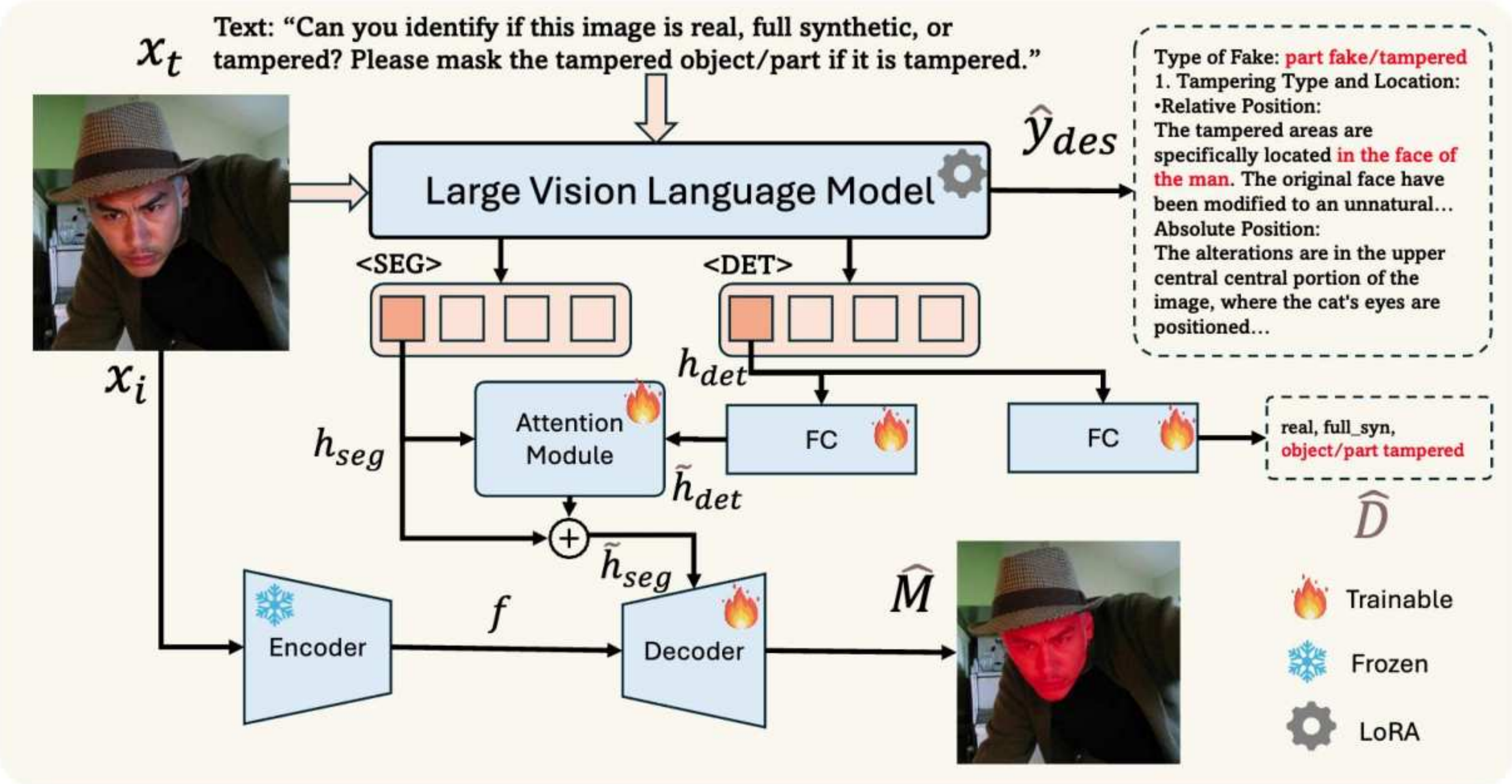
Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection, CVPR 2024

2 深伪检测研究现状



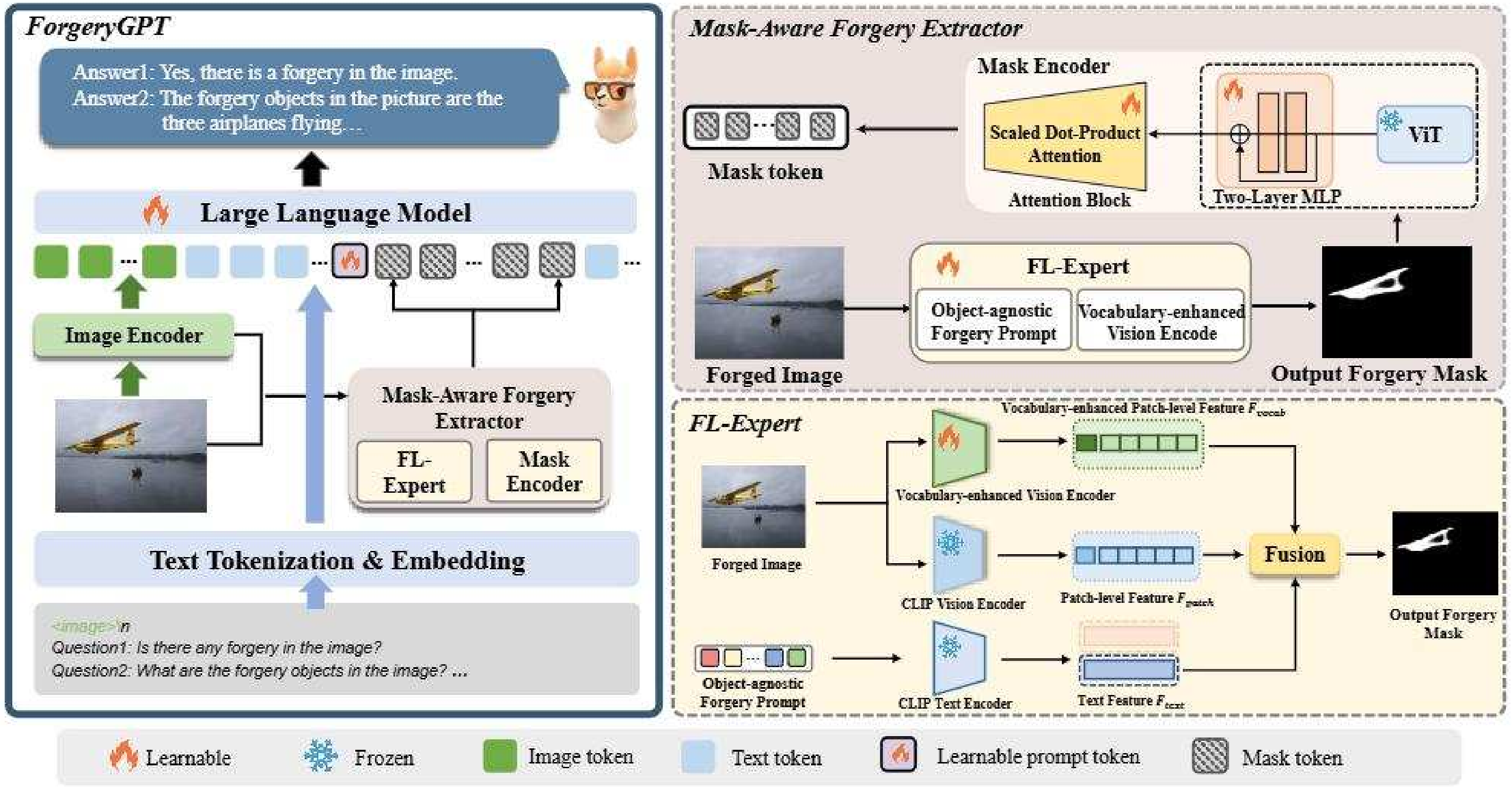
Fakeshield: Explainable image forgery detection and localization via multi-modal large language models, ICLR2025

