中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

人脸伪造合成度分析研究

李琦

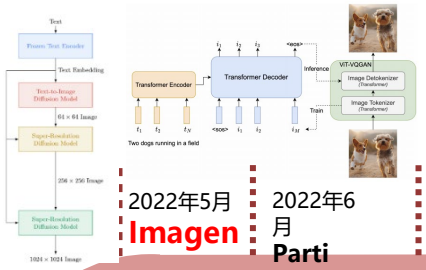
中国科学院自动化研究所

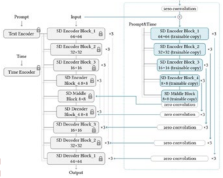
模式识别实验室

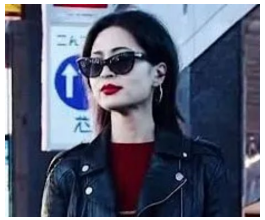
深度伪造→深度合成→ AIGC

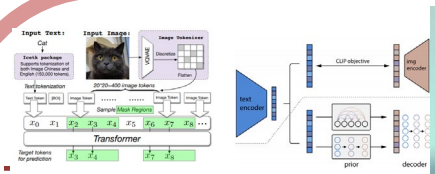
- 从2017年深度伪造（Deepfake）换脸技术突破以来，深度合成（Deep Synthesis）、人工智能生成内容（AIGC）发展迅猛，在文生图、文生视频等方面不断取得突破，真实与虚假的边界日益模糊。

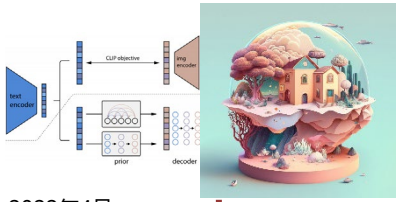
合成技术发展趋势：
专用→通用
写实→写意
单模态→多模态

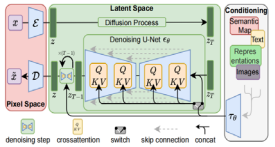


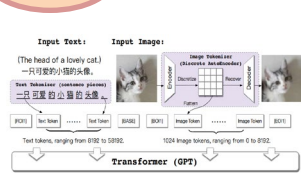












2017 Deepfake
2019 StyleGAN
2021
2022
2023
2024

2022年5月 **Imagen**
2022年6月 **Parti**
2022年8月 **Stable Diffusion**
2022年10月 **NovelAI AI绘画**
2023年2月 **ControlNet**
2024年2月 **Sora**
2022年4月 **CogView2**
2022年4月 **DALLE2**
2022年3月 **Midjourney**
2021年12月 **文心ERNIE-ViLG**
2021年12月 **Latent Diffusion**
2021年12月 **GLIDE**
2021年1月 **DALLE**
2021年2月 **CLIP公开**
2021年5月 **CogView**
2021年7月 **紫东太初**

深度伪造→深度合成→ AIGC

深度伪造(Deepfake)——深度合成的一个具体应用

借助深度合成技术，伪造以人为主体的视听觉内容，以达到欺骗的目的。

深度合成(Deep Synthesis)

深度合成技术，是指利用深度学习、虚拟现实等生成合成类算法制作文本、图像、音频、视频、虚拟场景等网络信息的技术。--《深度合成管理规定》

人工智能生成内容(AI Generated Content)

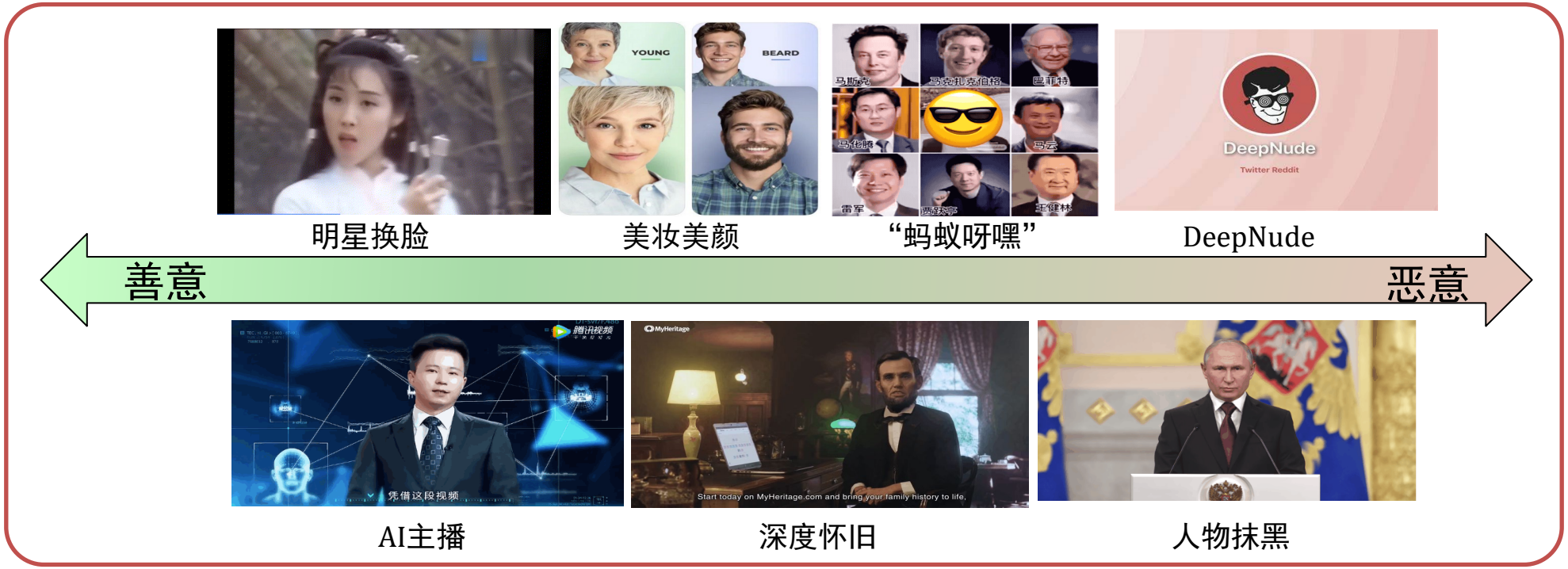
利用ChatGPT、Sora等生成式人工智能工具，根据用户输入的要求来生成新的内容 。

1. 以人为主体
2. 具有欺骗性

- 人脸
- 头部
- 肢体
- 声音

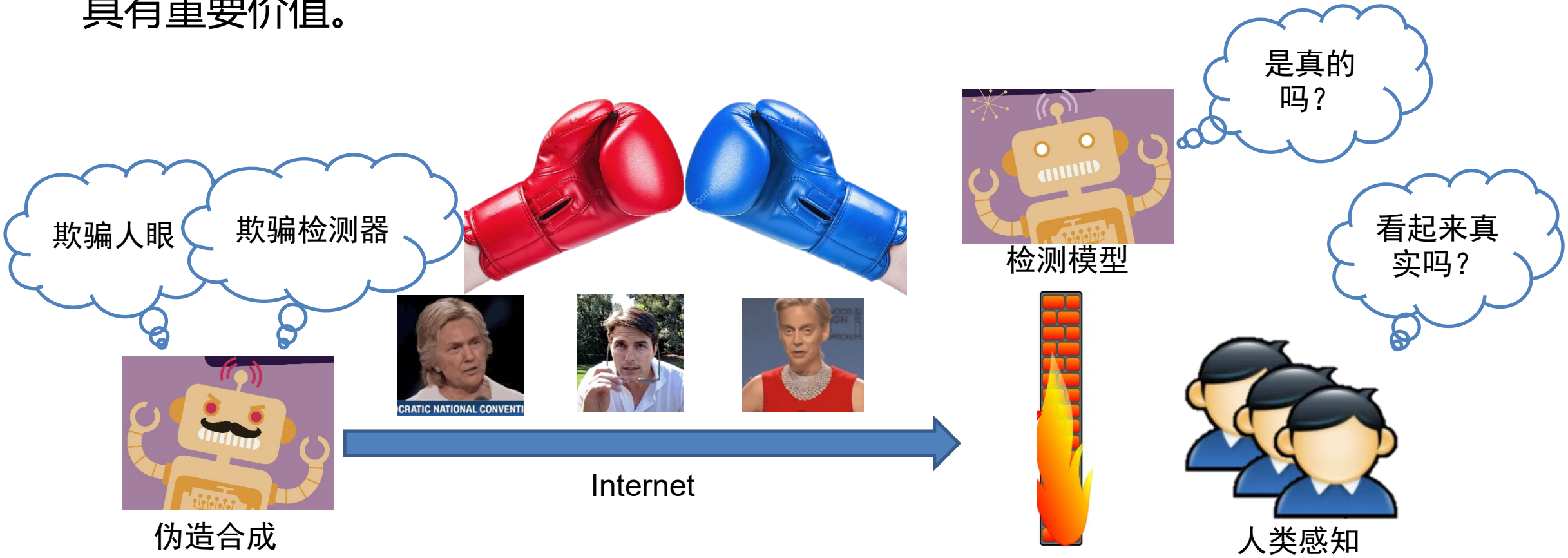
深度伪造→深度合成→ AIGC

- 深度合成技术是一把双刃剑，既有善意应用也有恶意应用，需要在促进技术健康发展的同时防范化解安全威胁。

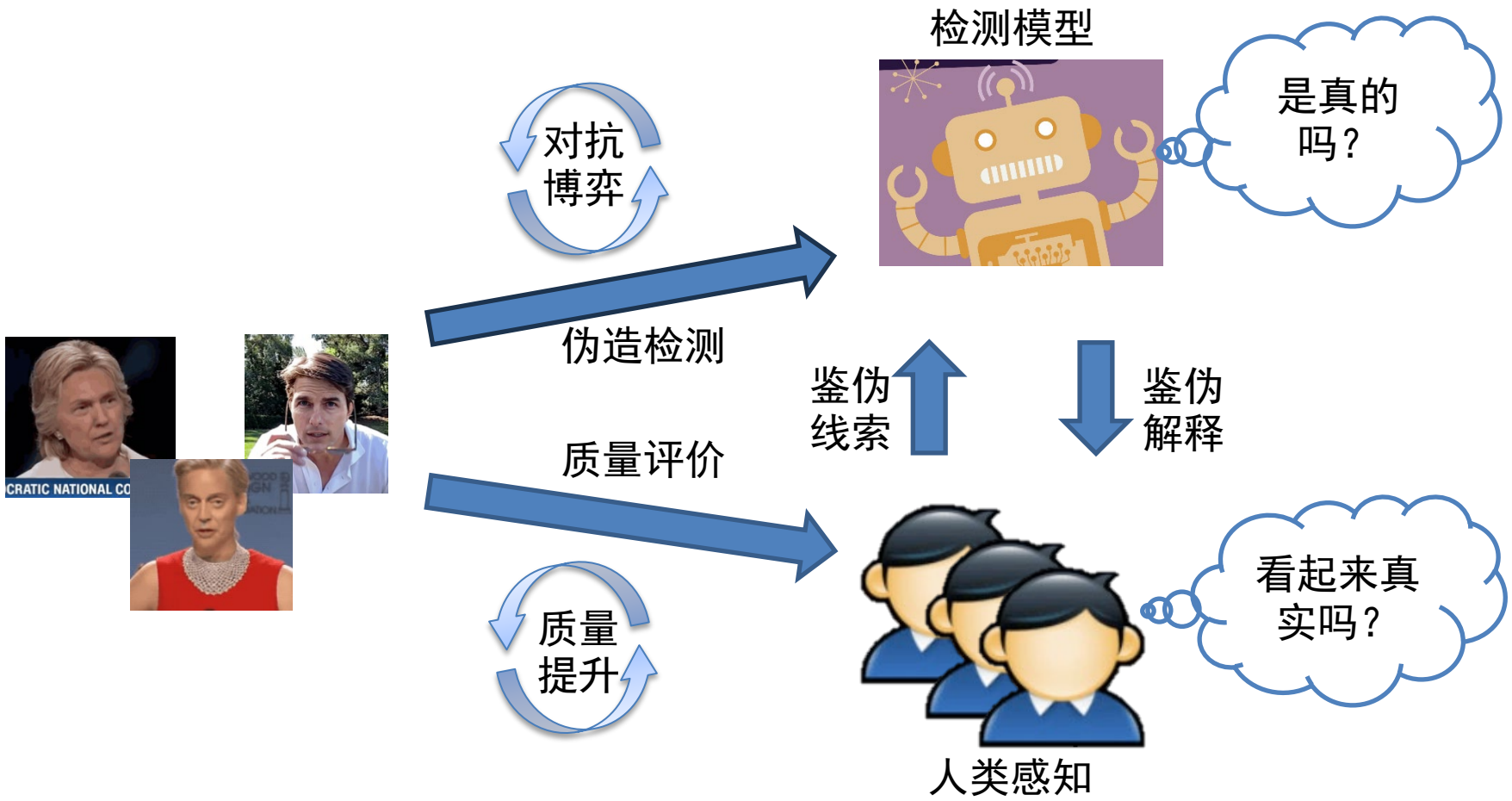


合成度分析

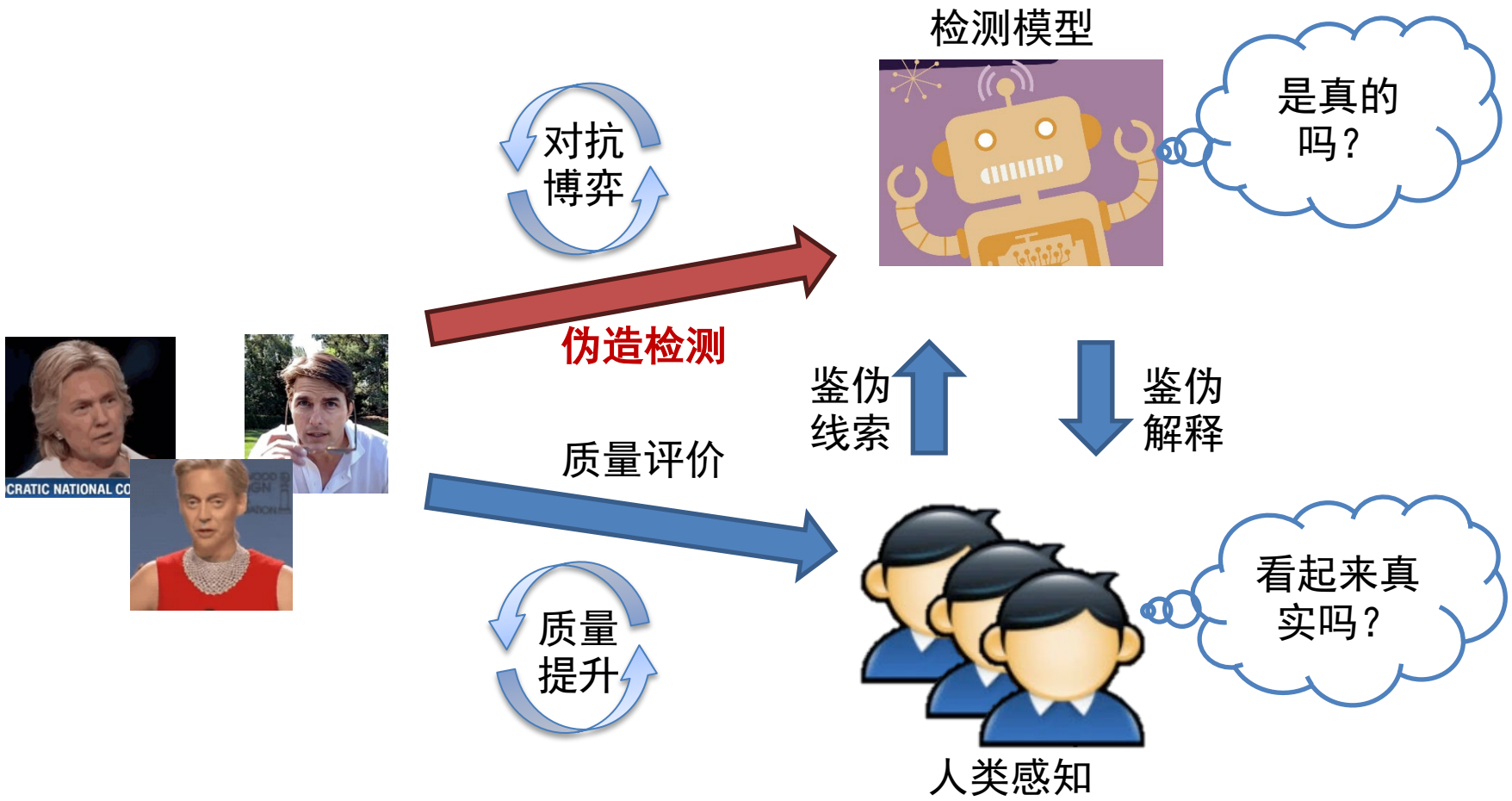
- 合成度分析（Synthesis Analysis）是在**检测模型**和**人类感知**两个视角下对合成内容的真实程度进行评估的研究，对于**防范虚假信息、促进合成质量提升**都具有重要价值。