2 深伪检测研究现状

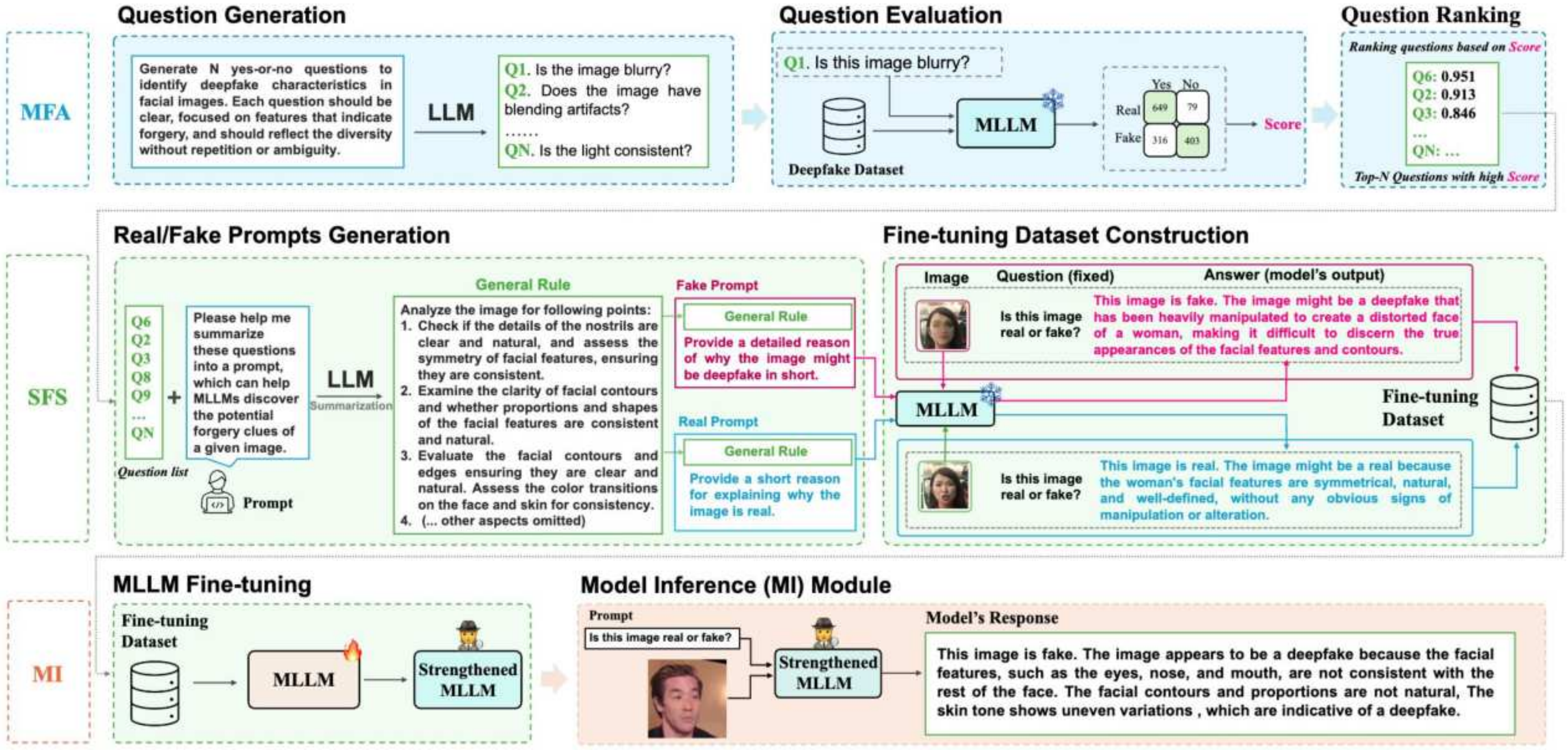


SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model, CVPR2025

2 深伪检测研究现状



ForgeryGPT: Multimodal Large Language Model For Explainable Image Forgery Detection and Localization



X 2-DFD: A framework for eplainable and etendable Deepfake Detection, arXiv2024

◆ 面向开放世界的深度伪造图像通用检测研究 ◆

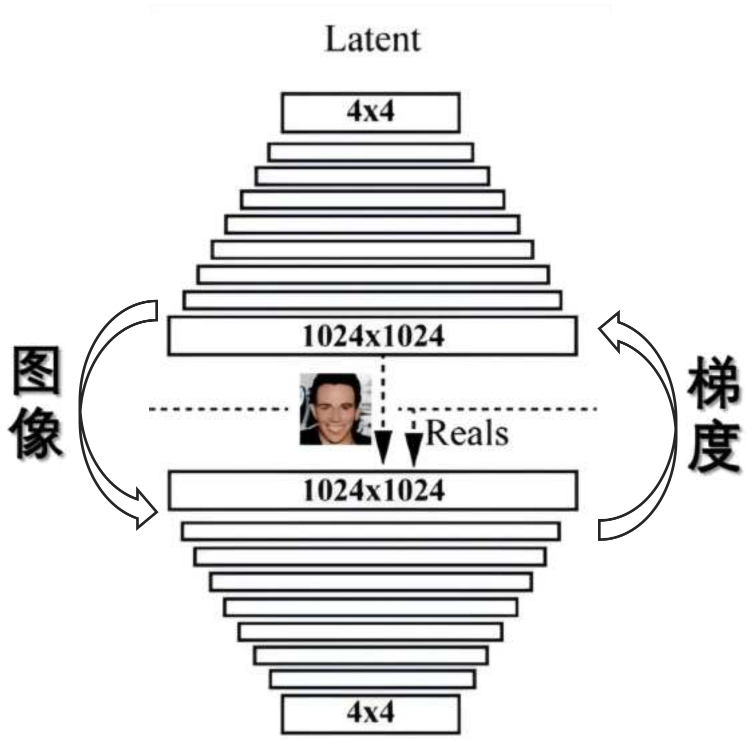
	伪造手段	伪造数据	深伪检测
研究对象	伪造手段	伪造数据	深伪检测
研究思路	剖析生成机理	归纳数据特性	解译伪造语义
研究成果	**首次挖掘**对抗梯度作为伪造痕迹，发表于CVPR2023	**首次提出**面向 AIGC的频域鉴伪方案，发表于AAAI2024	**首次揭示**预训练模型执行鉴伪机理，发表于AAAI2025
	首次提出AIGC通用伪影局部像素关系，发表于CVPR2024	提出频域信息指导预训练模型方案，发表于CVPR2024	

01

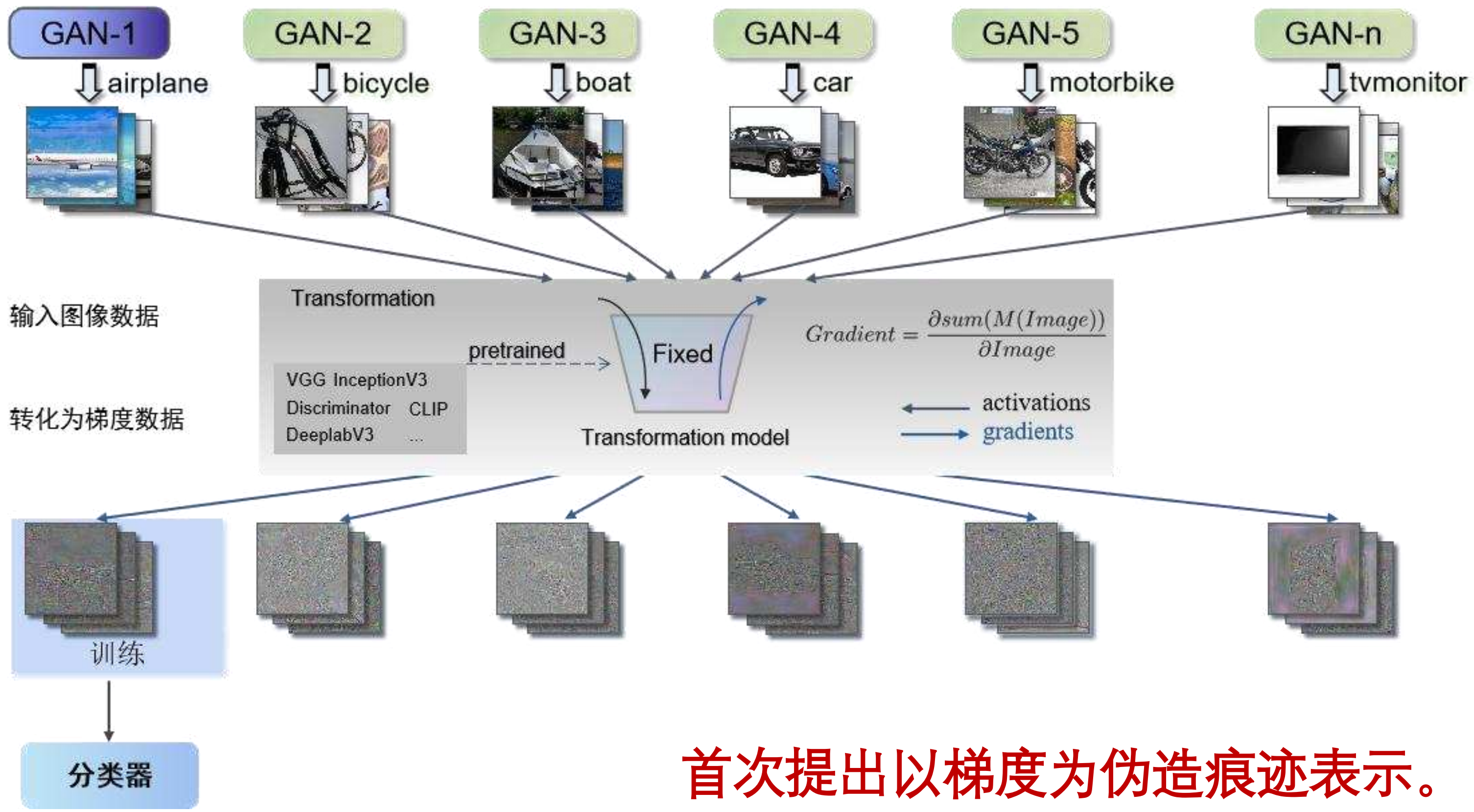
基于对抗梯度的GAN图像伪造检测方法

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Yunchao Wei: Learning on gradients: Generalized artifacts representation for gan-generated images detection. ***IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, pp. 12105-12114. 2023. **(CCF A类，计算机视觉领域顶级会议)**