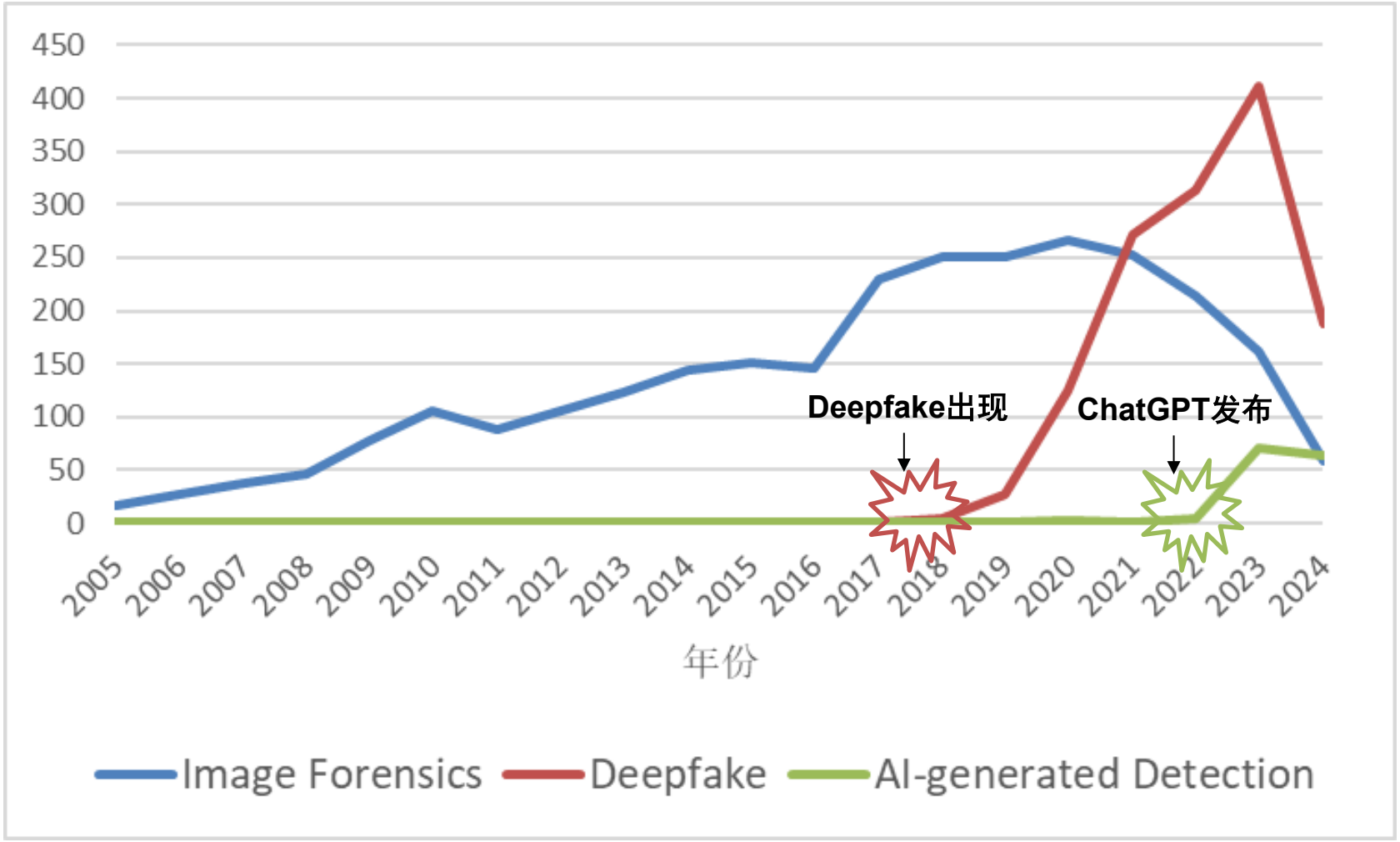
研究概览



研究概览

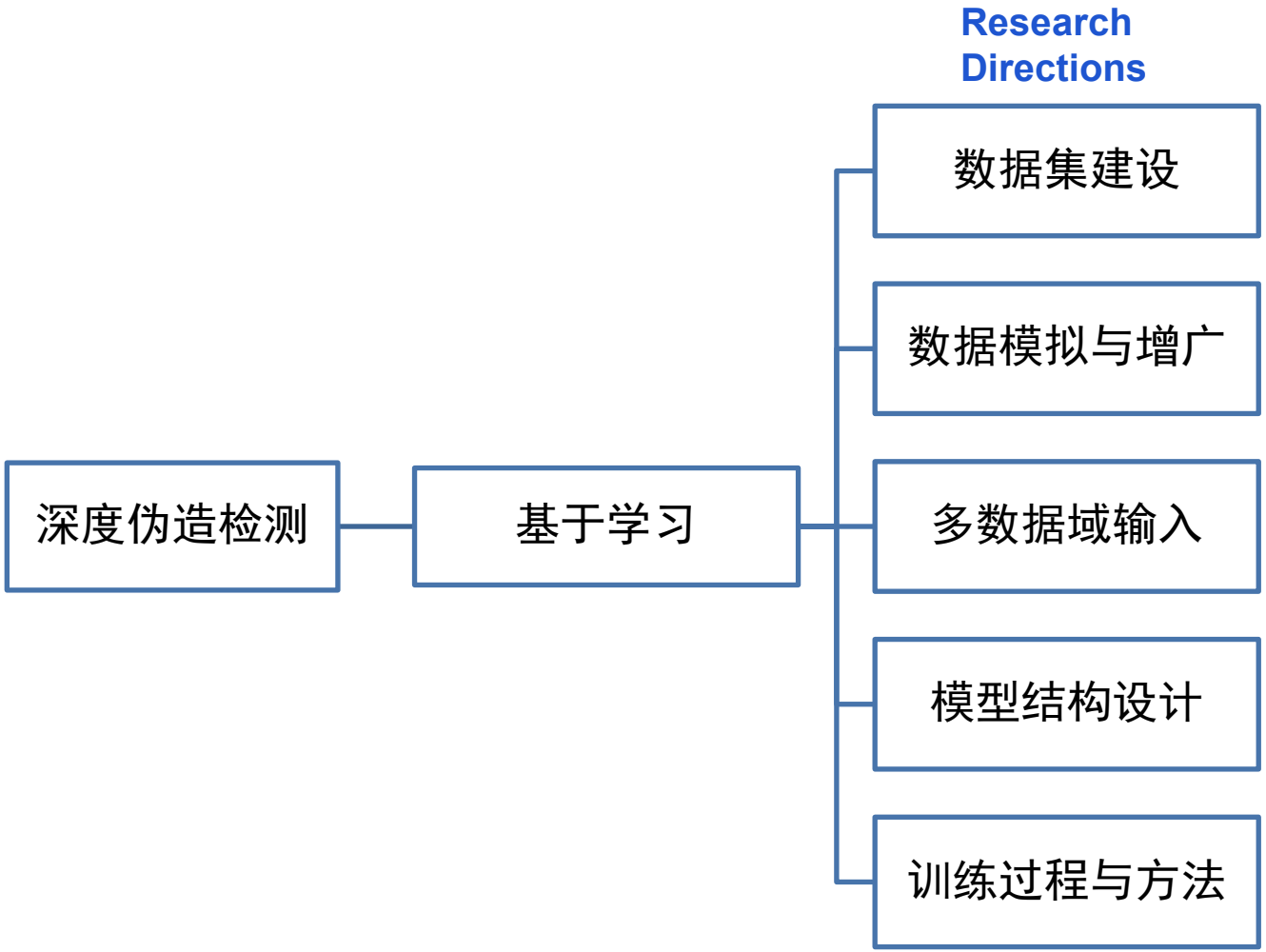


伪造检测研究趋势

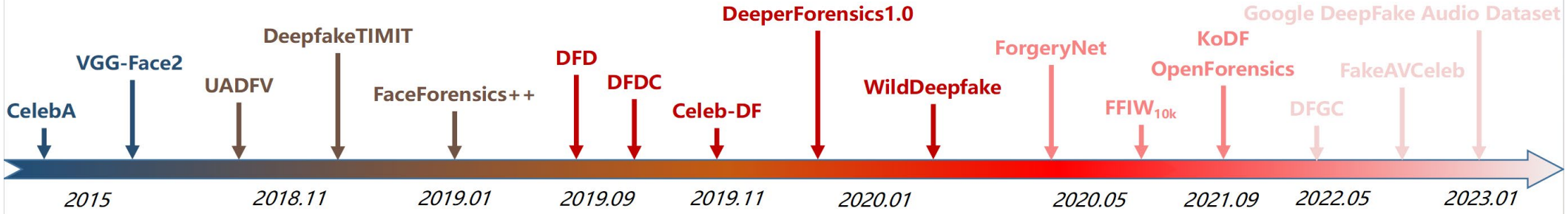


Source of data: https://www.aminer.cn/

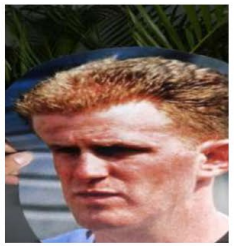
深度伪造检测方法


Research Directions
深度伪造检测
基于学习
数据集建设
数据模拟与增广
多数据域输入
模型结构设计
训练过程与方法

深度伪造检测数据集



早期数据集



特点：只覆盖了种族、姿势和背景等基础信息。

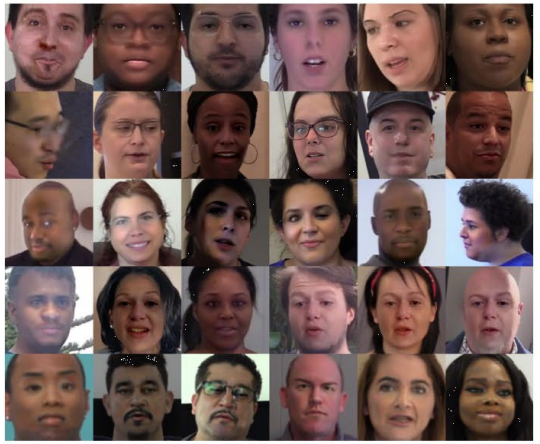
准确率:99%+

第一批数据集

特点:质量较差、场景单一，检测难度低

准确率:99%

第二批数据集



特点:质量较好、场景较复杂，但多样性不足

准确率:97%(DFDC)

第三批数据集

特点:质量较好、多样性较强，更贴近真实情况

准确率:80%(ForgeryNet)

深度伪造比赛

Media Forensics Challenge (MFC) 2017~2019

NIST
Information Technology Laboratory / Information Access Division
MULTIMODAL INFORMATION GROUP

- Manipulation Detection and Localization
- Splice Detection and Localization

安全AI挑战者计划：
伪造图像的篡改检测-长期赛

中国人工智能大赛
China Artificial Intelligence Competition

kaggle

Deepfake Detection Challenge (DFDC)

CSIG人工智能安全大赛

第二届CSIG图像图形技术挑战赛竞赛项目六
多媒体伪造取证大赛
Fake Media Forensic Challenge of CSIG

DeepFake Game Competition

深度伪造博弈比赛（DFGC 2021）

ForgeryNet: 2021
Face Forgery Analysis Challenge

广播电视和网络视听
人工智能应用创新大赛

国家广播电视总局
National Radio and Television Administration

DeepFake Game Competition

深度伪造博弈竞赛（DFGC 2022）

DeepFake Game Competition

深度伪造博弈竞赛（DFGC-VRA）

蚂蚁集团主办

inclusion 外滩大会
全球Deepfake
攻防挑战赛
百万数据集 百万奖金池

融新汇智 竞促发展
第五届中国人工智能大赛
The 5th China AI Competition

第五届中国人工智能大赛
- AIGC视频检测赛
- AIGC音频检测赛
- 大模型安全攻防赛

2017 2018 2019 2020 2021 2022 2023 2024

深度伪造博弈比赛（DFGC）

连续三年组织深度伪造博弈比赛，伪造与反伪造队伍进行多轮对抗博弈。

- DFGC-2021 构建了包含2万张换脸图像的数据集，涉及多种图像换脸方法，并携带高迁移性的对抗噪声。
- DFGC-2022 构建了包含4394段高质量视频的换脸数据集，涉及7种换脸模型、4大类后处理方法。
- DFGC-2023引入了主观视觉评价方法，提升鉴伪客观检测的可解释化依据指标。

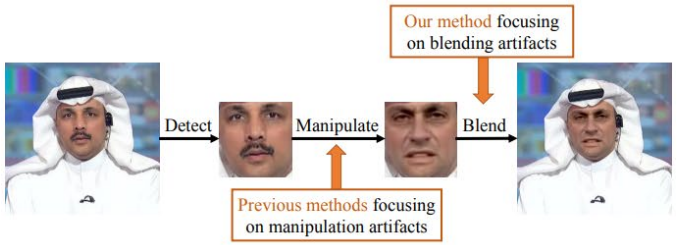
以赛促研：深度伪造博弈国际竞赛

依托比赛构造新数据集

Bo Peng, Hongxing Fan, Wei Wang, Jing Dong, Yuezun Li, Siwei Lyu, Qi Li, Zhenan Sun, "DFGC 2021: A DeepFake Game Competition", IJCB 2021
Bo Peng, Wei Xiang, Yue Jiang, Wei Wang, Jing Dong, Zhenan Sun, Zhen Lei, Siwei Lyu, “DFGC 2022: The Second DeepFake Game Competition", IJCB 2022.
Bo Peng, Xianyun Sun, Caiyong Wang, Wei Wang, Jing Dong, Zhenan Sun, "DFGC-VRA: DeepFake Game Competition on Visual Realism Assessment", IJCB 2023.

自监督数据增广—Face X-ray

- Face X-ray是一种对**人脸拼接边缘进行模拟**的方法，利用合成数据来自督训练，获得更具泛化性的检测能力。



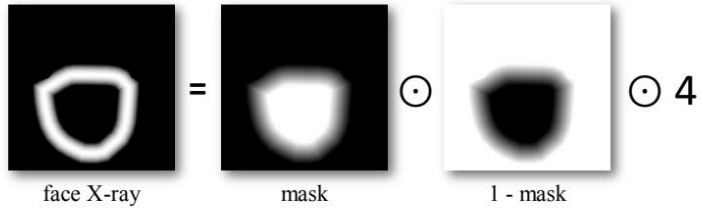
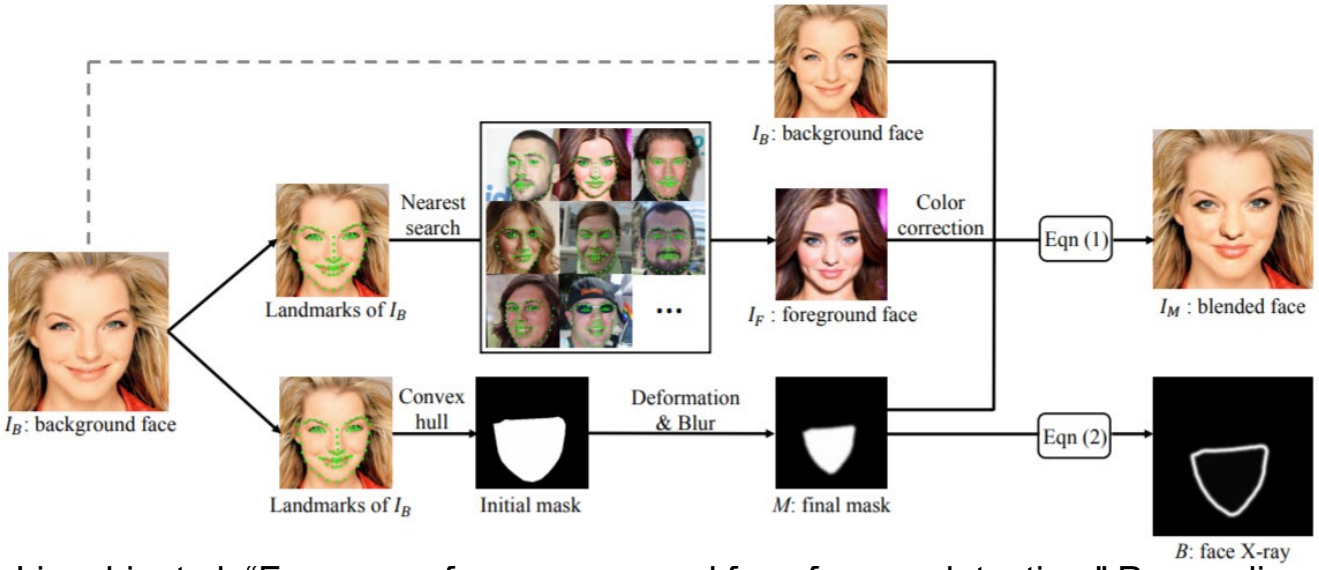
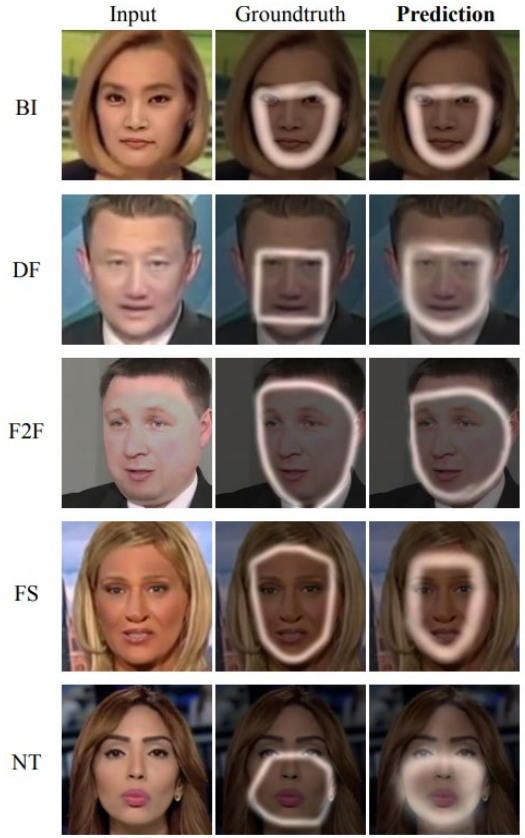


Figure 4. Illustrating the relationship between face X-ray and the

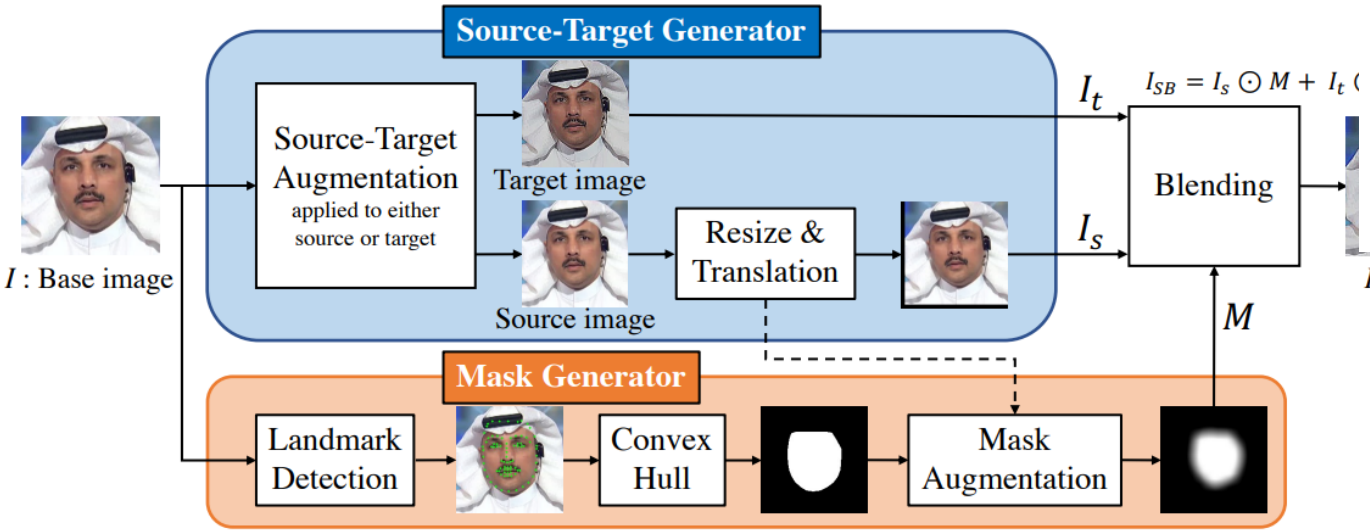


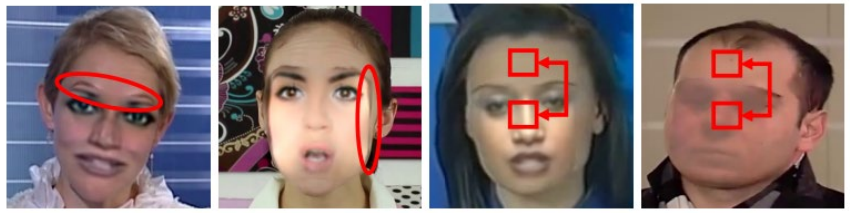


Li, Lingzhi, et al. “Face x-ray for more general face forgery detection.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

自监督数据增广—SBI

- 之前的方法（Face X-ray）基于两张不同的人脸图像产生模拟换脸数据，SBI方法只**使用一张图像产生模拟数据**，避免不同图像ID不匹配、相机不匹配等带来的数据偏斜，迫使模型关注更本质的**拼接融合痕迹。**





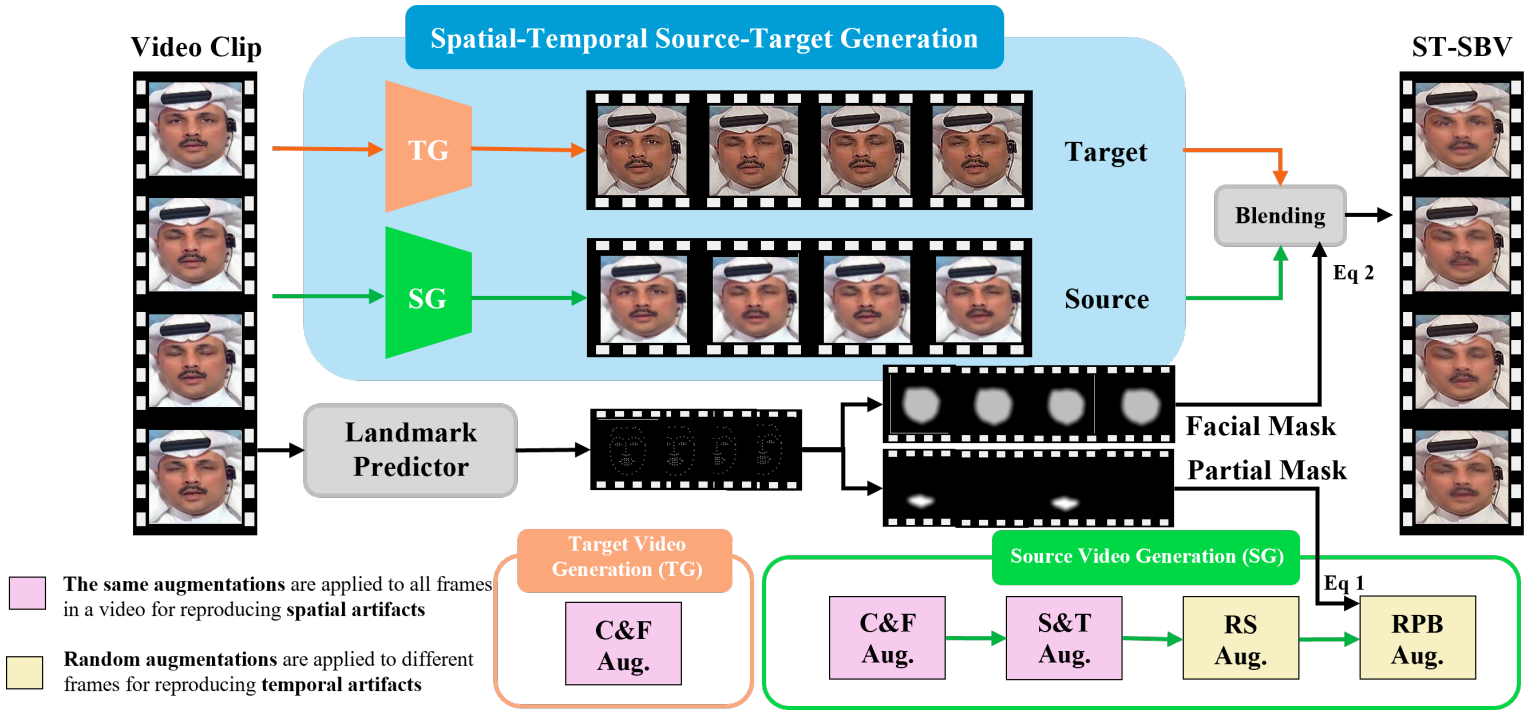
真实数据中的典型伪造痕迹

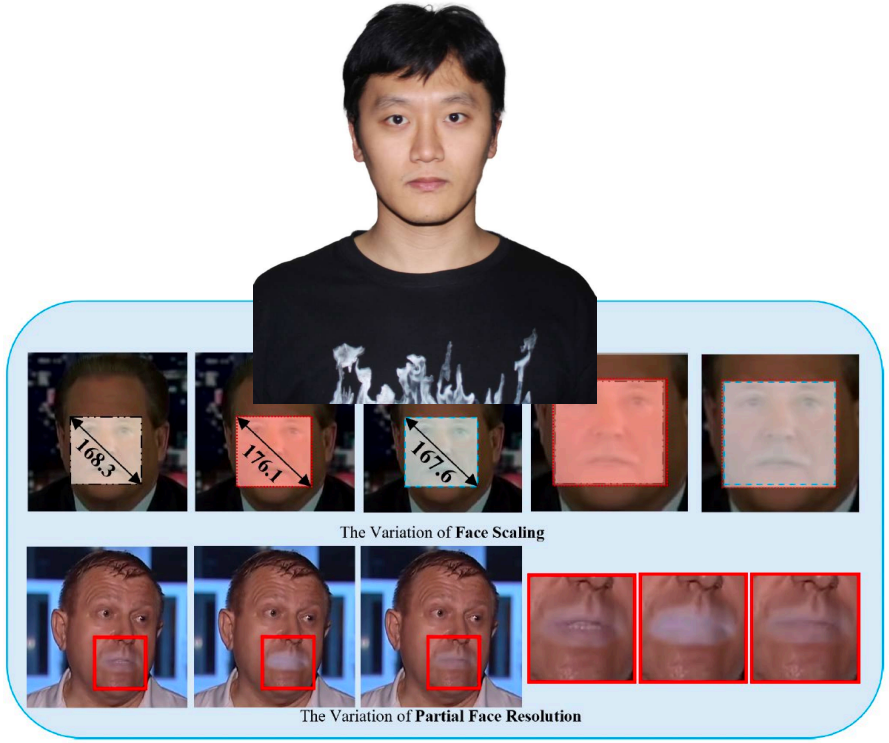
模拟产生的伪造痕迹

Shiohara, Kaede, and Toshihiko Yamasaki. "Detecting Deepfakes with Self-Blended Images." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

自监督数据增广—ST-SBV

- 现有方法大部分针对空域伪造痕迹进行模拟， ST-SBV方法对视频中时空域痕迹进行模拟和自监督数据增广，取得了更好的检测泛化性效果。

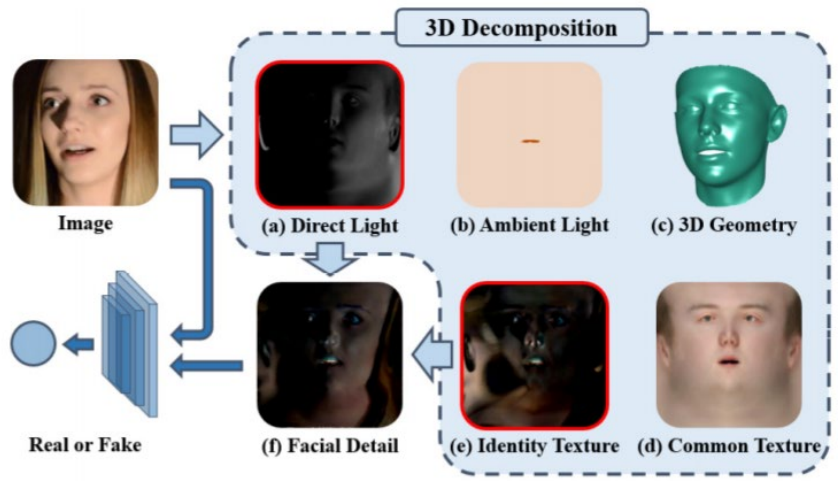


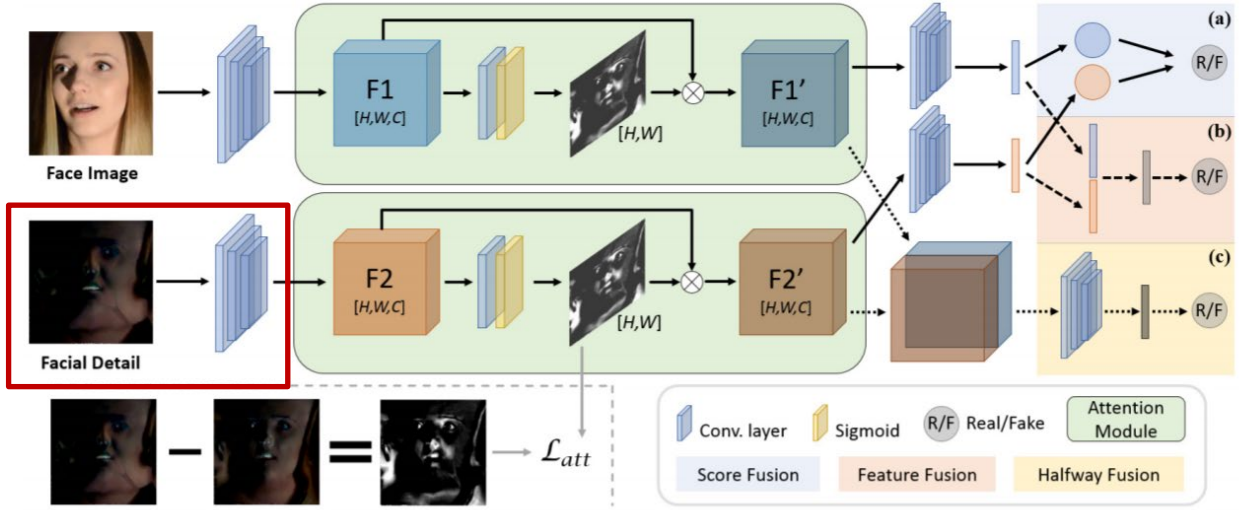


Weinan Guan, Wei Wang, Bo Peng, Jing Dong, Tieniu Tan, "ST-SBV: Spatial-Temporal Self-Blended Videos for Deepfake Detection", Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 2024.

空域数据归一化处理

- Zhu等人提出基于3D人脸模型得到的面部细节图像作为归一化处理手段，增强伪造线索来检测微弱的伪造痕迹特征。

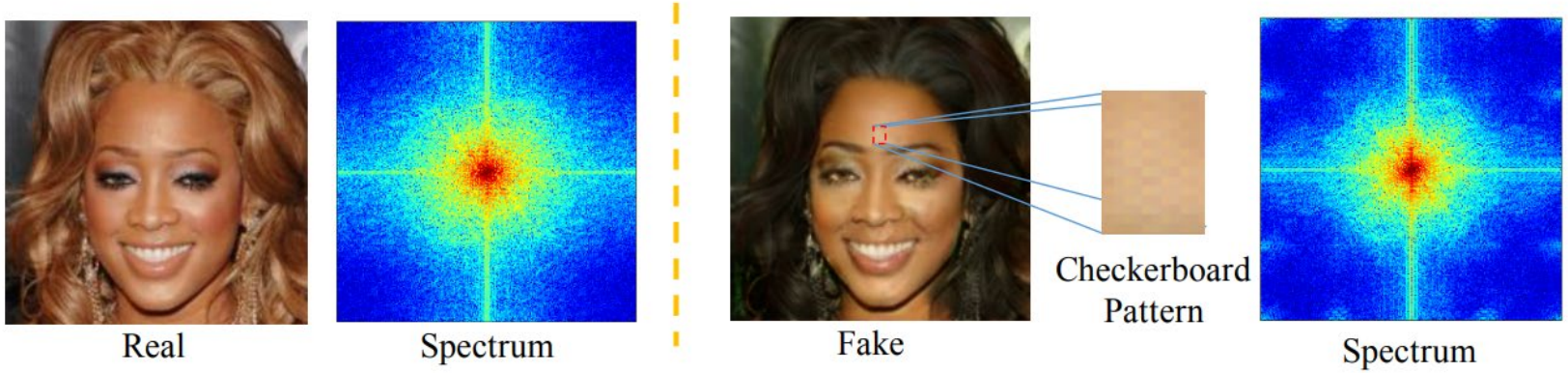


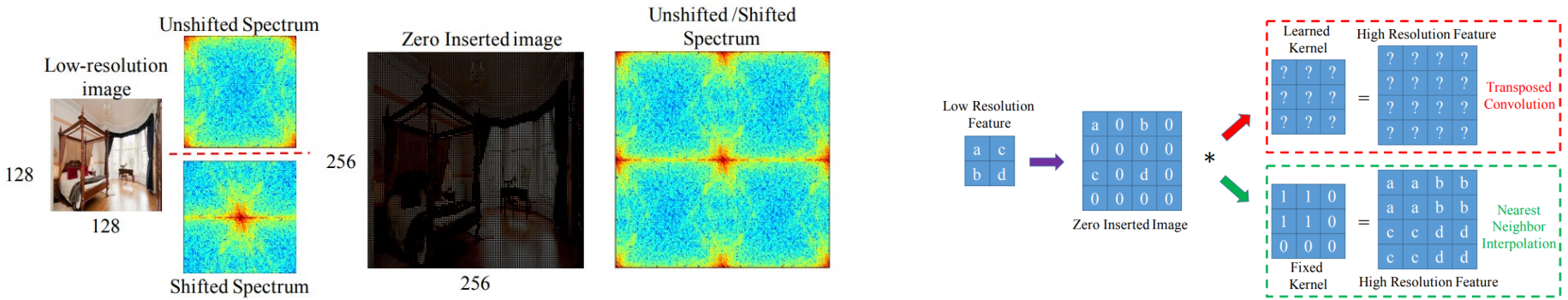


Zhu, Xiangyu, et al. "Face Forgery Detection by 3D Decomposition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

频率域方法—AutoGAN

- GAN生成的图像在相邻像素之间存在棋盘格模式像素伪影，导致频域中出现高频痕迹。AutoGAN提出在频域Spectrum图像上训练检测器。

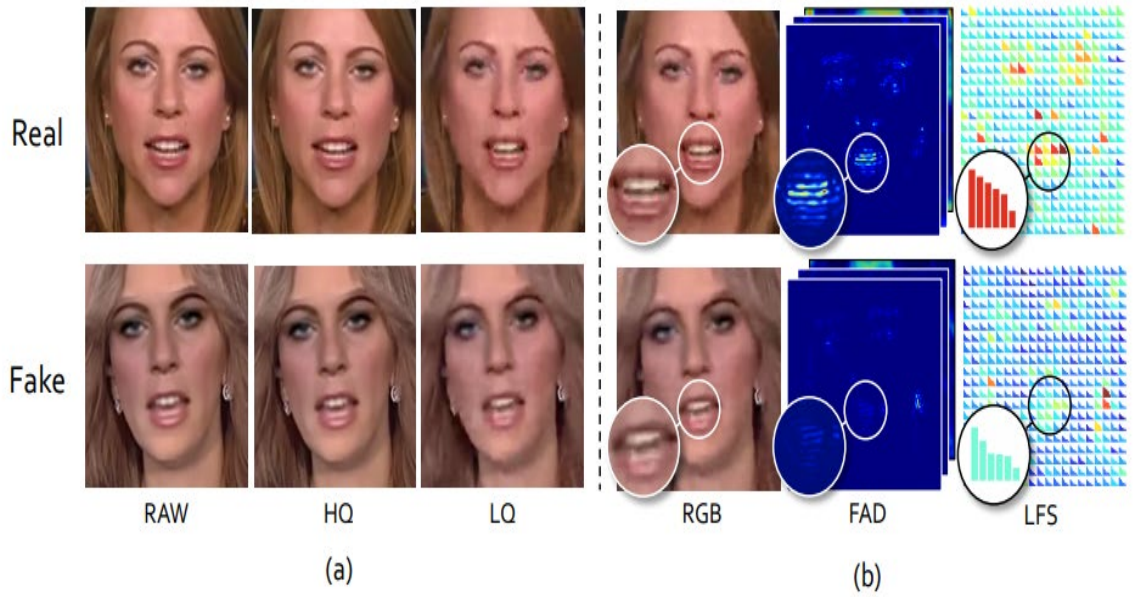




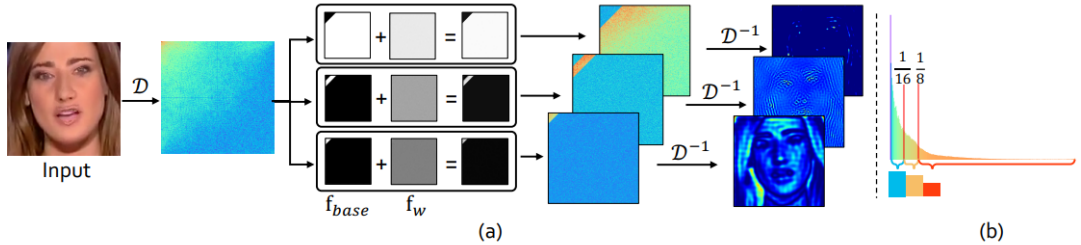
Zhang, Xu, Svebor Karaman, and Shih-Fu Chang. "Detecting and simulating artifacts in gan fake images." WIFS, 2019.