- 现象：GAN伪造图像检测泛化性不足，即面对未知伪造源图像的检测性能较低；
- 问题：对抗生成网络是否含有通用性伪造痕迹？



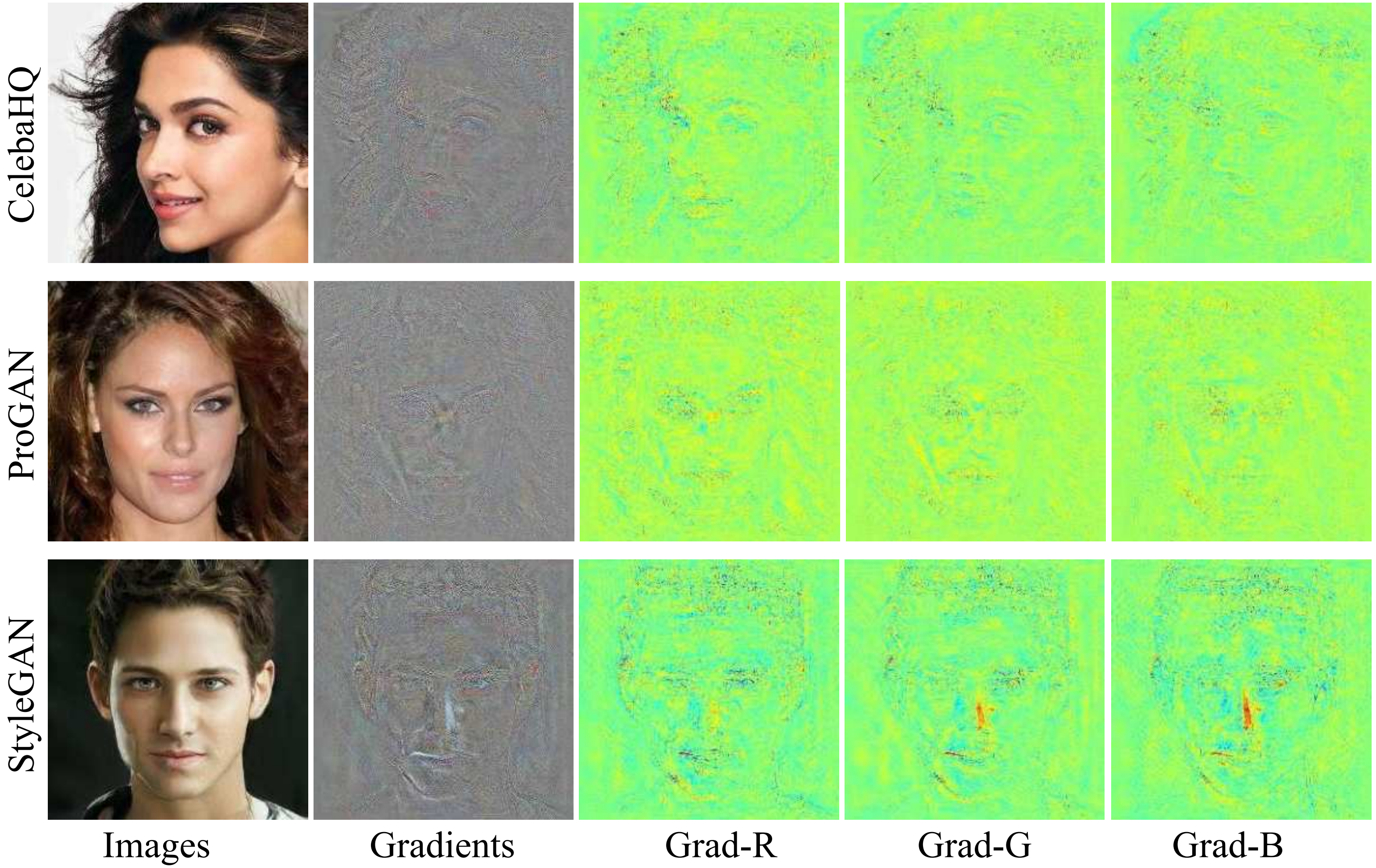
对抗生成网络结构



首次提出以梯度为伪造痕迹表示。

Tan Chuangchuang, Yao Zhao, Shikui Wei, Guanghua Gu, Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. **CVPR. 2023 (CCF-A).**

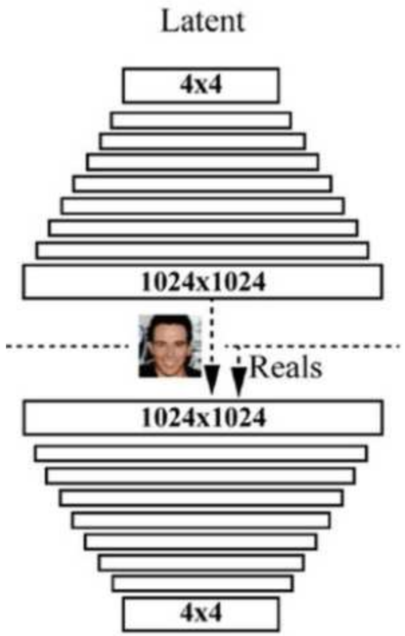
◻ 基于梯度的GAN伪影特征表示可视化



梯度特点：

1)突出对抗训练中学习的显著性信息，

2)去除图像内容，减少过拟合项.



Tan Chuangchuang, Yao Zhao, Shikui Wei, Guanghua Gu, Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. **CVPR. 2023 (CCF-A).**

北京交通大学

- 8种生成模型模拟开放世界，部分ProGAN数据作为训练数据; 检测性能提升11.4%。

Methods	Settings		Test Models																	
	Input	#class	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
			Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Wang	Image	1	50.4	63.8	50.4	79.3	68.2	94.7	50.2	61.3	50.0	52.9	50.0	48.2	50.3	67.6	50.1	51.5	52.5	64.9
Frank	Freq	1	78.9	77.9	69.4	64.8	67.4	64.0	62.3	58.6	67.4	65.4	60.5	59.5	67.5	69.1	52.4	47.3	65.7	63.3
Durall	Freq	1	85.1	79.5	59.2	55.2	70.4	63.8	57.0	53.9	66.7	61.4	99.8	99.6	58.7	54.8	53.0	51.9	68.7	65.0
BiHPF(WACV)	Freq	1	82.5	81.4	68.0	62.8	68.8	63.6	67.0	62.5	75.5	74.2	90.1	90.1	73.6	92.1	51.6	49.9	72.1	72.1
FrePGAN(AAAI22)	Image	1	95.5	99.4	80.6	90.6	77.4	93.0	63.5	60.5	59.4	59.9	99.6	100.0	53.0	49.1	70.4	81.5	74.9	79.3
LGrad(ProGAN-bedroom)	Grad	1	98.4	99.9	82.6	95.6	83.3	98.4	76.2	81.8	82.3	90.6	99.7	100.0	71.7	75.0	52.8	57.8	80.9	87.4
LGrad(StyleGAN-bedroom)	Grad	1	99.4	99.9	96.0	99.6	93.8	99.4	79.5	88.9	84.7	94.4	99.5	100.0	70.9	81.8	66.7	77.9	86.3(11.4↑)	92.7(13.4↑)
Wang	Image	4	91.4	99.4	63.8	91.4	76.4	97.5	52.9	73.3	72.7	88.6	63.8	90.8	63.9	92.2	51.7	62.3	67.1	86.9
Frank	Freq	4	90.3	85.2	74.5	72.0	73.1	71.4	88.7	86.0	75.5	71.2	99.5	99.5	69.2	77.4	60.7	49.1	78.9	76.5
Durall	Freq	4	81.1	74.4	54.4	52.6	66.8	62.0	60.1	56.3	69.0	64.0	98.1	98.1	61.9	57.4	50.2	50.0	67.7	64.4
BiHPF(WACV)	Freq	4	90.7	86.2	76.9	75.1	76.2	74.7	84.9	81.7	81.9	78.9	94.4	94.4	69.5	78.1	54.4	54.6	78.6	77.9
FrePGAN(AAAI22)	Image	4	99.0	99.9	80.7	89.6	84.1	98.6	69.2	71.1	71.1	74.4	99.9	100.0	60.3	71.7	70.9	91.9	79.4	87.2
LGrad(ProGAN-bedroom)	Grad	4	99.7	100.0	87.8	99.1	91.7	99.7	80.9	89.3	78.2	89.0	99.8	100.0	73.5	78.6	53.1	55.0	83.1	88.8
LGrad(StyleGAN-bedroom)	Grad	4	99.9	100.0	94.8	99.9	96.0	99.9	82.9	90.7	85.3	94.0	99.6	100.0	72.4	79.3	58.0	67.9	86.1	91.5

Tan Chuangchuang, Yao Zhao, Shikui Wei, Guanghua Gu, Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. **CVPR. 2023 (CCF-A).**