频率域方法—F3-Net

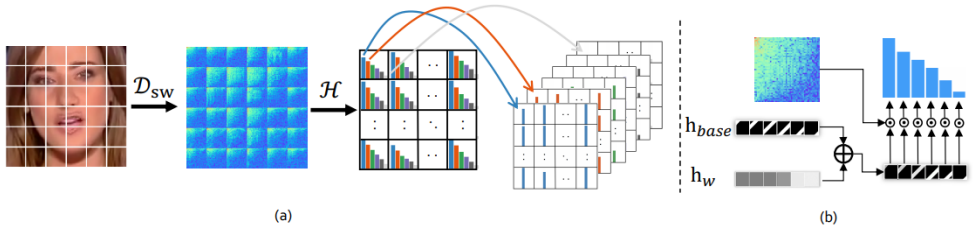
- F3-Net提出频率感知分解（Frequency-aware Decomposition，简称FAD）和局部频率统计（Local Frequency Statistics，简称LFS），可以在低质量图像中检测到频率感知的伪造线索。



- FAD采用带通滤波得到对应频段的图像分解。



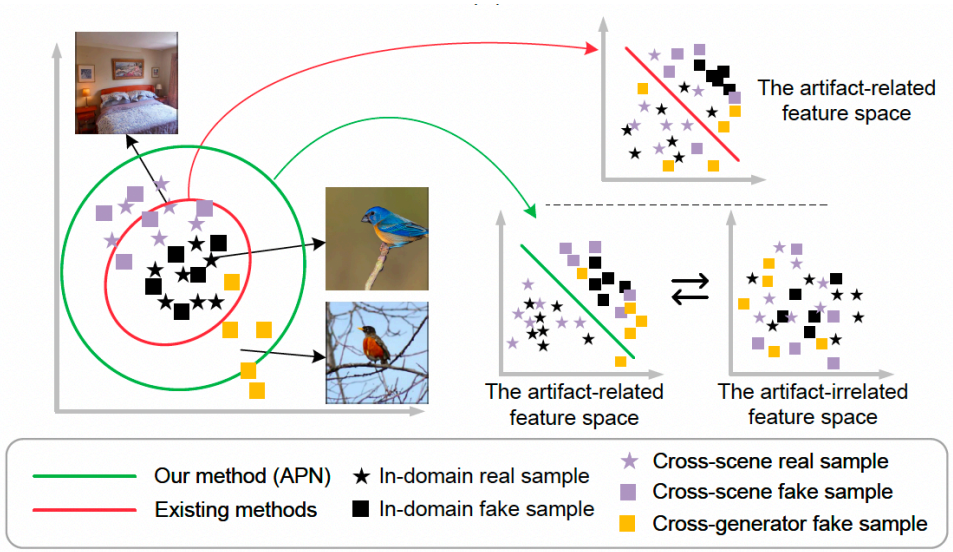
- LFS采用滑窗DCT变换得到局部频率统计信息。

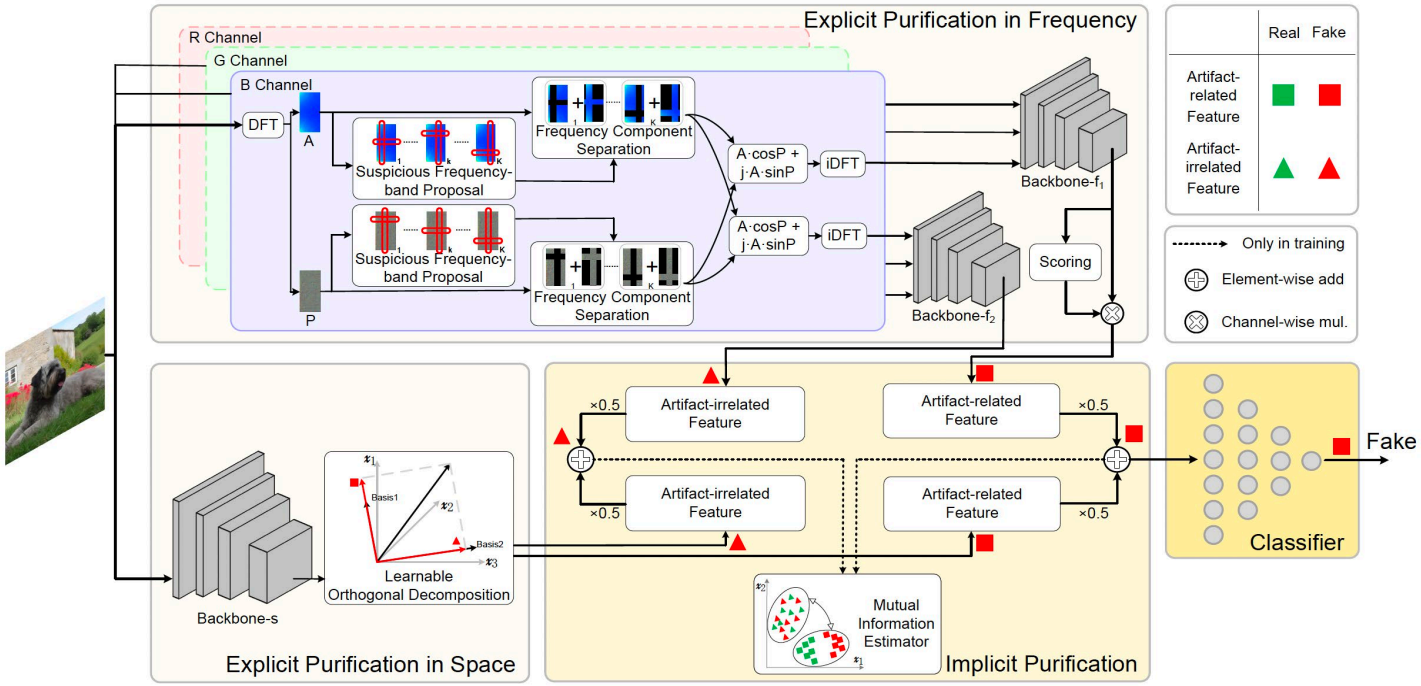


Qian, Yuyang, et al. "Thinking in frequency: Face forgery detection by mining frequency-aware clues." ECCV, 2020.

模型设计—APN网络

- 提出一种对AIGC生成图像特征进行萃取纯化的APN网络，从频域和空域分别提取解耦的生成痕迹特征，有效提升了面对未知伪造模型和新内容场景的检测泛化能力。

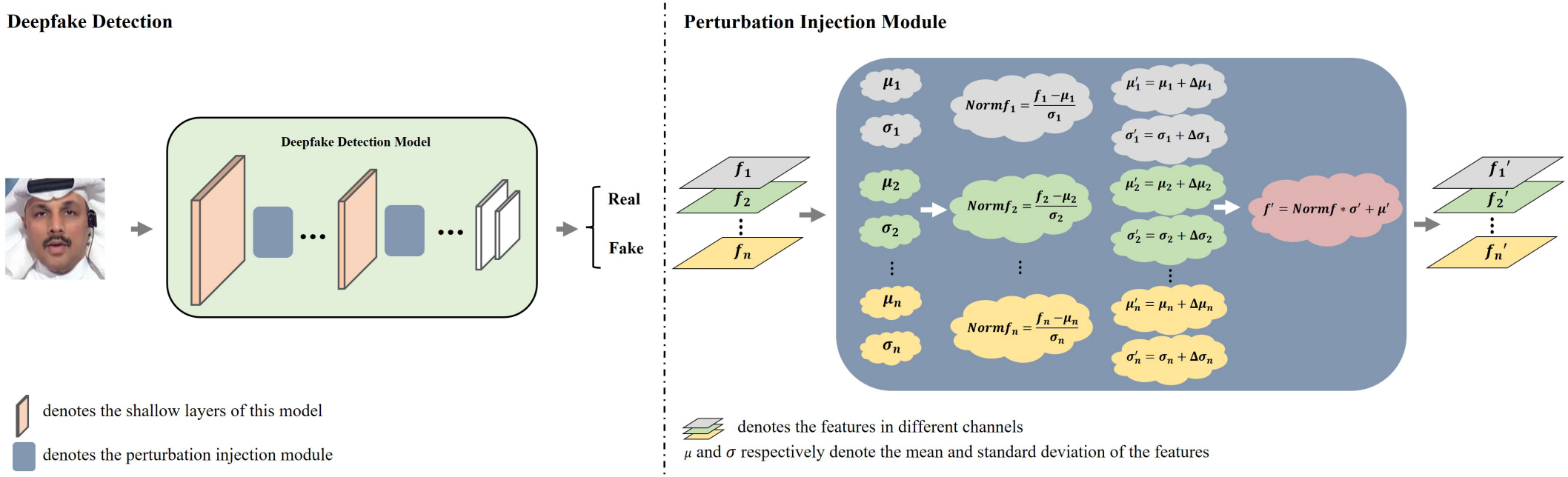




Zheling Meng, Bo Peng, Jing Dong, Tieniu Tan, Haonan Cheng, "Artifact Feature Purification for Cross-domain Detection of AI-generated Images", Computer Vision and Image Understanding (CVIU), 2024.

损失函数设计—梯度正则化

- 先前研究表明检测器容易捕捉到特定于深度伪造方法的纹理特征（浅层特征）。
- 所提方法核心在于抑制检测器对浅层特征统计量变化的敏感度，从而增强检测器的泛化能力。



Weinan Guan, Wei Wang, Jing Dong, Bo Peng, "Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization", IEEE Transactions on Information Forensics & Security (TIFS), 2024.

损失函数设计—梯度正则化

- 在二分类经验损失基础上，加入浅层特征敏感度正则损失。

经验损失函数　　损失对浅层特征统计量的梯度　　浅层特征　　深层网络参数

$$\min_{\theta} L(x, y, \theta) + \lambda ||\nabla_{\mu_s, \sigma_s} L_E(f_{\theta_s}(x), y, \theta_d)||_2.$$

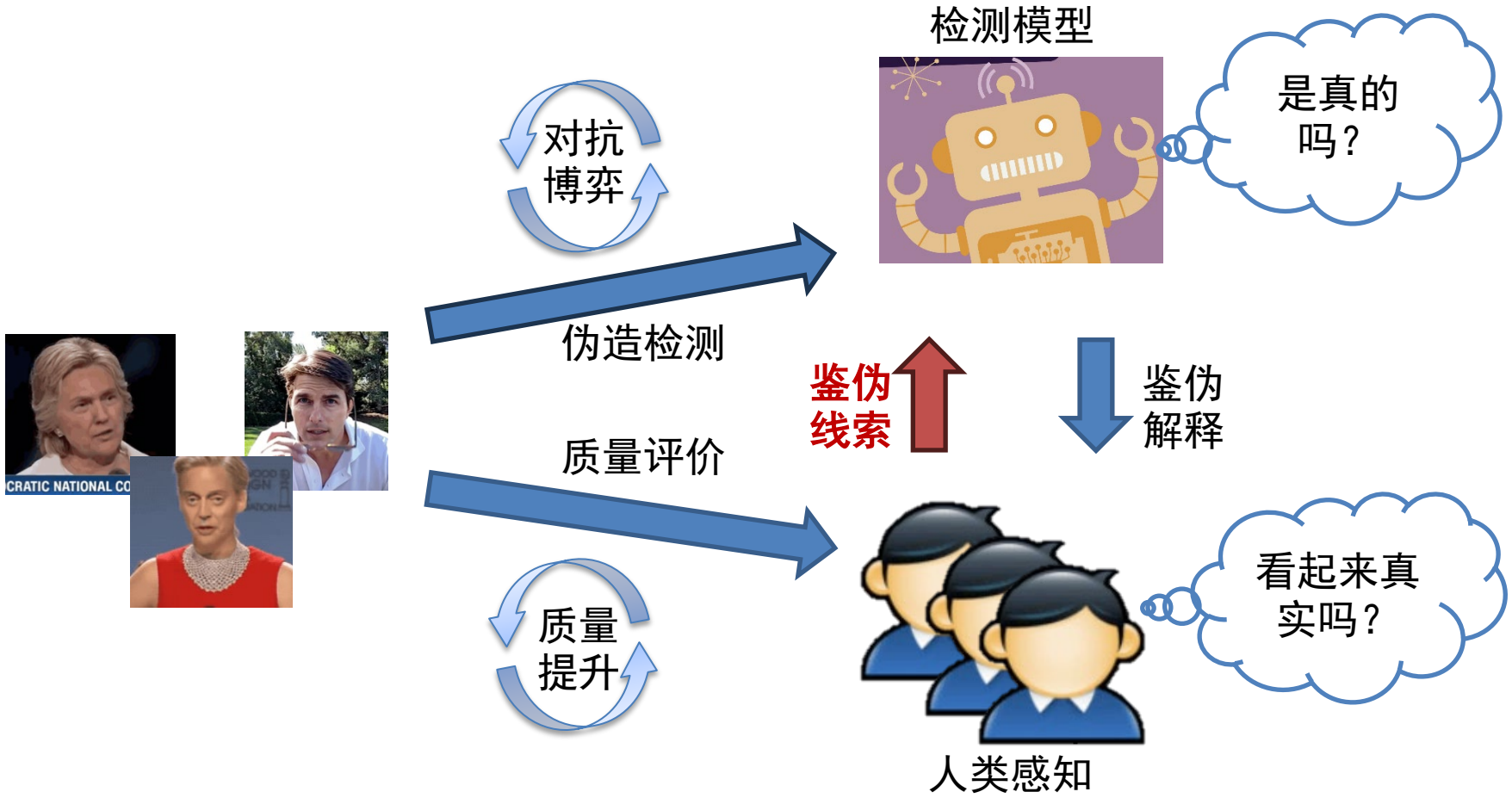
目前的形式上，直接优化过于困难，涉及到Hessian矩阵的求解问题，为此我们通过Taylor展开的方式，将其近似为以下可计算方式。

$$\min_{\theta}(1-\alpha)L(x, y, \theta) + \alpha L_E(f'_{\theta_s}(x), y, \theta_d)$$

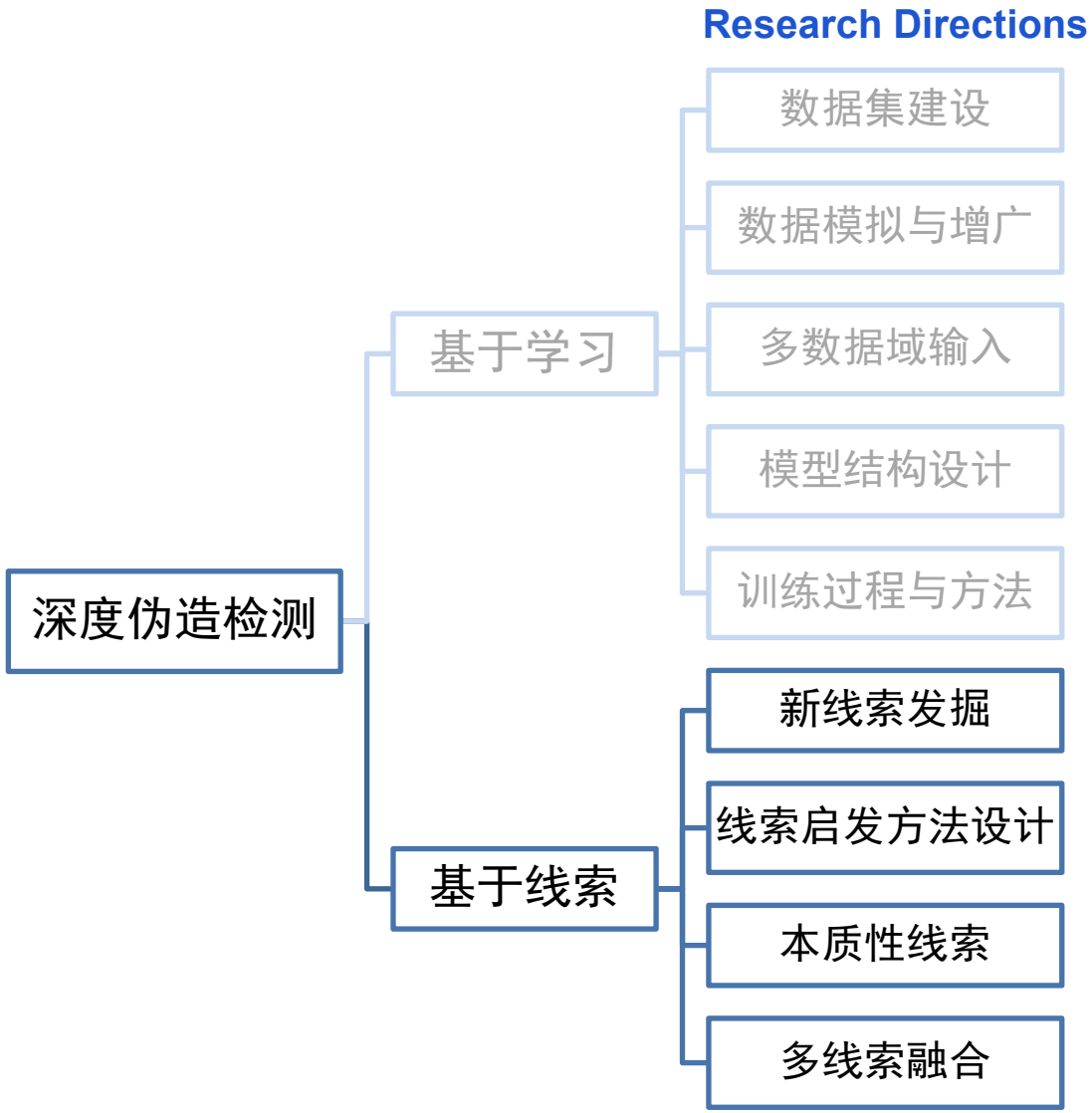
扰动后的浅层特征

Weinan Guan, Wei Wang, Jing Dong, Bo Peng, "Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization", IEEE Transactions on Information Forensics & Security (TIFS), 2024.

研究概览



深度伪造检测方法


Research Directions
深度伪造检测
基于学习
数据集建设
数据模拟与增广
多数据域输入
模型结构设计
训练过程与方法
基于线索
新线索发掘
线索启发方法设计
本质性线索
多线索融合

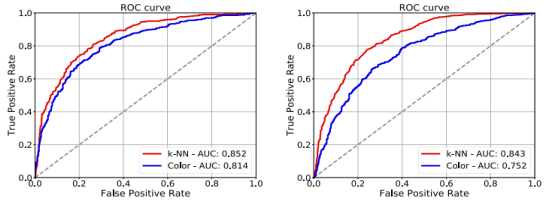
视觉瑕疵线索

- 早期深度伪造结果中存在一些视觉瑕疵，如瞳色不一致、局部细节缺失、拼接边缘等，针对每种瑕疵，设计和提取专门的手工特征，基于此训练分类器实现伪造人脸检测。

Different Eye Color

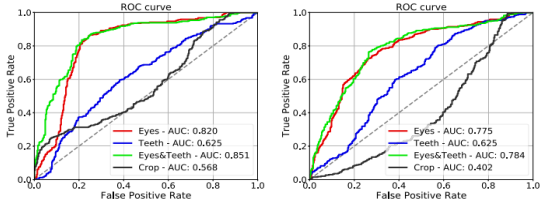
Dissimilarity in eye colors

$$\text{Dist}_H = \min(|l_H - r_H|, 360 - |l_H - r_H|)$$
$$\text{Dist}_S = |l_S - r_S|$$
$$\text{Dist}_V = |l_V - r_V|$$
$$\text{Dist}_{HSV} = \text{Dist}_H + \text{Dist}_S + \text{Dist}_V \ .$$



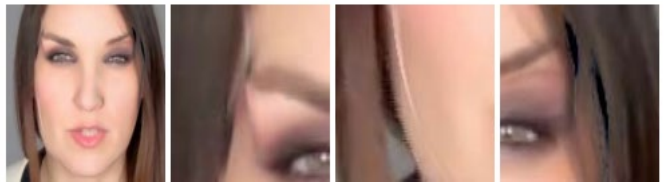
(a) Faces generated by ProGAN (b) Faces generated by Glow

Missing Details &



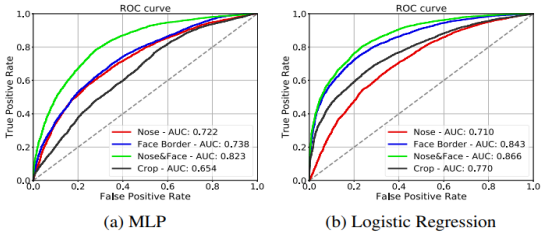
(a) MLP (b) Logistic Regression

Splicing Edges



提取纹理特征

We choose the **texture energy** approach [23] by Laws to generate features that **describe the complexity of the texture**

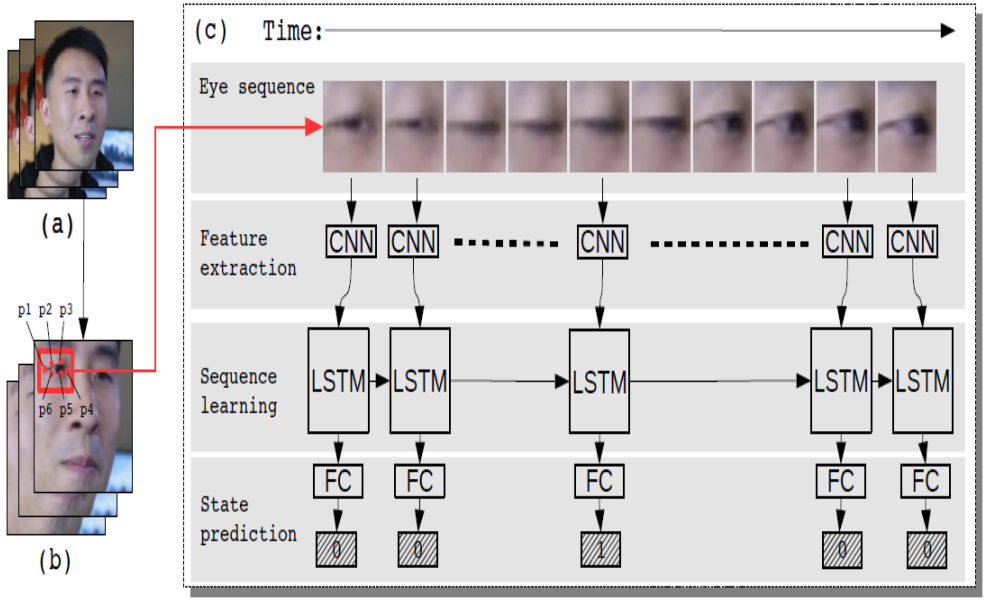


(a) MLP (b) Logistic Regression

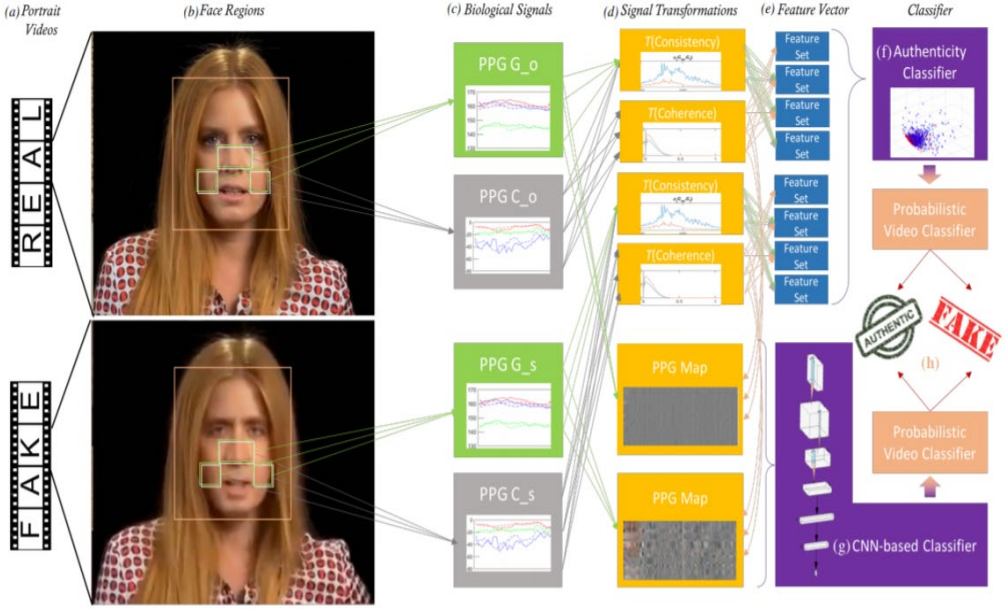
Matern, Falko, Christian Riess, and Marc Stamminger. "Exploiting visual artifacts to expose deepfakes and face manipulations." 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019.

生理信号线索

- 真实人脸视频中存在的眨眼、脉搏等生理信号指征可以用于鉴别伪造人脸视频。



The Eye-blinking clue, Y. Li et al., WIFS 2018

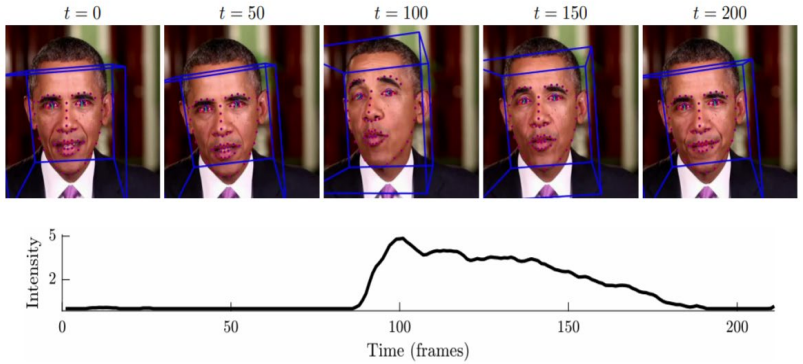


The pulse clue, (Ciftci et al. 2020)

- Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. "In ictu oculi: Exposing ai created fake videos by detecting eye blinking." 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018.
- Ciftci, Umur Aybars, Ilke Demir, and Lijun Yin. "Fakecatcher: Detection of synthetic portrait videos using biological signals." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

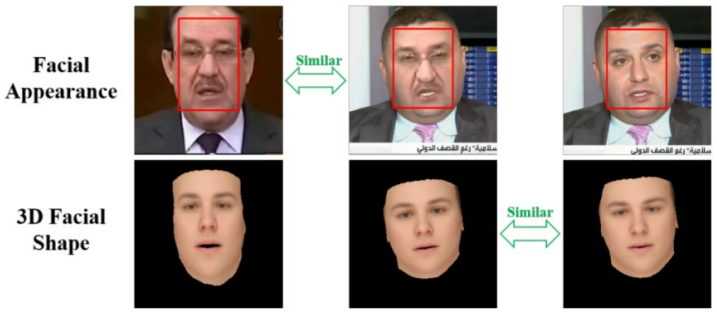
身份一致性线索

- 可以将人脸表观身份这种硬生物识别特征与行为运动特征、三维形状特征、面部周遭特征等软生物识别特征所确定的身份进行比对，以发现深度伪造中隐含的身份不一致问题。



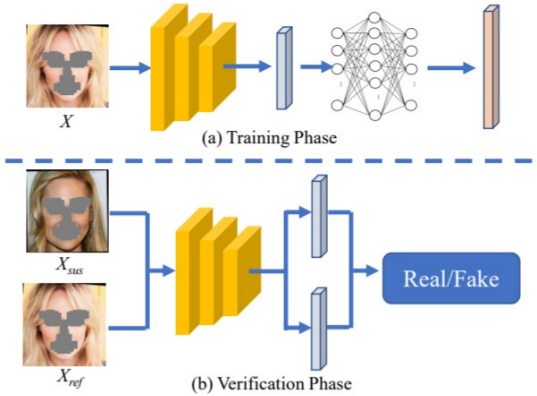
Facial Action Unit （FAU）intensity along time

表观身份信息与头面部运动特质（一种软生物特征）不一致。



Mismatch between appearance and shape

表观身份信息与面部3D形状（一种软生物特征）不一致。



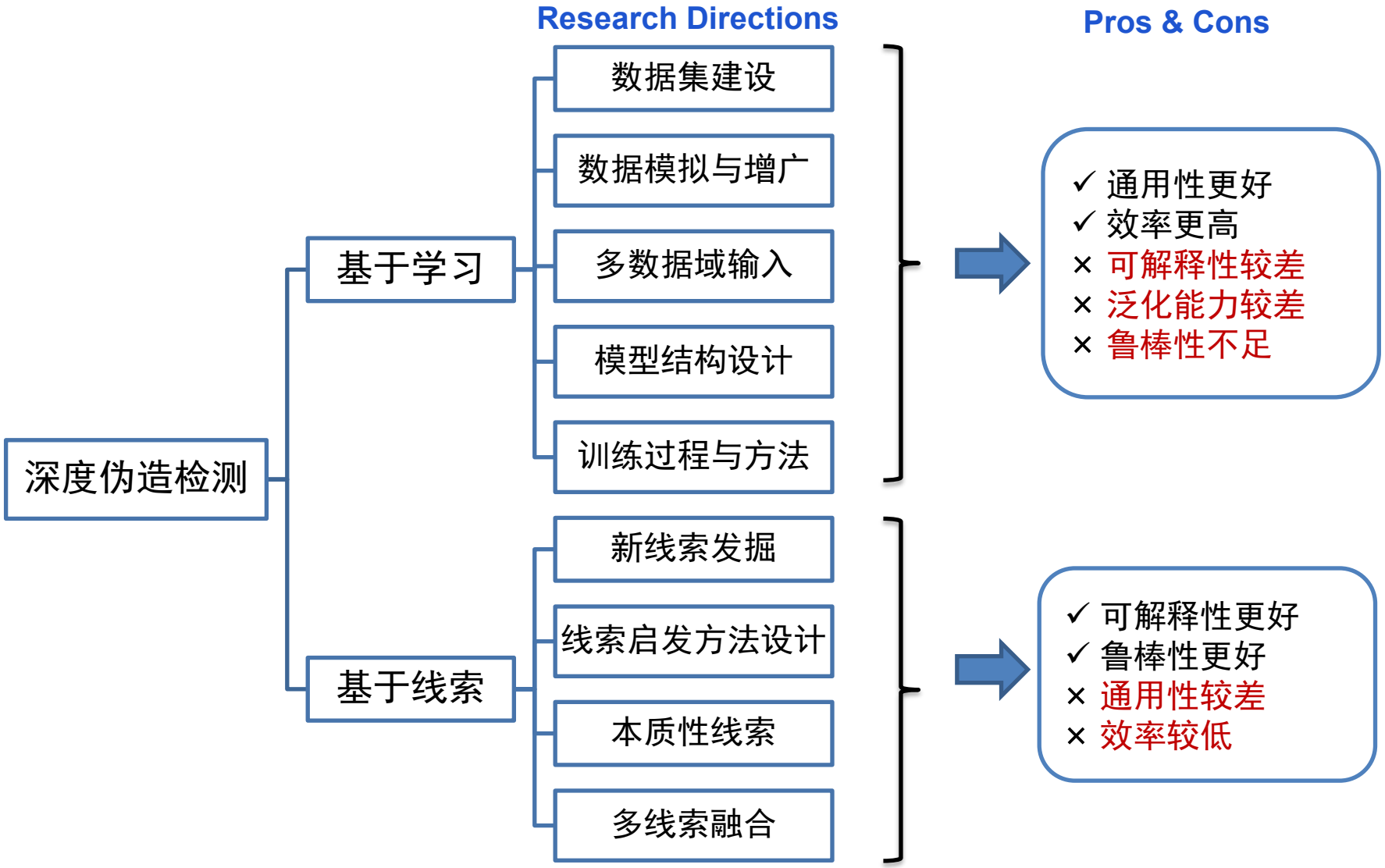
表观身份信息与周遭身份信息（一种软生物特征）不一致。

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019, June). Protecting World Leaders Against Deep Fakes. In CVPR workshops