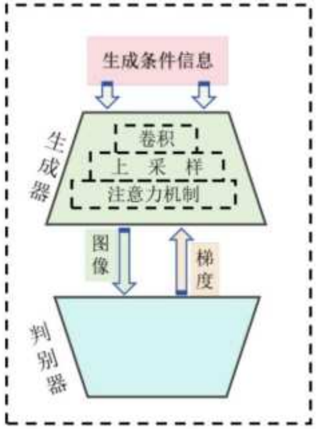
02

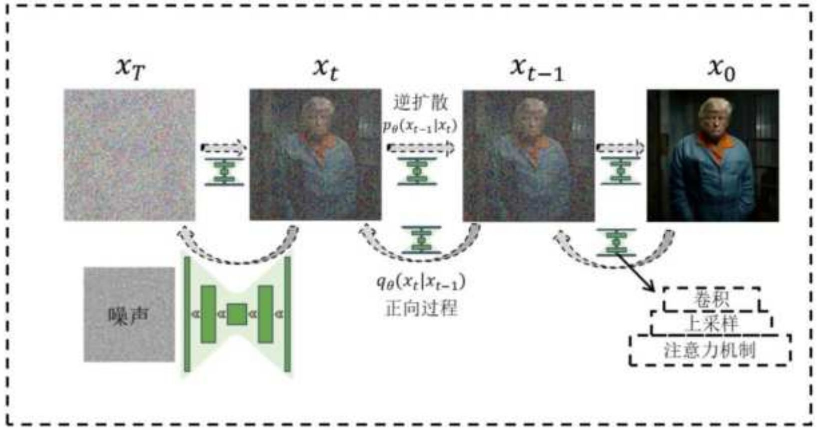
基于局部像素关联的通用伪造检测方法

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, Yunchao Wei: Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. ***IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, pp. 28130-28139. 2024. **(CCF A类，计算机视觉领域顶级会议)**

- 现象：GAN、Diffusion等伪造源层出不穷，检测器面对**未知伪造源**图像失效；
- 问题：生成网络是否含有**基因式伪造痕迹**？

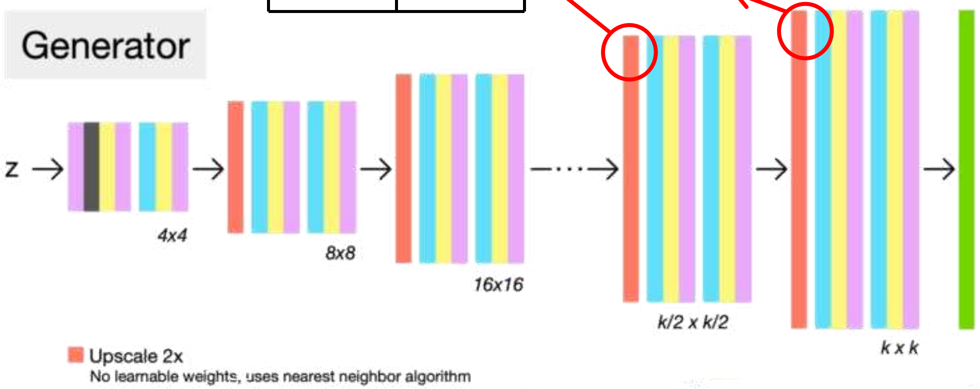
上采样广泛存在于典型生成方法中





0.3481	0.3481
0.3481	0.3481

0.5462	0.4679
0.4337	0.5022



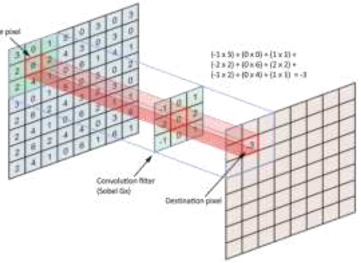
通用部件

通用操作

卷积

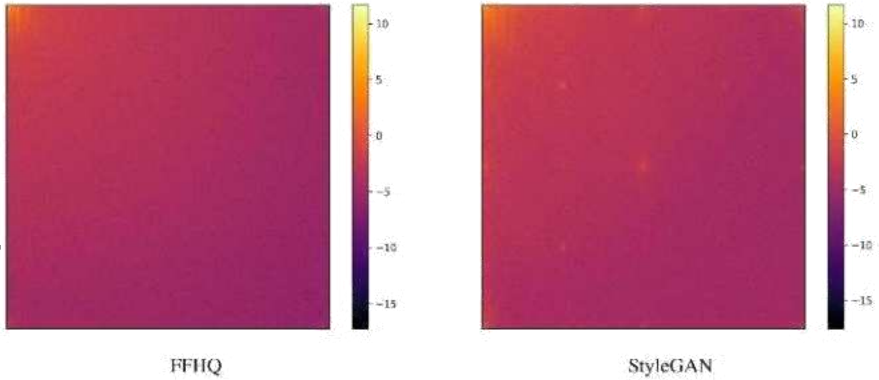
上 采 样

注意力机制



2020

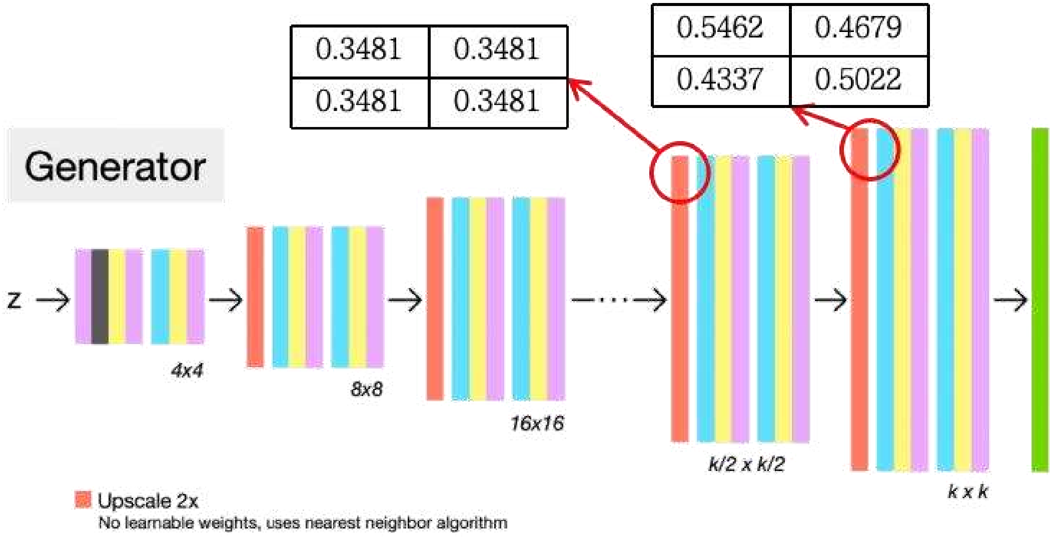
CVPR，ICML 工作提出频域特征作为上采样的伪影。

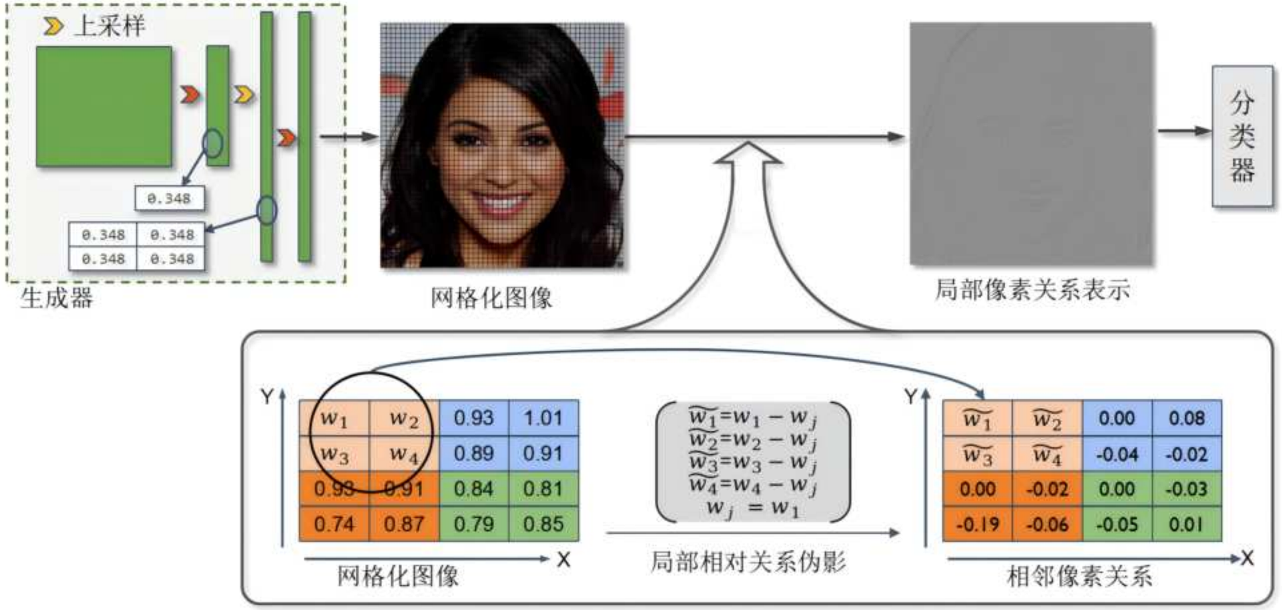


Tan Chuangchuang, Yao Zhao, S Wei, G Gu, P Liu, Y Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. **CVPR. 2024 (CCF-A).**