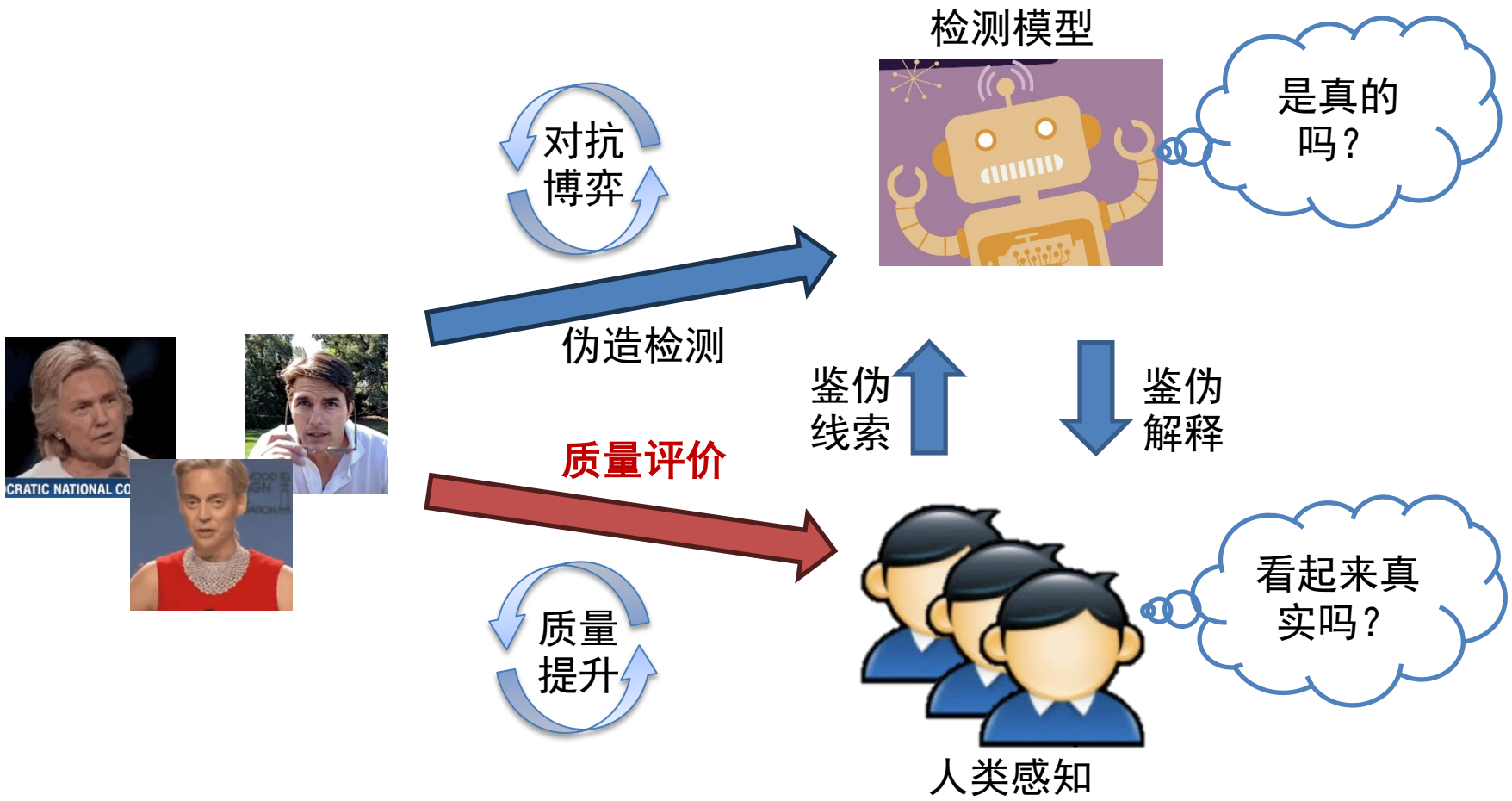
Guan, W., Wang, W., Dong, J., Peng, B., & Tan, T. (2022, August). Robust face-swap detection based on 3d facial shape information. In CAAI international conference on artificial intelligence (pp. 404-415).

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Chen, D., ... & Guo, B. (2020). Identity-Driven DeepFake Detection. arXiv preprint arXiv:2012.03930.

深度伪造检测方法小结


Research Directions
Pros & Cons
深度伪造检测
基于学习
数据集建设
数据模拟与增广
多数据域输入
模型结构设计
训练过程与方法
✓ 通用性更好
✓ 效率更高
× 可解释性较差
× 泛化能力较差
× 鲁棒性不足
基于线索
新线索发掘
线索启发方法设计
本质性线索
多线索融合
✓ 可解释性更好
✓ 鲁棒性更好
× 通用性较差
× 效率较低

研究概览



视觉质量评估

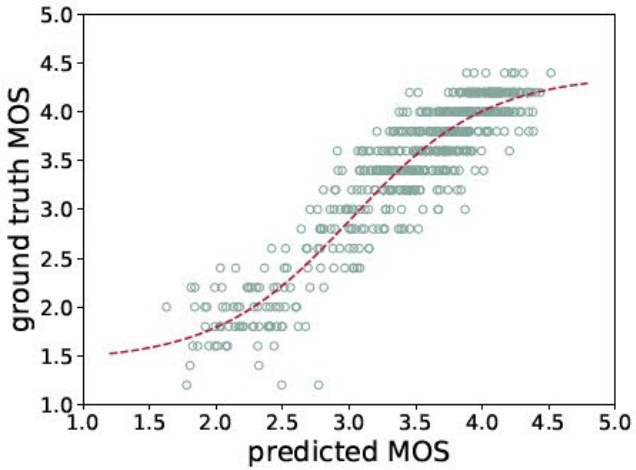
- 图像/视频质量评估（IQA/VQA）是一个经典的研究课题，主要评估由于拍摄和传输中的压缩或拍摄条件所引起的自然图像/视频的整体质量退化程度，评价主体是人，客观评价指标或方法需要逼近人的打分。

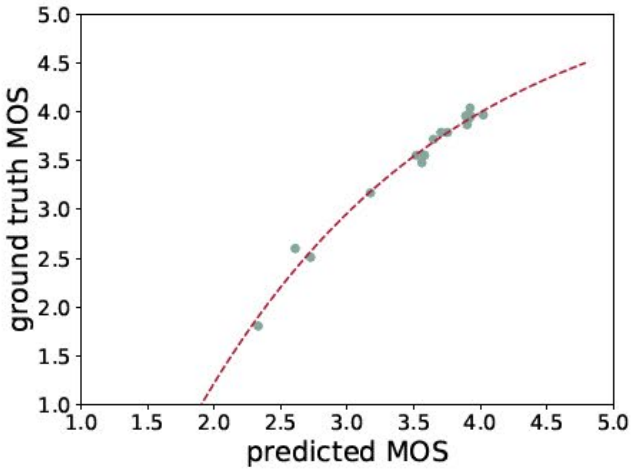
(a) LIVE-VQC (b) KoNViD-1k (c) YouTube-UGC

Tu, Zhengzhong, et al. "UGC-VQA: Benchmarking blind video quality assessment for user generated content." IEEE Transactions on Image Processing 30 (2021): 4449-4464.

深度伪造视频真实感评估

- 不同于自然图像，合成数据中影响视觉质量的因素更多体现在**局部伪造痕迹**上，视觉真实感是评价深度伪造视频质量的一个重要方面。
- 构建了首个用于面部交换视频的视觉真实感评估的基准。尝试回答: **HOW PERCEPTIVELY REAL ARE THE FACE-SWAP VIDEOS?**





Xianyun Sun, Beibei Dong, Caiyong Wang, Bo Peng, Jing Dong, “Visual Realism Assessment for Face-Swap Videos”, International Conference on Image and Graphics (ICIG), 2023

深度伪造视频真实感评估

- 在所构建的数据集（DFGC-2022）测试取得了较好的表现，尤其是用于深度伪造检测的预训练特征最为有效，证明了鉴伪和评估两个任务的强相关性。

(a) Performance under video level facial-id split

Metric	SRCC↑(std)	PLCC↑(std)	RMSE ↓(std)
BRISQUE	0.2646(.104)	0.4185(.124)	0.6473(.055)
GM-LOG	0.4324(.097)	0.5630(.088)	0.5907(.053)
FRIQUEE	0.5281(.084)	0.6926(.078)	0.5134(.059)
TLVQM	0.3988(.081)	0.5586(.096)	0.5923(.058)
V-BLIINDS	0.4042(.114)	0.6251(.123)	0.5502(.071)
VIDEVAL	0.3277(.124)	0.4521(.104)	0.6376(.054)
ensemble	0.6364(.063)	0.7979(.052)	0.4298(.052)
ResNet50	0.6006(.083)	0.7827(.059)	0.4420(.049)
VGGFace	0.5814(.111)	0.7710(.078)	0.4486(.054)
DFDC-ispl	0.5641(.092)	0.7868(.061)	0.4380(.047)
DFGC-1st	**0.7952(.051)**	**0.8975(.028)**	**0.3132(.030)**

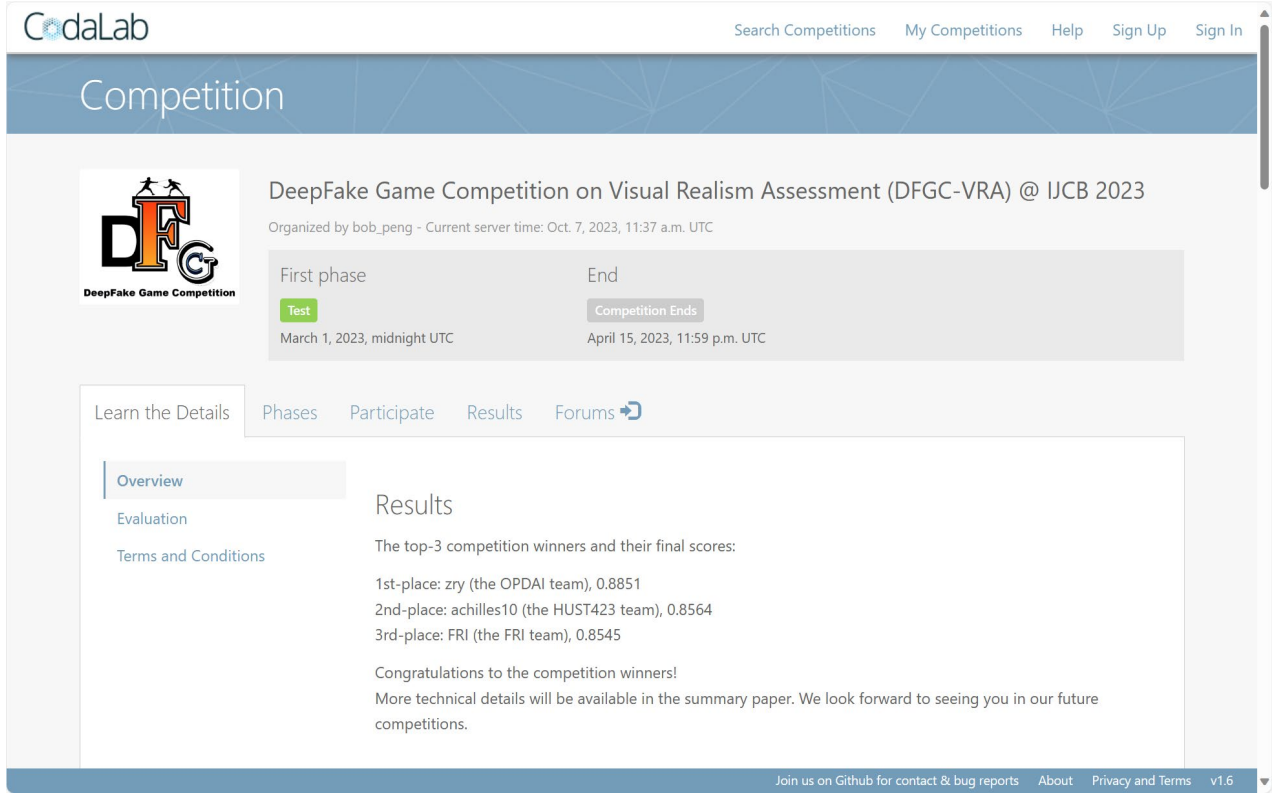
(b) Performance under video level submit-id split

Metric	SRCC↑(std)	PLCC↑(std)	RMSE↓(std)
BRISQUE	0.5379(.202)	0.5803(.198)	0.4208(.135)
GM-LOG	0.5160(.229)	0.5657(.226)	0.4152(.114)
FRIQUEE	0.6481(.165)	0.6928(.175)	0.3536(.082)
TLVQM	0.5593(.195)	0.6165(.203)	0.3097(.096)
V-BLIINDS	0.4851(.235)	0.5316(.247)	0.4166(.096)
VIDEVAL	0.5438(.201)	0.6014(.202)	0.4047(.119)
ensemble	0.7211(.142)	0.7628(.152)	0.3020(.048)
ResNet50	0.7423(.126)	0.7868(.132)	0.2905(.043)
VGGFace	0.7673(.100)	0.7922(.113)	0.3049(.094)
DFDC-ispl	0.7582(.115)	0.8009(.129)	0.2825(.050)
DFGC-1st	**0.8081(.096)**	**0.8356(.106)**	**0.2540(.037)**

Xianyun Sun, Beibei Dong, Caiyong Wang, Bo Peng, Jing Dong, “Visual Realism Assessment for Face-Swap Videos”, International Conference on Image and Graphics (ICIG), 2023

深度伪造视频真实感评估

- DFGC-VRA: DeepFake Game Competition on Visual Realism Assessment 提供比赛数据、评价指标、baseline方法，吸引更多更有效的参赛方案。



	Method-1 Method-2 ··· Method-25	Method-26 ··· Method-35
ID Pair-1 ID Pair-2 ··· ID Pair-14	Train set (700 videos)	Test-2 set (280 videos)
ID Pair-15 ··· ID Pair-20	Test-1 set (300 videos)	Test-3 set (120 videos)

Bo Peng, Xianyun Sun, Caiyong Wang, Wei Wang, Jing Dong, Zhenan Sun, "DFGC-VRA: DeepFake Game Competition on Visual Realism Assessment", IJCB 2023.

深度伪造视频真实感评估

- 获胜方案采用端到端微调训练、鉴伪任务预训练、质量评估领域损失，取得远超baseline方法的表现。

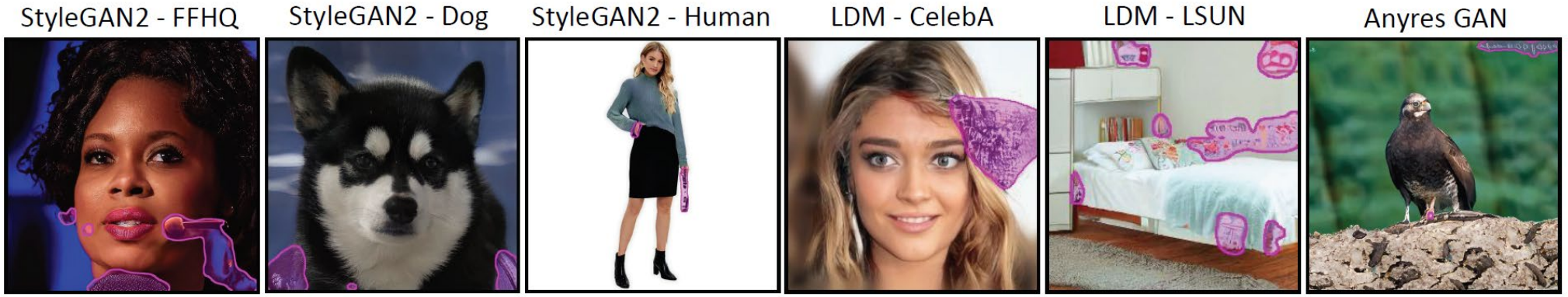
Table 1: Overview of competition results. LB stands for leaderboard results obtained by re-running the inference code and the teams' model checkpoints, and RP stands for reproduced results by re-running the training/fine-tuning codes by organizers.

Team	LB	RP	Backbone	Pre-train	Fine-tune	Loss	Inference
OPDAI	0.8851	0.8825	Swin-transformer	DFDC Det	DFGC-22	Norm-in-norm, KL divergence	3 frames score fusion
HUST	0.8564	0.8474	ConvNeXt, LSTM	ImageNet	DFGC-22 & extra data	MAE, rank, PLCC	20 frames score fusion, 5 models ensemble
UNILJ	0.8545	0.8501	ConvNeXt, Eva	Deepfake Det, ImageNet	DFGC-22	RMSE	10 clips score fusion, 2 models ensemble
USTC	0.8360	0.8116	ResNet152	self-collected face-swap data	DFGC-22	rank	8 frames score fusion
INT&NUST	0.8257	0.8146	ResNext-Transformer hybrid	ImageNet	DFGC-22	PLCC, MSE	4 frames score fusion, 2 streams fusion
Baseline [28]	0.5470	0.5470	ConvNeXt & Swin-transformer	Deepfake Det	SVR on DFGC-22	MSE	regression on video feature

Bo Peng, Xianyun Sun, Caiyong Wang, Wei Wang, Jing Dong, Zhenan Sun, "DFGC-VRA: DeepFake Game Competition on Visual Realism Assessment", IJCB 2023.

AIGC视觉质量评估

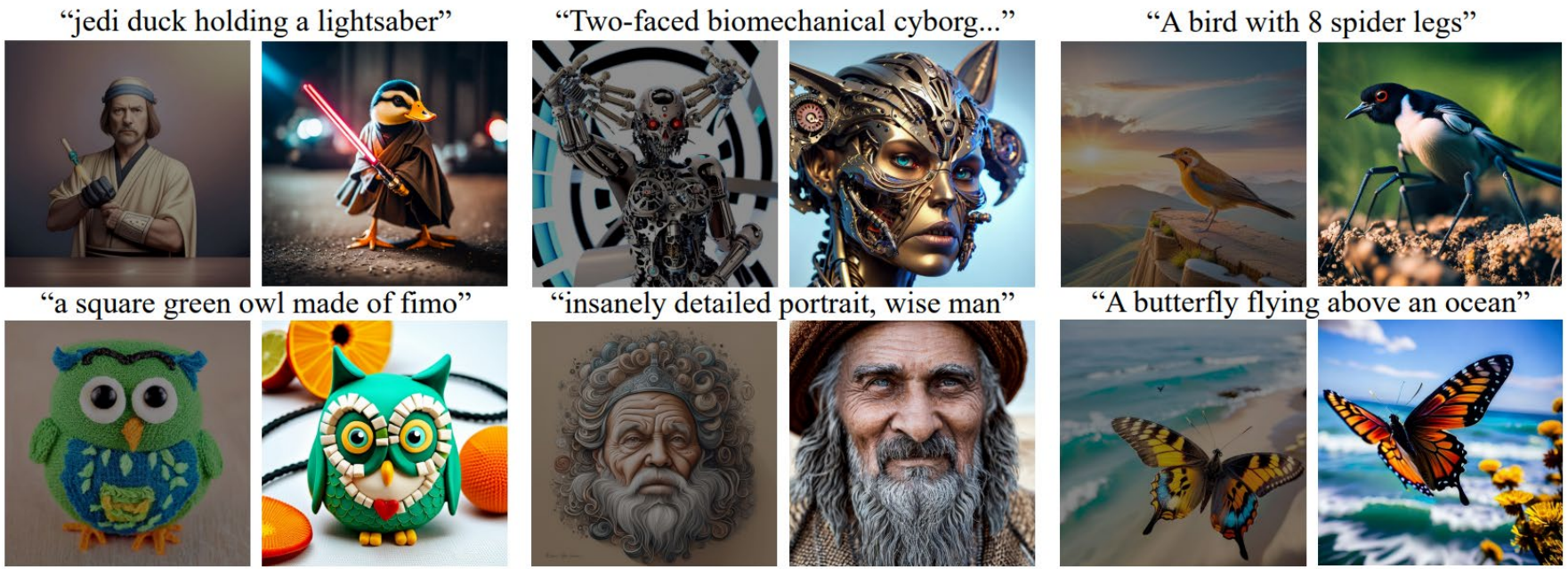
- 最近，AIGC生成图像的视觉质量评估引起了广泛关注。



Perceptual Artifacts Localization for Image Synthesis Tasks, CVPR'2023

FaceScore: Benchmarking and Enhancing Face Quality in Human Generation, arXiv' 2024

AIGC视觉质量评估



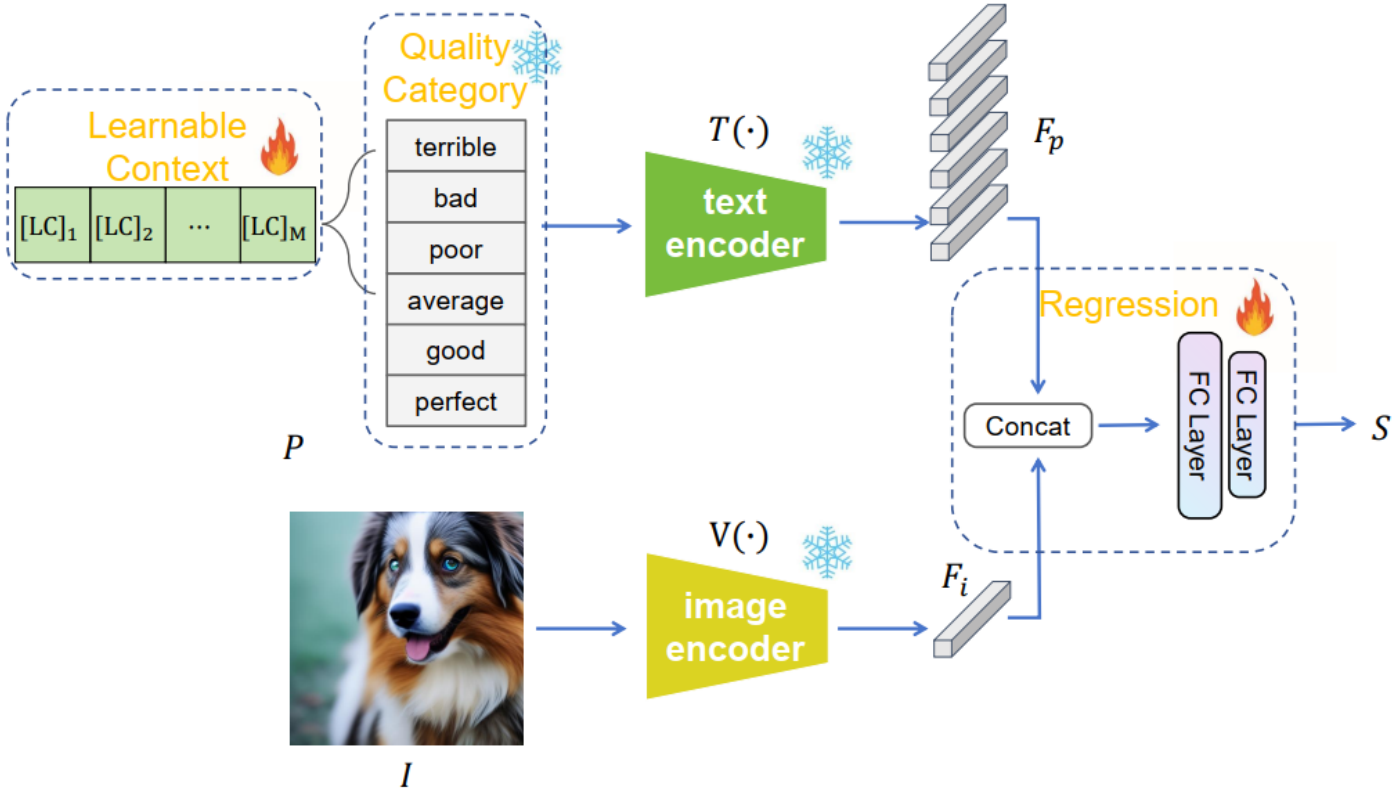
"Pick-a-pic: An open dataset of user preferences for text-to-image generation." NeurIPS 2023.

	Score	Dimension	Image	Ratings	Generation	Public Available
DiffsionDB [15]	No	No	1,819,808	0	Diffusion (1)	Yes
AGIQA-1K [13]	MOS	Perception	1,080	23,760	Diffusion (2)	Yes
Pick-A-Pic [16]	Preference	Overall	500,000	500,000	Diffusion (3)	Yes
HPS [17]	Preference	Overall	98,807	98,807	Diffusion (1)	Yes
ImageReward [14]	Seven Point Likert	Perception; Alignment	136,892	410,676	Auto Regressive; Diffusion (6)	No
AGIQA-3K	MOS	Perception; Alignment	2,982	125,244	GAN; Auto Regressive; Diffusion (6)	Yes

"Agiqa-3k: An open database for ai-generated image quality assessment.“ TCSVT 2023

基于CLIP模型的AIGC视觉质量评估

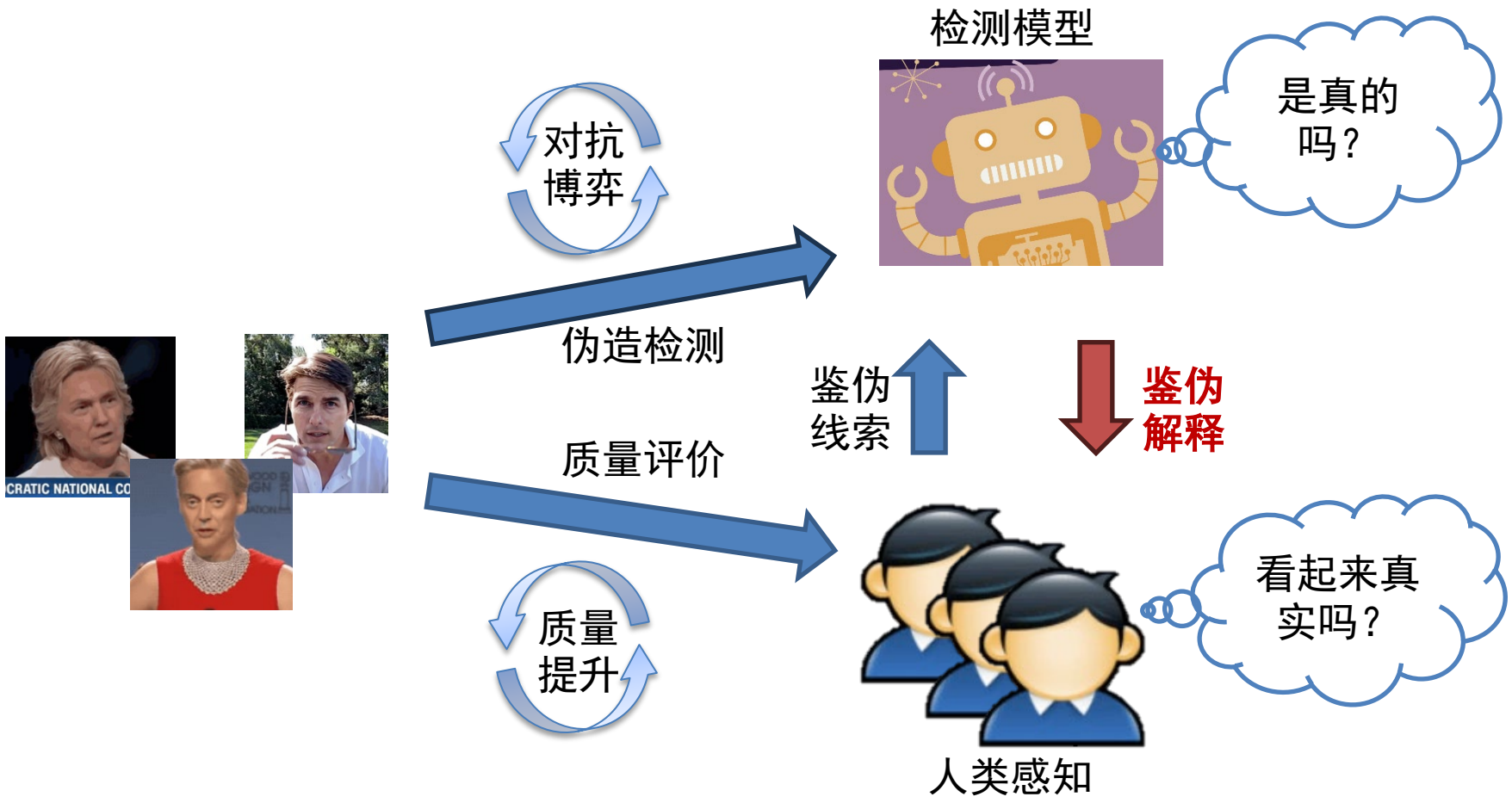
- 利用CLIP中包含的丰富视觉及语言知识，提出多类别可学习提示方法用于AI生成图像的质量评估，在AGIQA-3K与AIGCIQA2023数据集上取得SOTA效果。



Methods	PLCC	SRCC	KRCC
FID [7]	0.1860	0.1733	0.1158
CEIQ [32]	0.4166	0.3228	0.2220
NIQE [17]	0.5171	0.5623	0.3876
GMLF [31]	0.8181	0.6987	0.5119
CNNIQA [9]	0.8469	0.7478	0.5580
DBCNN [35]	0.8759	0.8207	0.6336
CLIP-IQA [24]	0.8053	0.8426	0.6468
CLIPAGIQA(Ours)	0.8978	0.8618	0.6776

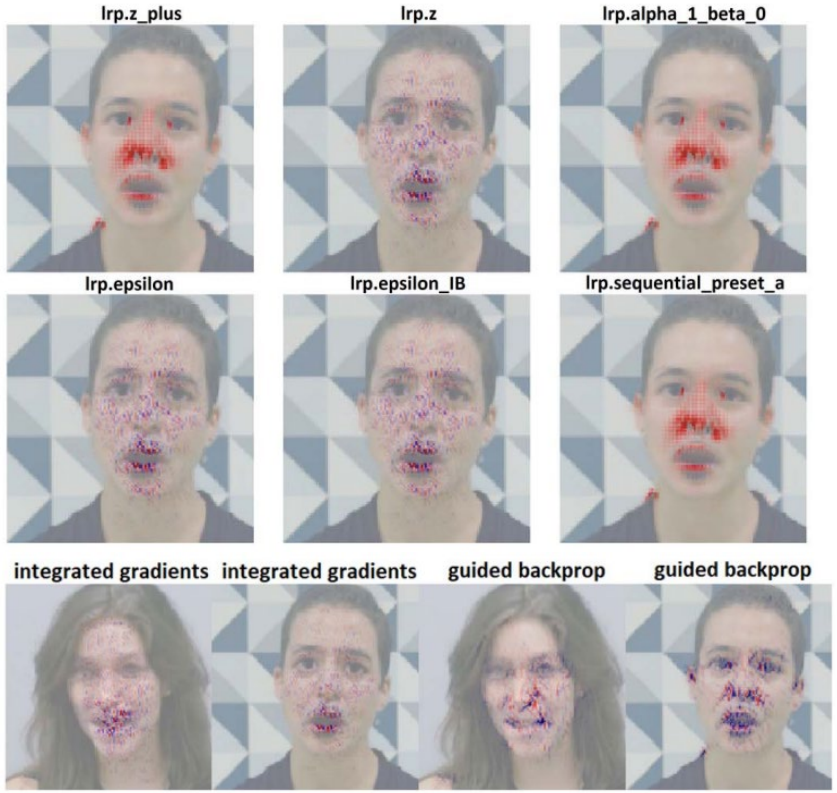
Zhenchen Tang, Zichuan Wang, Bo Peng, Jing Dong, "CLIP-AGIQA: Boosting the Performance of AI-Generated Image Quality Assessment with CLIP", International Conference on Pattern Recognition (ICPR), 2024.

研究概览



深度伪造检测结果的可解释性

- 神经网络鉴伪模型是一个黑盒子，其检测结果往往难以被人类理解。一些研究利用XAI中的归因方法得到对模型决策最为重要的图像区域供人类观察。

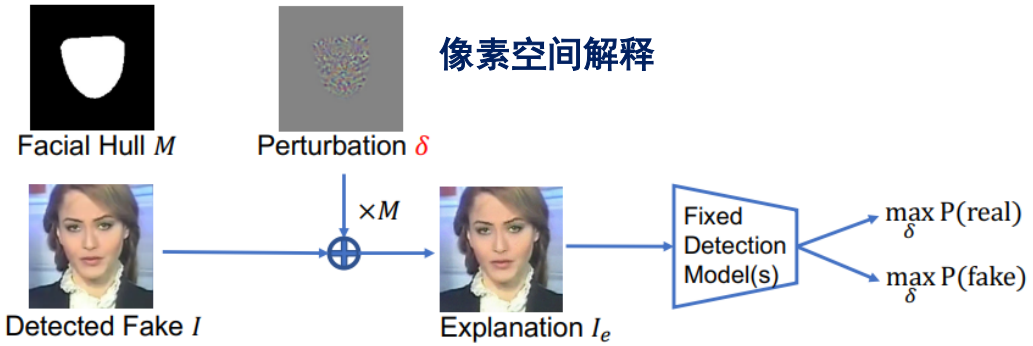


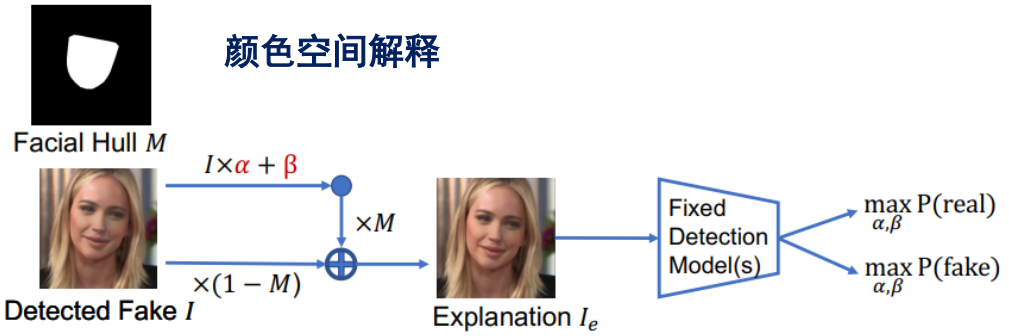
(a) GradCAM (b) LTPA lv. 2 (c) SHAP (d) Bonettini

- Malolan, Badhrinarayan, Ankit Parekh, and Faruk Kazi. "Explainable deep-fake detection using visual interpretability methods." 2020 3rd International Conference on Information and Computer Technologies (ICICT). IEEE, 2020.
- Pino, Samuele, Mark James Carman, and Paolo Bestagini. "What's wrong with this video? Comparing Explainers for Deepfake Detection." arXiv preprint arXiv:2105.05902 (2021).

深度伪造检测的反事实解释

- 区域归因法仍然难以直观解释高质量伪造图像，我们提出一种基于反事实增强的深伪解释方法，分别从像素空间和颜色空间强化或弱化潜在的伪造痕迹，使得人类**更易察觉到换脸图像中的像素伪造痕迹与色彩失调痕迹**。





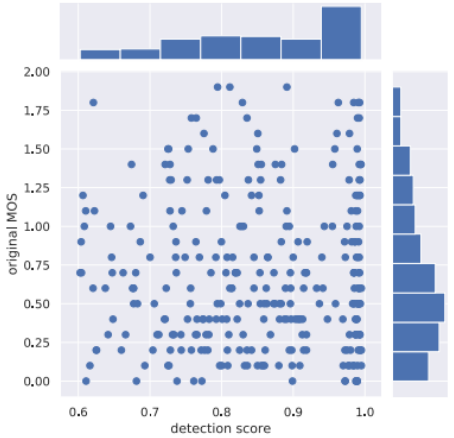
More real　Original　More fake

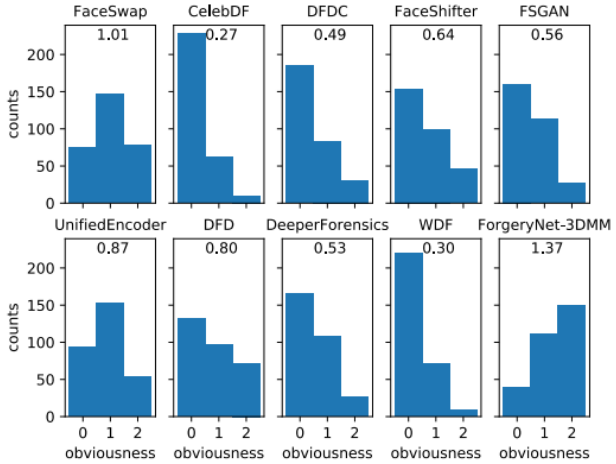
More real　Original　More fake

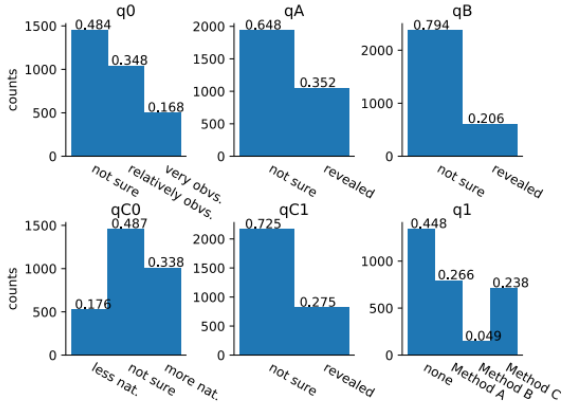
Bo Peng, Siwei Lyu, Wei Wang, Jing Dong. "Counterfactual image enhancement for explanation of face swap deepfakes." Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer Nature Switzerland, 2022.