3.2 研究工作-局部像素关系伪造线索

- ❑ 现象：GAN、Diffusion等伪造源层出不穷，检测器面对**未知伪造源**图像失效；
- ❑ 问题：生成网络是否含有**基因式伪造痕迹**？





- ❑ 上采样与卷积结合考虑。
- ❑ 首次提出局部像素相关信息作为通用伪影表示。

Tan Chuangchuang, Yao Zhao, S Wei, G Gu, P Liu, Y Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. **CVPR. 2024 (CCF-A).**

3.2 研究工作-局部像素关系伪造线索

实验结果

方法	AttGAN		BEGAN		CramerGAN		InfoMaxGAN		MMDGAN		RelGAN		S3GAN		SNGAN		STGAN		平均	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
CNNDetection[99]	51.1	83.7	50.2	44.9	81.5	97.5	71.1	94.7	72.9	94.4	53.3	82.1	55.2	66.1	62.7	90.4	63.0	92.7	62.3	82.9
Frank[73]	65.0	74.4	39.4	39.9	31.0	36.0	41.1	41.0	38.4	40.5	69.2	96.2	69.7	81.9	48.4	47.9	25.4	34.0	47.5	54.7
Durall[74]	39.9	38.2	48.2	30.9	60.9	67.2	50.1	51.7	59.5	65.5	80.0	88.2	87.3	97.0	54.8	58.9	62.1	72.5	60.3	63.3
Patchfor[114]	68.0	92.9	97.1	100.0	97.8	99.9	93.6	98.2	97.9	100.0	99.6	100.0	66.8	68.1	97.6	99.8	92.7	99.8	90.1	95.4
F3Net[75]	85.2	94.8	87.1	97.5	89.5	99.8	67.1	83.1	73.7	99.6	98.8	100.0	65.4	70.0	51.6	93.6	60.3	99.9	75.4	93.1
SelfBland[144]	63.1	66.1	56.4	59.0	75.1	82.4	79.0	82.5	68.6	74.0	73.6	77.8	53.2	53.9	61.6	65.0	61.2	66.7	65.8	69.7
GANDetection[145]	57.4	75.1	67.9	100.0	67.8	99.7	67.6	92.4	67.7	99.3	60.9	86.2	69.6	83.5	66.7	90.6	69.6	97.2	66.1	91.6
LGrad[83]	68.6	93.8	69.9	89.2	50.3	54.0	71.1	82.0	57.5	67.3	89.1	99.1	78.5	86.0	78.0	87.4	54.8	68.0	68.6	80.8
UniFD[86]	78.5	98.3	72.0	98.9	77.6	99.8	77.6	98.9	77.6	99.7	78.2	98.7	85.2	98.1	77.6	98.7	74.2	97.8	77.6	98.8
NPR(本章所提方法)	83.0	96.2	99.0	99.8	98.7	99.0	94.5	98.3	98.6	99.0	99.6	100.0	79.0	80.0	88.8	97.4	98.0	100.0	93.2	96.6

方法	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		平均	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
CNNDetection[99]	91.4	99.4	63.8	91.4	76.4	97.5	52.9	73.3	72.7	88.6	63.8	90.8	63.9	92.2	51.7	62.3	67.1	86.9
Frank[73]	90.3	85.2	74.5	72.0	73.1	71.4	88.7	86.0	75.5	71.2	99.5	99.5	69.2	77.4	60.7	49.1	78.9	76.5
Durall[74]	81.1	74.4	54.4	52.6	66.8	62.0	60.1	56.3	69.0	64.0	98.1	98.1	61.9	57.4	50.2	50.0	67.7	64.4
Patchfor[114]	97.8	100.0	82.6	93.1	83.6	98.5	64.7	69.5	74.5	87.2	100.0	100.0	57.2	55.4	85.0	93.2	80.7	87.1
F3Net[75]	99.4	100.0	92.6	99.7	88.0	99.8	65.3	69.9	76.4	84.3	100.0	100.0	58.1	56.7	63.5	78.8	80.4	86.2
SelfBland[144]	58.8	65.2	50.1	47.7	48.6	47.4	51.1	51.9	59.2	65.3	74.5	89.2	59.2	65.5	93.8	99.3	61.9	66.4
GANDetection[145]	82.7	95.1	74.4	92.9	69.9	87.9	76.3	89.9	85.2	95.5	68.8	99.7	61.4	75.8	60.0	83.9	72.3	90.1
BiHPF[90]	90.7	86.2	76.9	75.1	76.2	74.7	84.9	81.7	81.9	78.9	94.4	94.4	69.5	78.1	54.4	54.6	78.6	77.9
FrePGAN[81]	99.0	99.9	80.7	89.6	84.1	98.6	69.2	71.1	71.1	74.4	99.9	100.0	60.3	71.7	70.9	91.9	79.4	87.2
LGrad[83]	99.9	100.0	94.8	99.9	96.0	99.9	82.9	90.7	85.3	94.0	99.6	100.0	72.4	79.3	58.0	67.9	86.1	91.5
UniFD[86]	99.7	100.0	89.0	98.7	83.9	98.4	90.5	99.1	87.9	99.8	91.4	100.0	89.9	100.0	80.2	90.2	89.1	98.3
NPR(本章所提方法)	99.8	100.0	96.3	99.8	97.3	100.0	87.5	94.5	95.0	99.5	99.7	100.0	86.6	88.8	77.4	86.2	92.5	96.1

GAN 图像检测性能：

- 只在ProGAN上训练；
- 在17种GAN上实现泛化检测，ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, Deepfake, AttGAN, BEGAN, CramerGAN, InfoMaxGAN, MMDGAN, RelGAN, S3GAN, SNGAN, STGAN；
- 平均正确率达到92.8%。

Tan Chuangchuang, Yao Zhao, S Wei, G Gu, P Liu, Y Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. **CVPR. 2024 (CCF-A).**

3.2 研究工作-局部像素关系伪造线索

实验结果

方法	ADM		DDPM		IDDPM		LDM		PNDM		VQ-Diffusion		Stable Diffusion v1		Stable Diffusion v2		平均	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
CNNDetection[99]	53.9	71.8	62.7	76.6	50.2	82.7	50.4	78.7	50.8	90.3	50.0	71.0	38.0	76.7	52.0	90.3	51.0	79.8
Frank[73]	58.9	65.9	37.0	27.6	51.4	65.0	51.7	48.5	44.0	38.2	51.7	66.7	32.8	52.3	40.8	37.5	46.0	50.2
Durall[74]	39.8	42.1	52.9	49.8	55.3	56.7	43.1	39.9	44.5	47.3	38.6	38.3	39.5	56.3	62.1	55.8	47.0	48.3
Patchfor[114]	77.5	93.9	62.3	97.1	50.0	91.6	99.5	100.0	50.2	99.9	100.0	100.0	90.7	99.8	94.8	100.0	78.1	97.8
F3Net[75]	80.9	96.9	84.7	99.4	74.7	98.9	100.0	100.0	72.8	99.5	100.0	100.0	73.4	97.2	99.8	100.0	85.8	99.0
SelfBland[144]	57.0	59.0	61.9	49.6	63.2	66.9	83.3	92.2	48.2	48.2	77.2	82.7	46.2	68.0	71.2	73.9	63.5	67.6
GANDetection[145]	51.1	53.1	62.3	46.4	50.2	63.0	51.6	48.1	50.6	79.0	51.1	51.2	39.8	65.6	50.1	36.9	50.8	55.4
LGrad[83]	86.4	97.5	99.9	100.0	66.1	92.8	99.7	100.0	69.5	98.5	96.2	100.0	90.4	99.4	97.1	100.0	88.2	98.5
UniFD[86]	78.4	92.1	72.9	78.8	75.0	92.8	82.2	97.1	75.3	92.5	83.5	97.7	56.4	90.4	71.5	92.4	74.4	91.7
NPR(本章所提方法)	88.6	98.9	99.8	100.0	91.8	99.8	100.0	100.0	91.2	100.0	100.0	100.0	97.4	99.8	93.8	100.0	95.3	99.8

方法	DALLE		Glide_100_10		Glide_100_27		Glide_50_27		ADM		LDM_100		LDM_200		LDM_200_cfg		平均	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
CNNDetection[99]	51.8	61.3	53.3	72.9	53.0	71.3	54.2	76.0	54.9	66.6	51.9	63.7	52.0	64.5	51.6	63.1	52.8	67.4
Frank[73]	57.0	62.5	53.6	44.3	50.4	40.8	52.0	42.3	53.4	52.5	56.6	51.3	56.4	50.9	56.5	52.1	54.5	49.6
Durall[74]	55.9	58.0	54.9	52.3	48.9	46.9	51.7	49.9	40.6	42.3	62.0	62.6	61.7	61.7	58.4	58.5	54.3	54.0
Patchfor[114]	79.8	99.1	87.3	99.7	82.8	99.1	84.9	98.8	74.2	81.4	95.8	99.8	95.6	99.9	94.0	99.8	86.8	97.2
F3Net[75]	71.6	79.9	88.3	95.4	87.0	94.5	88.5	95.4	69.2	70.8	74.1	84.0	73.4	83.3	80.7	89.1	79.1	86.5
SelfBland[144]	52.4	51.6	58.8	63.2	59.4	64.1	64.2	68.3	58.3	63.4	53.0	54.0	52.6	51.9	51.9	52.6	56.3	58.7
GANDetection[145]	67.2	83.0	51.2	52.6	51.1	51.9	51.7	53.5	49.6	49.0	54.7	65.8	54.9	65.9	53.8	58.9	54.3	60.1
LGrad[83]	88.5	97.3	89.4	94.9	87.4	93.2	90.7	95.1	86.6	100.0	94.8	99.2	94.2	99.1	95.9	99.2	90.9	97.2
UniFD[86]	89.5	96.8	90.1	97.0	90.7	97.2	91.1	97.4	75.7	85.1	90.5	97.0	90.2	97.1	77.3	88.6	86.9	94.5
NPR (本章所提方法)	94.5	99.5	98.2	99.8	97.8	99.7	98.2	99.8	75.8	81.0	99.3	99.9	99.1	99.9	99.0	99.8	95.2	97.4

Diffusion图像检测性能：

- 只在ProGAN上训练；
- 在11种Diffusion上实现泛化检测，DDPM, IDDPM, ADM, LDM, PNDM, VQDiffusion, Glide, Stable Diffusion v1, Stable Diffusion v2, DALLE, and Midjourney；
- 平均正确率达到95.3%。

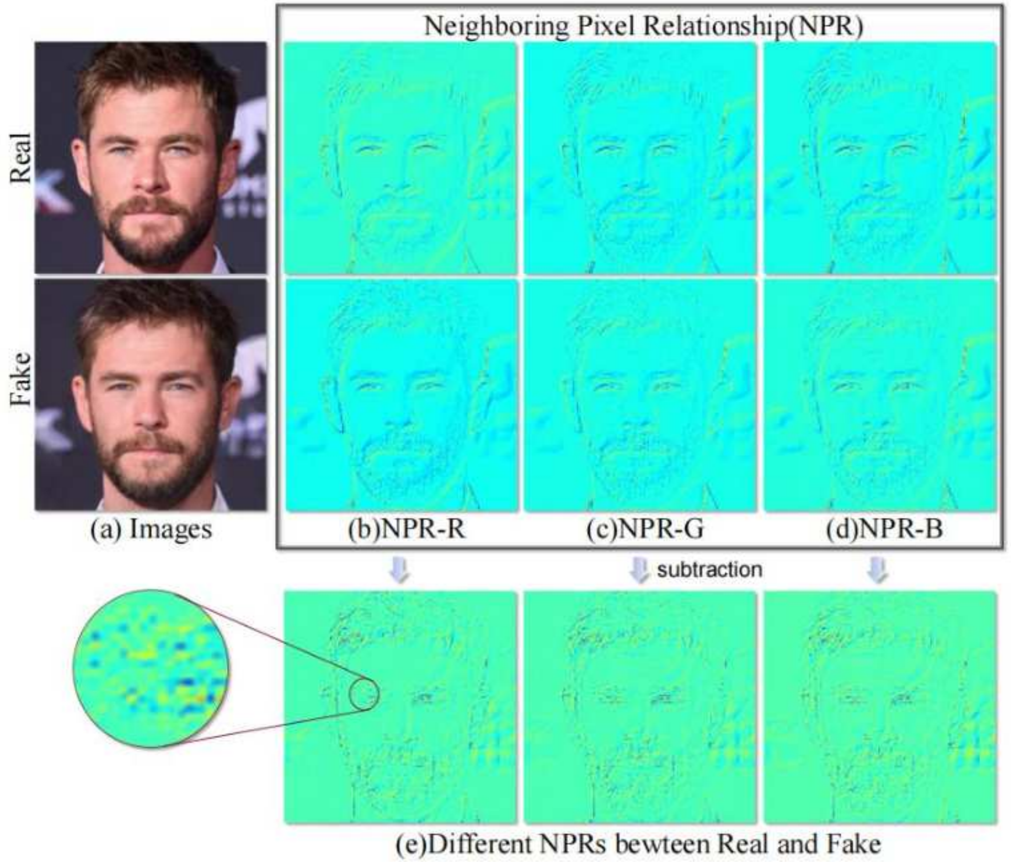
Tan Chuangchuang, Yao Zhao, S Wei, G Gu, P Liu, Y Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. **CVPR. 2024 (CCF-A).**

实验结果

- 在28种生成模型、38个子数据集的开放世界下验证其深伪检测的泛化性和有效性；

ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, Deepfake, AttGAN, BEGAN, CramerGAN, InfoMaxGAN, MMDGAN, RelGAN, S3GAN, SNGAN, STGAN, DDPM, IDDPM, ADM, LDM, PNDM, VQDiffusion, Glide, Stable Diffusion v1, Stable Diffusion v2, DALLE, and Midjourney.

方法	38 个子测试集的平均准确率
CNNDetection[99]	57.3
Frank[73]	56.8
Durall[74]	56.6
Patchfor[114]	80.6
F3Net[75]	78.1
SelfBland[144]	61.2
GANDetection[145]	59.5
LGrad[83]	80.5
UniFD[86]	79.8
NPR (本章所提方法)	92.2



Tan Chuangchuang, Yao Zhao, S Wei, G Gu, P Liu, Y Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. **CVPR. 2024 (CCF-A).**

03

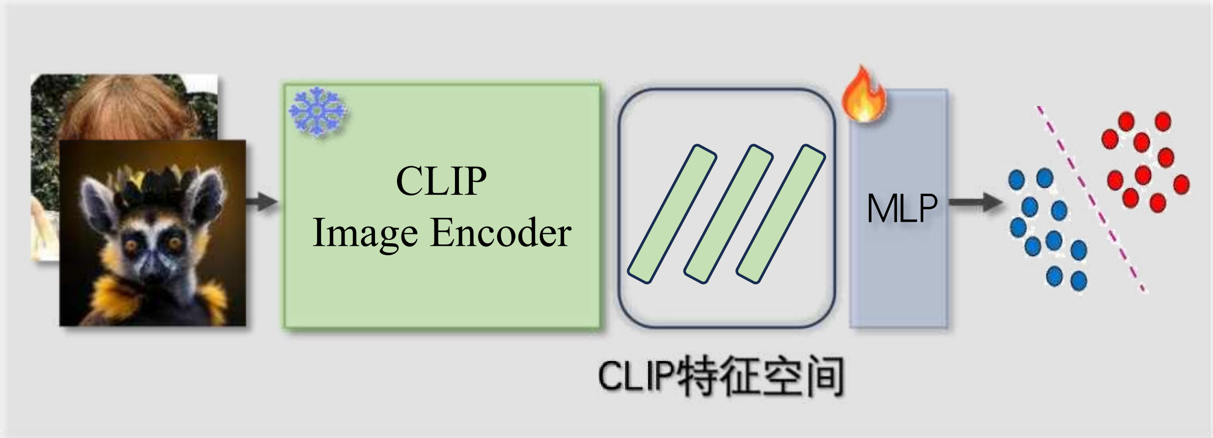
基于伪造概念注入的泛化性深伪检测方法

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, Yunchao Wei: C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection. ***Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)***, pp. 28130-28139. 2025. **(CCF A类，人工智能领域顶级会议)**