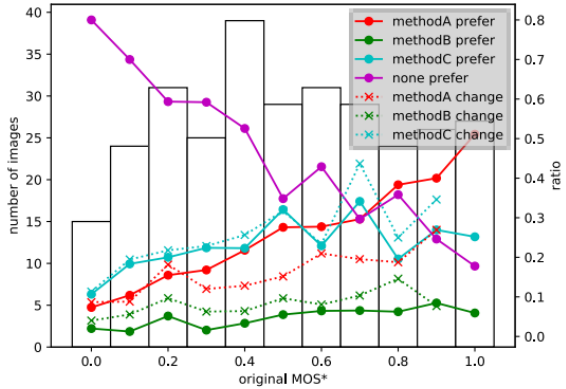
深度伪造检测的反事实解释

- 实验采用了来自10个换脸数据集的伪造图像，共300张.
- 基于DFDC 第一名的预训练模型生成反事实增强结果图像。







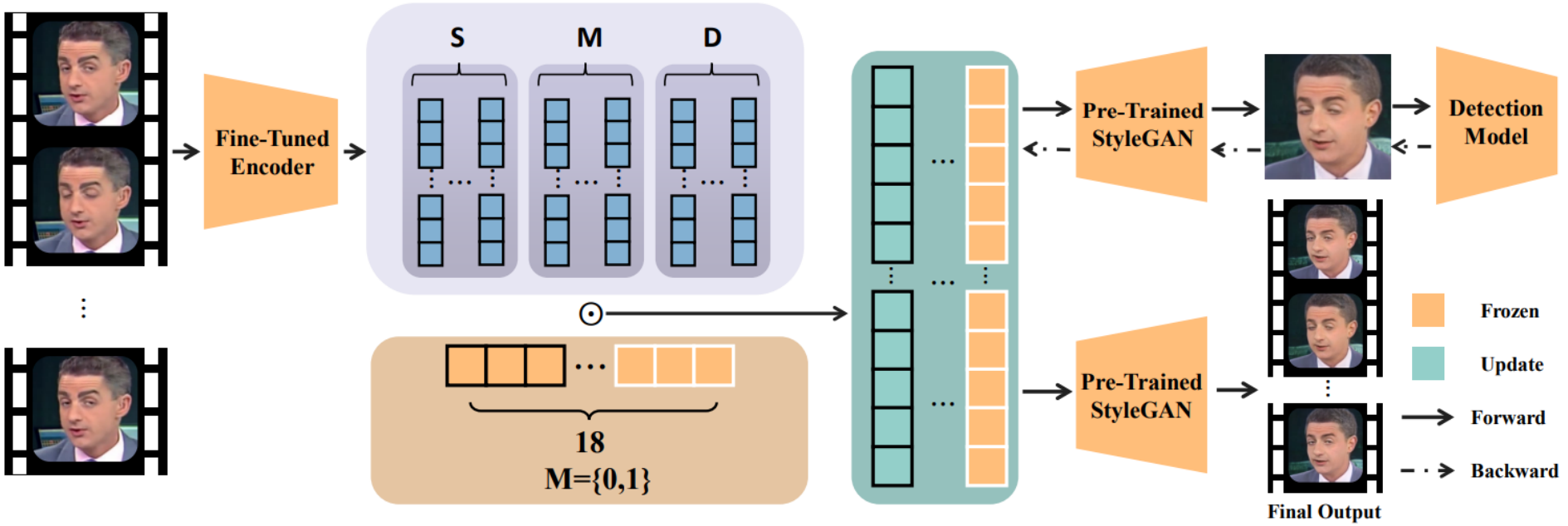


- 招募了10个普通人对300张换脸图像的解释结果进行了主观评估.
- 每张图像采用3种方法进行解释（Grad-CAM, Pixel Space, Color Space），每组结果回答6个问题。
- 实验结果表明所提颜色空间解释方法对于高质量换脸的解释效果最佳。

Bo Peng, Siwei Lyu, Wei Wang, Jing Dong. "Counterfactual image enhancement for explanation of face swap deepfakes." Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer Nature Switzerland, 2022.

深度伪造检测的反事实解释

- 也可以利用微调的编码器和预训练 StyleGAN 生成器来优化伪造人脸图像在StyleGAN隐空间的编码，生成去除伪造痕迹的版本，以进行反事实解释。

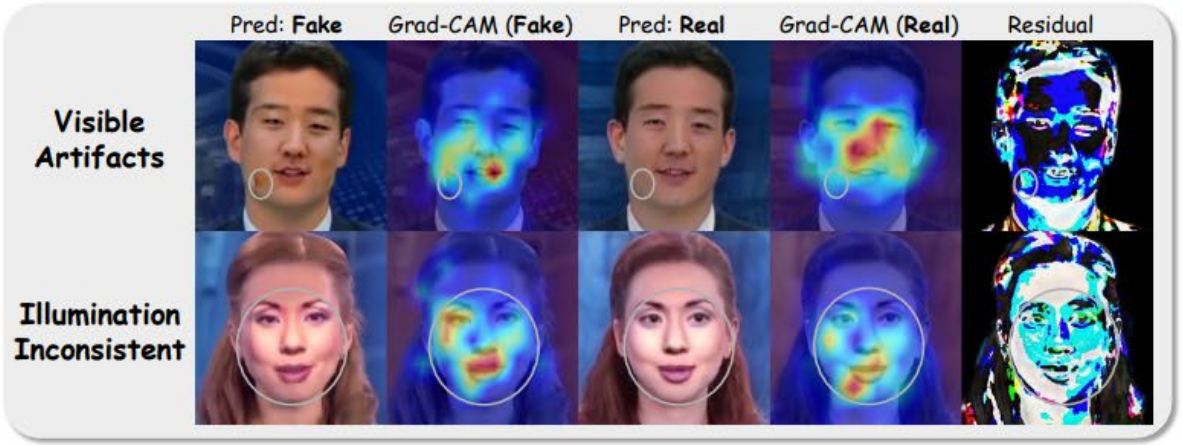


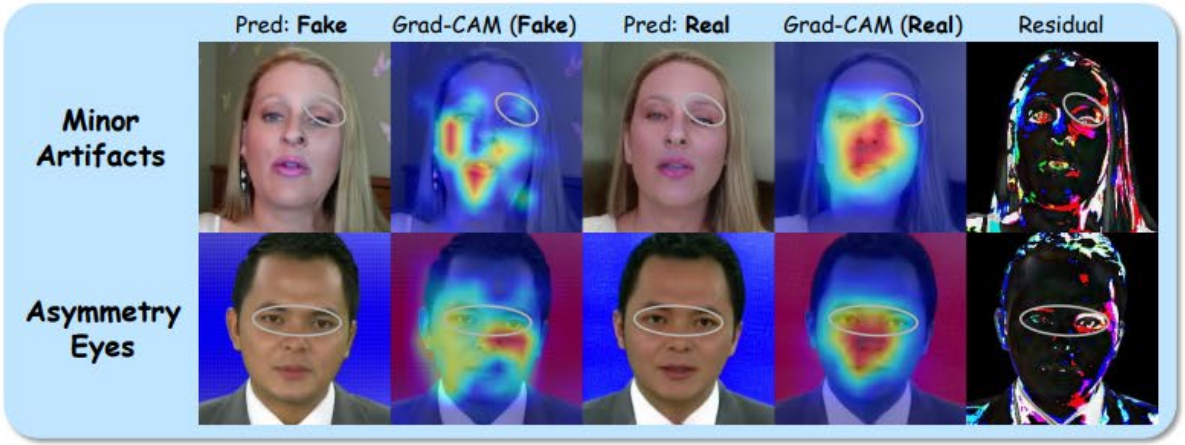
$$\arg\min_{\mathbf{W}_i^{adv}} \mathcal{L}_{adv}(\boldsymbol{D}(\boldsymbol{G}(\mathbf{W}_i^{adv})), y_t),$$

Yang Li, Songlin Yang, Wei Wang, Ziwen He, Bo Peng, Jing Dong, “Counterfactual Explanations for Face Forgery Detection via Adversarial Removal of Artifacts”, IEEE Conference on Multimedia Expo 2024 (ICME).

深度伪造检测的反事实解释

- 基于StyleGAN生成的反事实解释视觉效果更好，但解释结果的Residual稀疏性不够好。

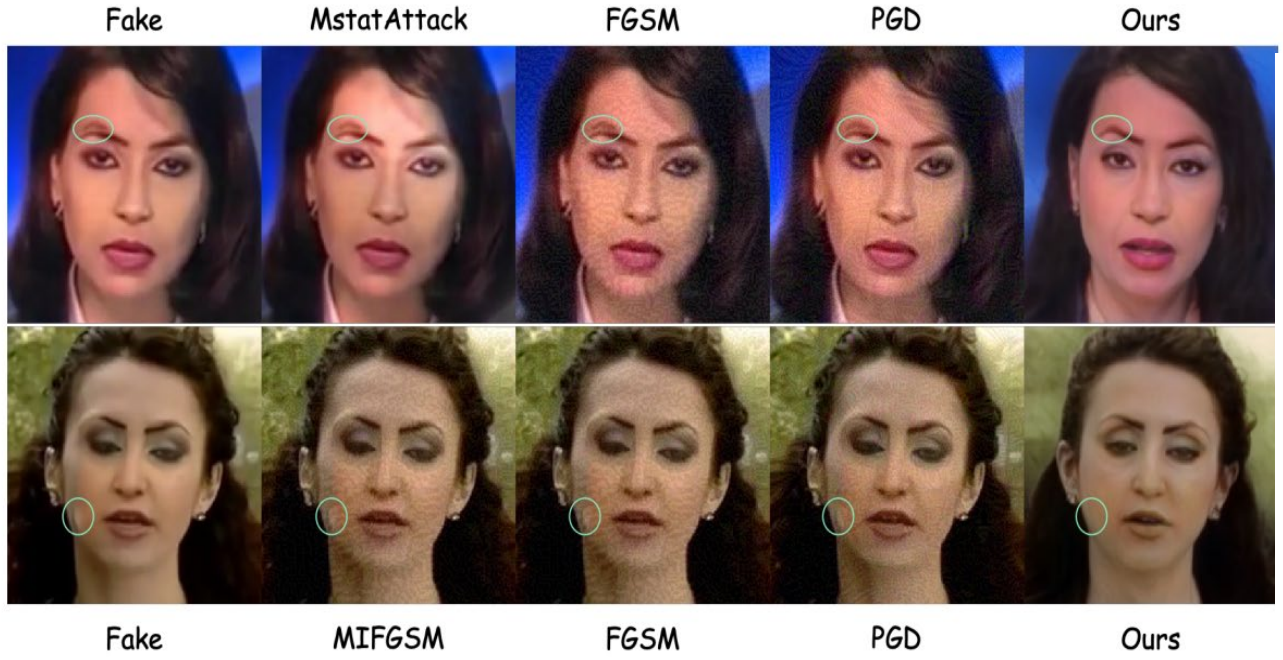




Yang Li, Songlin Yang, Wei Wang, Ziwen He, Bo Peng, Jing Dong, “Counterfactual Explanations for Face Forgery Detection via Adversarial Removal of Artifacts”, IEEE Conference on Multimedia Expo 2024 (ICME).

深度伪造检测的反事实解释

- 与传统对抗样本方法生成的解释结果相比，视觉结果更好，无噪声干扰。
- 本方法产生的解释结果作为对抗样本的迁移性也更好，说明该反事实解释方法确实擦除了具有泛化性的伪造痕迹。

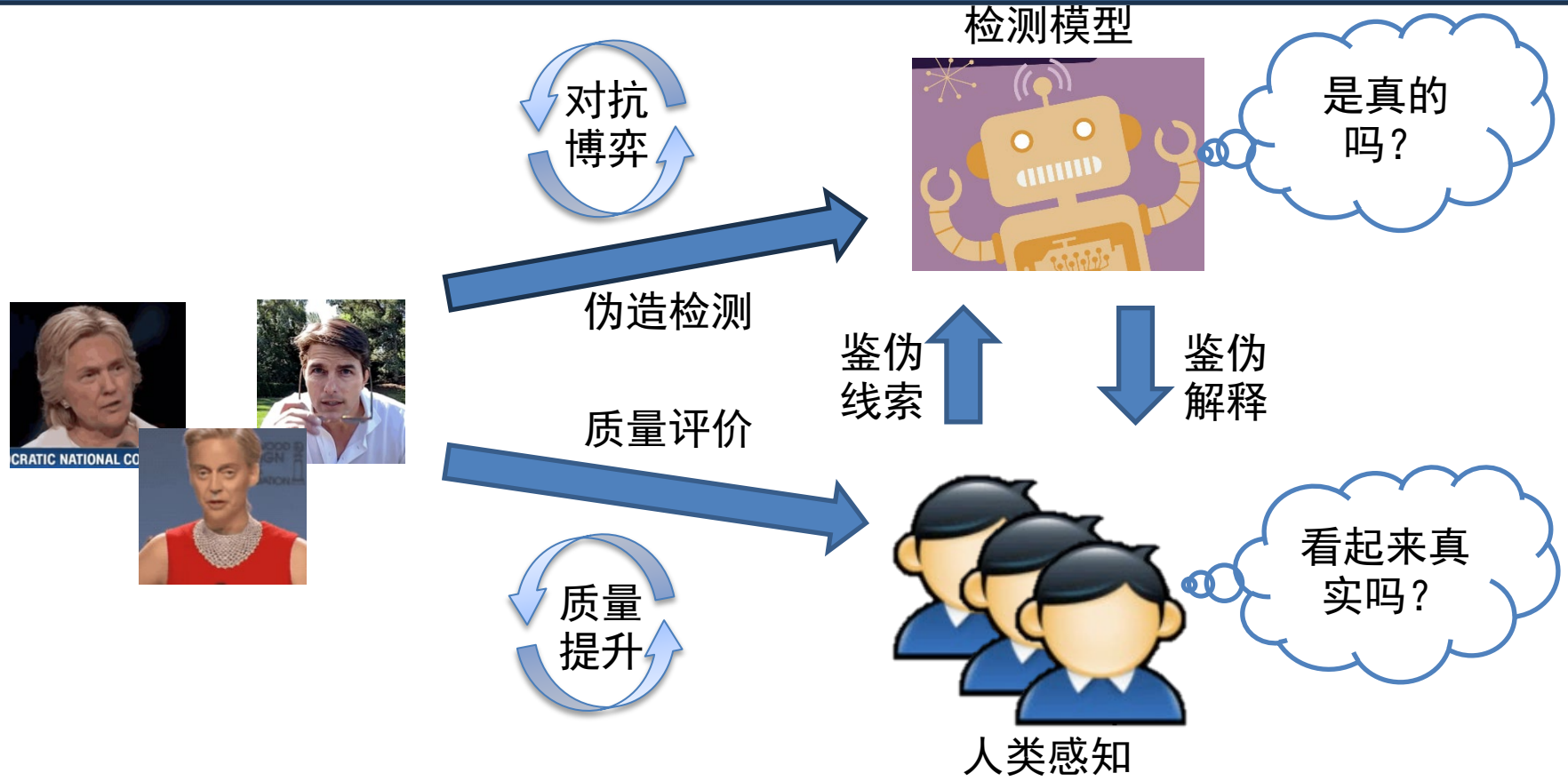


Model	Attack	Efficient-b4	Xception	MAT	RECCE
Efficient-b4	FGSM	99.2	17.2	30.3	12.3
	PGDl_{inf}	**99.9**	18.5	31.2	12.2
	Ours	98.9	**85.7**	**75.8**	**54.0**
Xception	FGSM	4.36	99.3	7.15	26.5
	PGDl_{inf}	4.86	**100**	7.15	26.5
	Ours	**86.1**	99.2	**73.9**	**73.0**
MAT	FGSM	18.0	21.4	99.8	32.9
	PGDl_{inf}	18.0	21.4	**100**	32.8
	Ours	**80.6**	**82.0**	90.6	**70.0**
RECCE	FGSM	9.74	26.0	26.7	99.5
	PGDl_{inf}	9.62	26.0	28.2	**100**
	Ours	**89.6**	**95.0**	**90.0**	98.8

Yang Li, Songlin Yang, Wei Wang, Ziwen He, Bo Peng, Jing Dong, “Counterfactual Explanations for Face Forgery Detection via Adversarial Removal of Artifacts”, IEEE Conference on Multimedia Expo 2024 (ICME).

总结

- 伪造检测主要介绍了基于学习的深度鉴伪方法
- 人类评价经验可以作为鉴伪线索指导检测器设计
- 与主观评价一致的客观质量评估方法随AIGC备受关注
- 黑盒鉴伪模型如何给人类解释亟需进一步研究



中国科学院
自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

谢谢！ Thanks!

Any Questions?