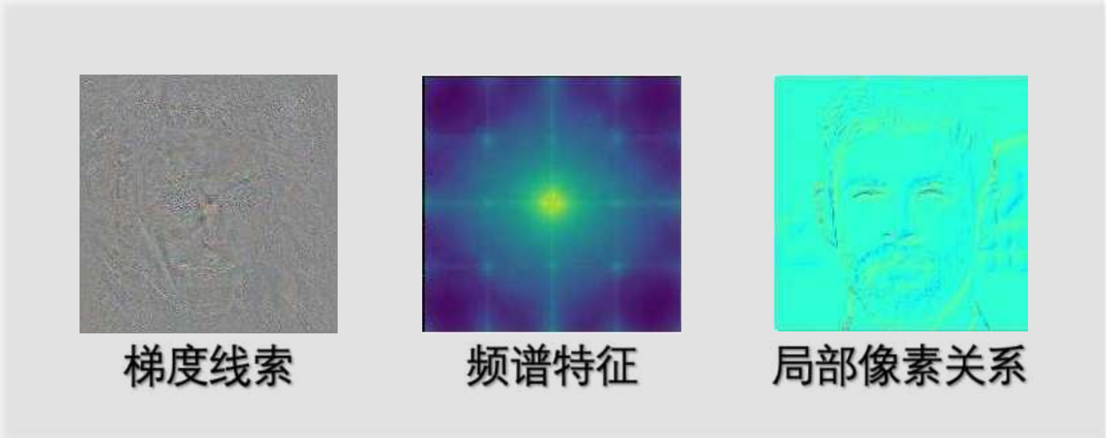
3.3 研究工作-伪造概念注入

- 现象：大规模预训练模型(如CLIP)可通过线性分类器实现泛化性伪造检测；
- 问题：实现线性检测的机制是什么？如何进一步挖掘CLIP的伪造检测潜力？

高层伪造语义特征



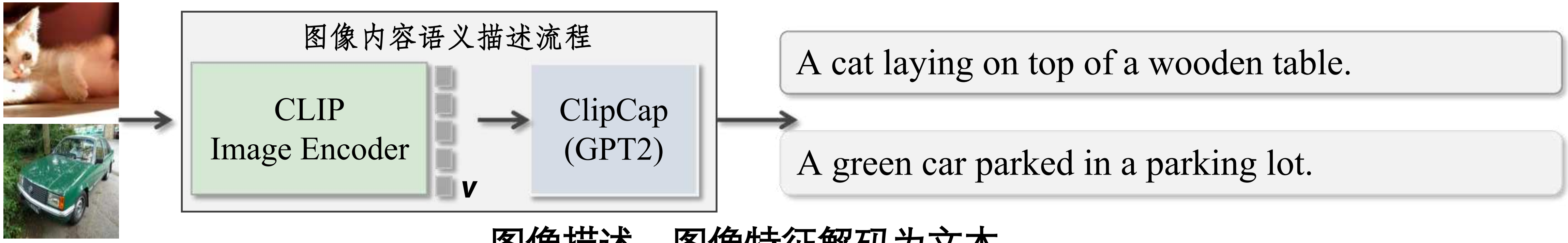
像素级伪造线索表示



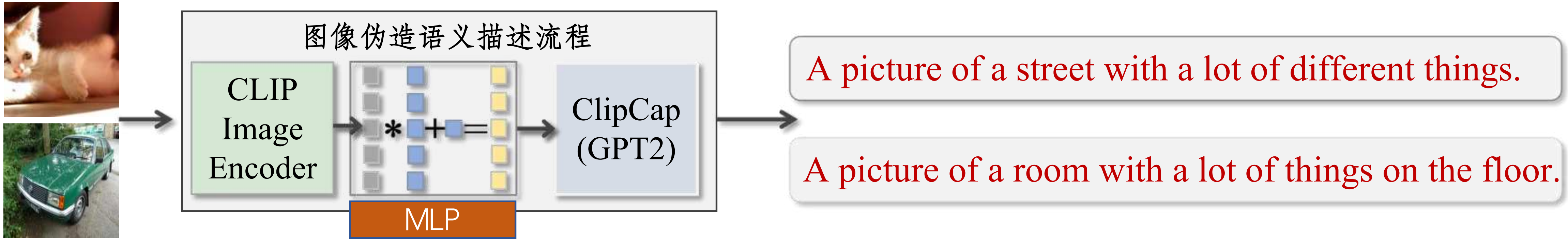
Chuangchuang Tan, R Tao, H Liu, G Gu, B Wu, Y Zhao, etal. C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection. **AAAI. 2025 (CCF-A).**

3.3 研究工作-伪造概念注入

研究思路 □ 高维特征转化为文本探知检测机制



图像描述：图像特征解码为文本

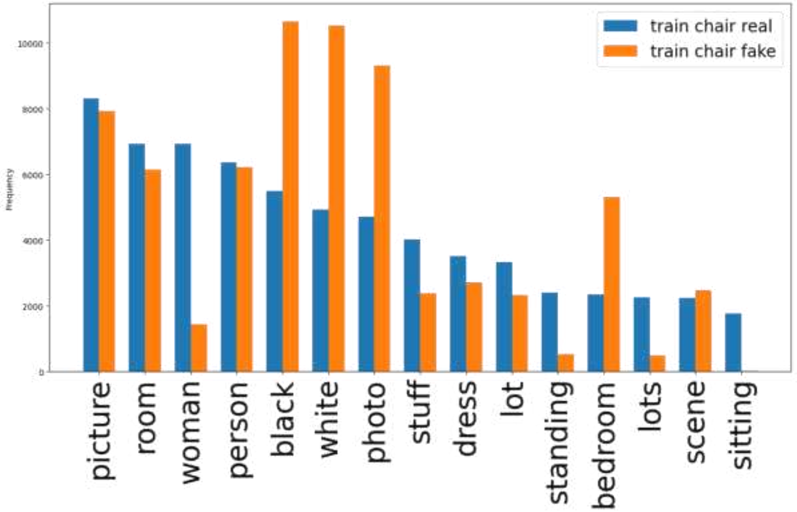


检测描述：检测特征解码为文本

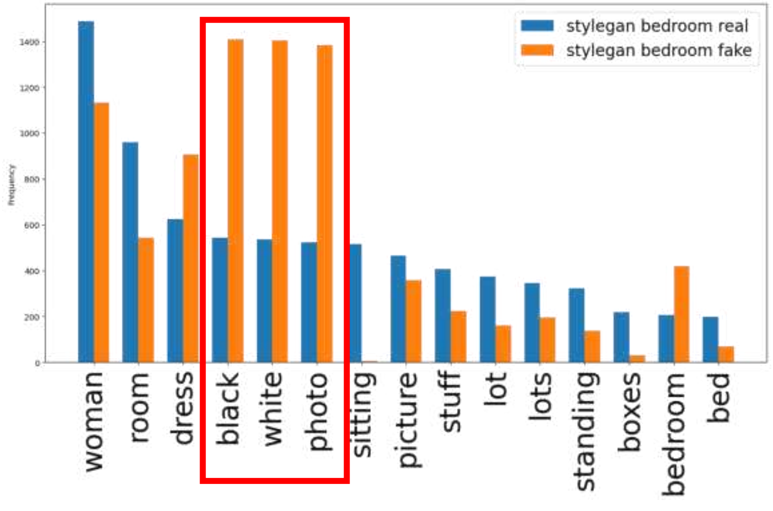
Chuangchuang Tan, R Tao, H Liu, G Gu, B Wu, Y Zhao, etal. C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection. **AAAI. 2025 (CCF-A).**

3.3 研究工作-伪造概念注入

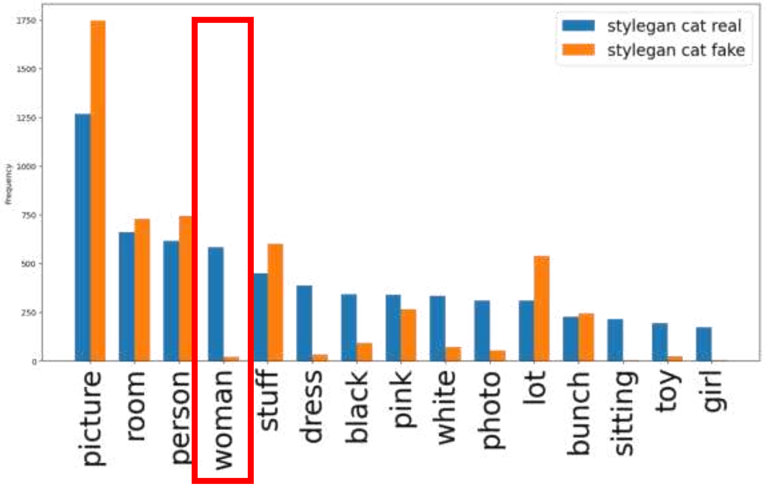
研究思路 □ 高维特征转化为文本探知检测机制＋词频分析



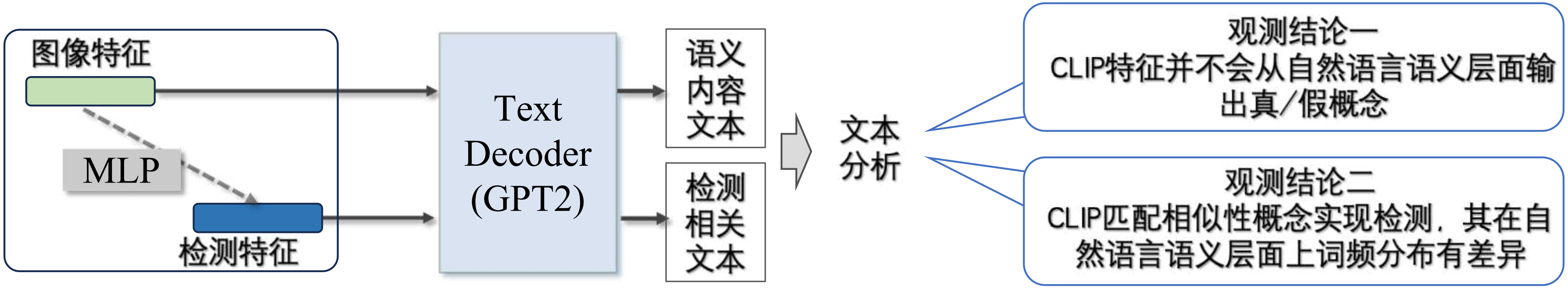
ProGAN chair 训练集



StyleGAN bedroom 测试集

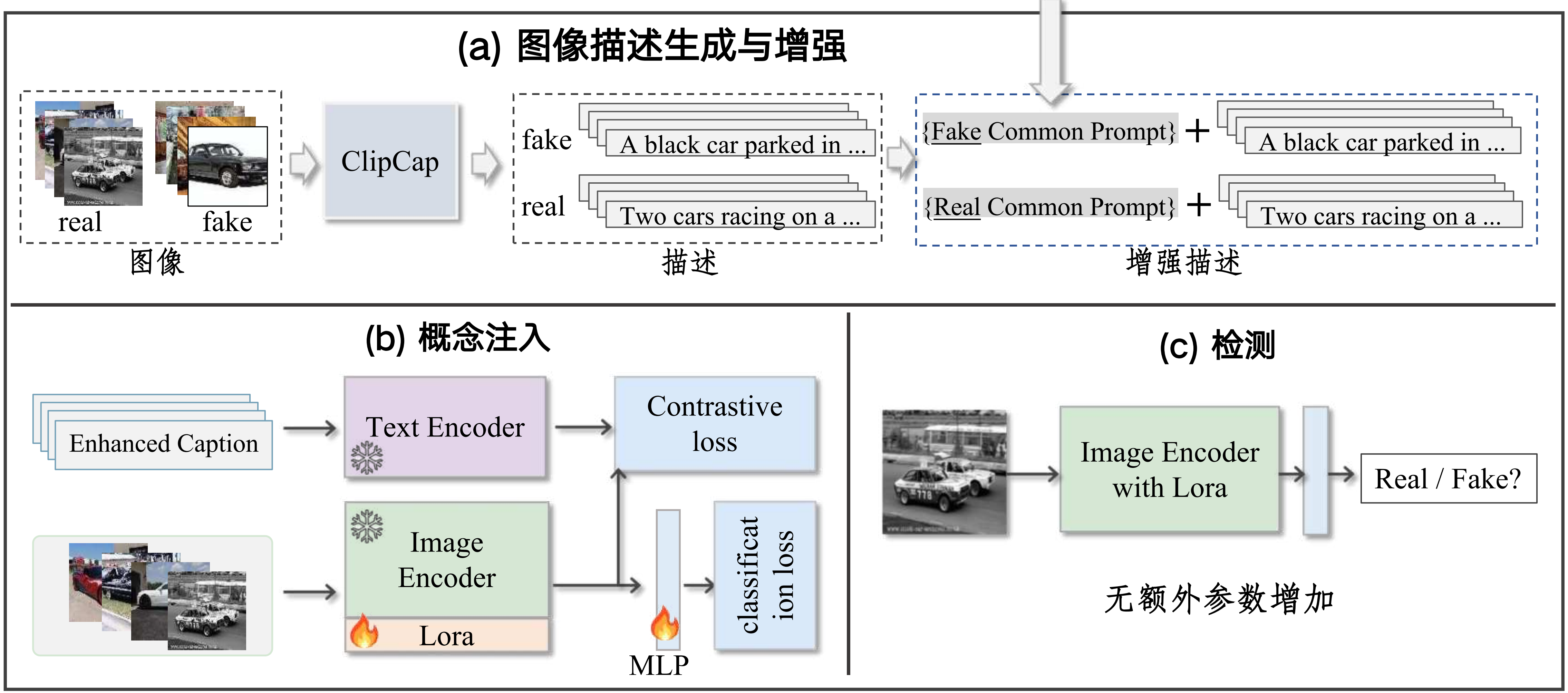


StyleGAN cat 测试集



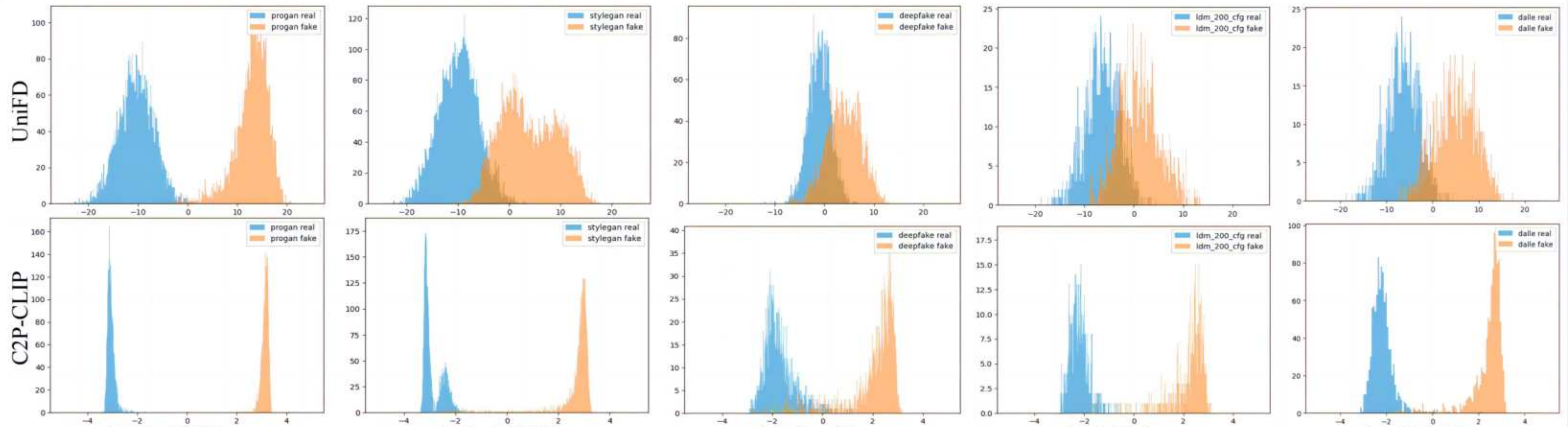
Chuangchuang Tan, R Tao, H Liu, G Gu, B Wu, Y Zhao, etal. C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection. **AAAI. 2025 (CCF-A).**

□ 通过 Prompt 将真假概念注入 CLIP 中。 {Deepfake，Camera} {Trump，Biden}



Chuangchuang Tan, R Tao, H Liu, G Gu, B Wu, Y Zhao, etal. C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection. **AAAI. 2025 (CCF-A).**

Methods	Ref	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAcc
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
CNN-Spot	CVPR2020	99.99	85.20	70.20	85.7	78.95	91.7	53.47	66.67	48.69	86.31	86.26	60.07	54.03	54.96	54.14	60.78	63.8	65.66	55.58	69.58
Patchfor	ECCV2020	75.03	68.97	68.47	79.16	64.23	63.94	75.54	75.14	75.28	72.33	55.3	67.41	76.5	76.1	75.77	74.81	73.28	68.52	67.91	71.24
Co-occurence	Elect. Imag.	97.70	63.15	53.75	92.50	51.1	54.7	57.1	63.06	55.85	65.65	65.80	60.50	70.7	70.55	71.00	70.25	69.60	69.90	67.55	66.86
Freq-spec	WIFS2019	49.90	99.90	50.50	49.90	50.30	99.70	50.10	50.00	48.00	50.60	50.10	50.90	50.40	50.40	50.30	51.70	51.40	50.40	50.00	55.45
F3Net	ECCV2020	99.38	76.38	65.33	92.56	58.10	100.0	63.48	54.17	47.26	51.47	51.47	69.20	68.15	75.35	68.80	81.65	83.25	83.05	66.30	71.33
UniFD	CVPR2023	100.0	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.5	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
LGrad	CVPR2023	99.84	85.39	82.88	94.83	72.45	99.62	58.00	62.50	50.00	50.74	50.78	77.50	94.20	95.85	94.80	87.40	90.70	89.55	88.35	80.28
FreqNet	AAAI2024	97.90	95.84	90.45	97.55	90.24	93.41	97.40	88.92	59.04	71.92	67.35	86.70	84.55	99.58	65.56	85.69	97.40	88.15	59.06	85.09
NPR	CVPR2024	99.84	95.00	87.55	96.23	86.57	99.75	76.89	66.94	98.63	50.00	50.00	84.55	97.65	98.00	98.20	96.25	97.15	97.35	87.15	87.56
FatFormer	CVPR2024	99.89	99.32	99.50	97.15	99.41	99.75	93.23	81.11	68.04	69.45	69.45	76.00	98.60	94.90	98.65	94.35	94.65	94.20	98.75	90.86
Ours	Trump,Biden	99.71	90.69	95.28	99.38	95.26	96.60	89.86	98.33	64.61	90.69	90.69	77.80	99.05	98.05	98.95	94.65	94.20	94.40	98.80	93.00
Ours	Deepfake,Camera	99.98	97.31	99.12	96.44	99.17	99.60	93.77	95.56	64.38	93.29	93.29	69.10	99.25	97.25	99.30	95.25	95.25	96.10	98.55	**93.79**



(a) ProGAN (b) StyleGAN (c) Deepfake (d) LDM (e) DALLE

Chuangchuang Tan, R Tao, H Liu, G Gu, B Wu, Y Zhao, etal. C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection. **AAAI. 2025 (CCF-A).**

以论文为目的

继承自Baseline

模糊

堆模块

增量提升

侧重

凑，乱

A CV Paper

现象

问题

动机

方法

SOTA

故事

以任务为目的

观察总结

定义与分析

清晰

契合任务及问题

水到渠成

侧重

顺畅，自洽

Image → LLM → Detection

Aritifact → LLM → Reasons

- 语义不合理定义难
- 像素级伪影语义化难
- 相关数据缺乏
- 可解释性检测评估难

谢谢！ 敬请各位老师同学批评指正

报告人：谭创创