深度换脸 !?

> 深度伪造与检测的浅入浅出

> 2101212850 干皓丞

$ man deepfake

北京大学
PEKING UNIVERSITY

文献军火库 x 资源包 x 感谢

本次报告为科普导向的讲义，但仍有特地搜集相关的研究文献军火库，有兴趣的同学可以根据此 GitHub 连接，进行检索。

另外本次报告与同为课堂的郭博菲同学合作连动，故部分内容需要参考郭同学报告部分。

最后感谢在此次撰写档时，台湾大学李宏毅老师在此次报告给学生的建议，让学生此撰写期末报告获得研究上启发。

目 录

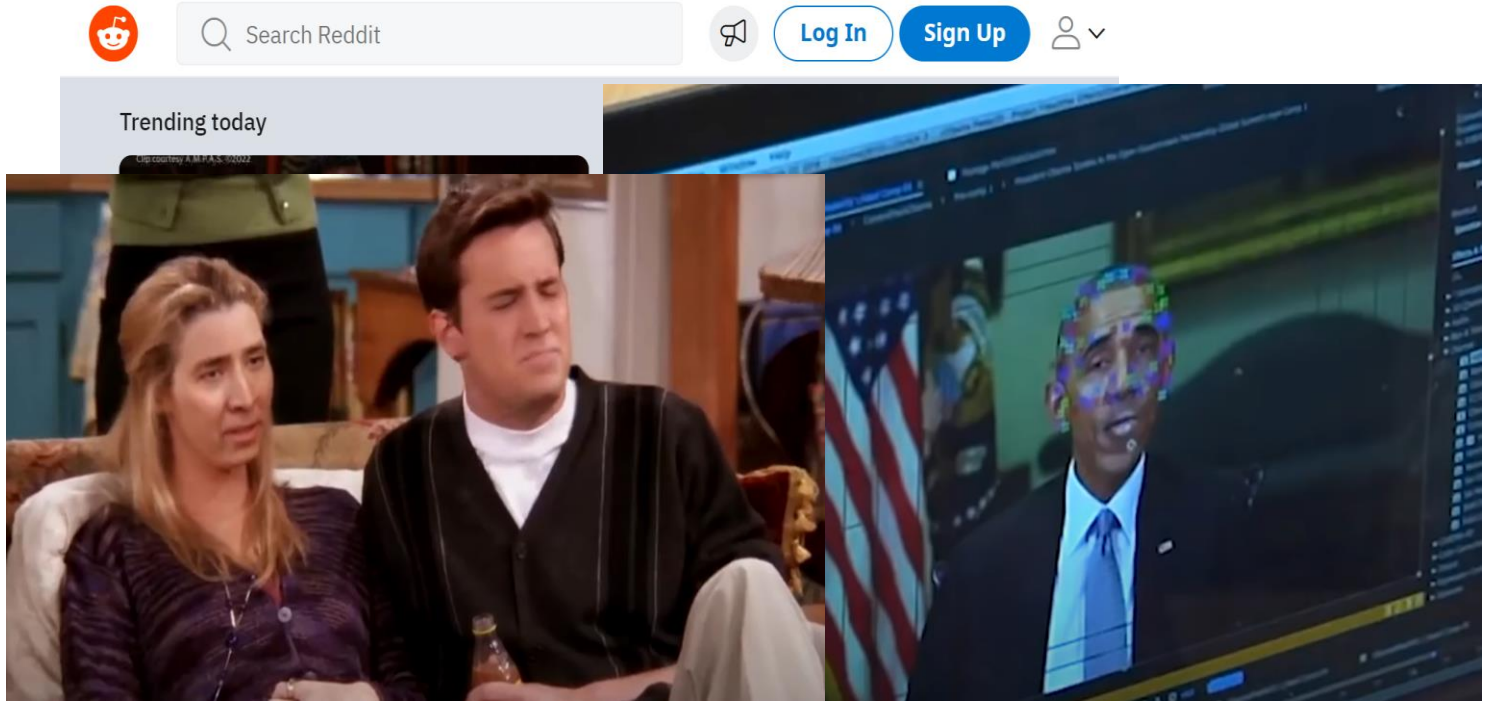
CONTENTS

1. 深度伪造概况

1. 2017 年初现在所谓的 Deepfake 的词出现于 reddit 上。

2. 由英文深度 Deep 和伪造 Fake 所组成。

3. 最早由影星替換換到所有電影與美劇角色尼可拉斯凯吉的搞笑影片与美国前总统欧巴马的伪造影像。



https://www.youtube.com/watch?v=WgpMz7PEZnc

1. 深度伪造概况

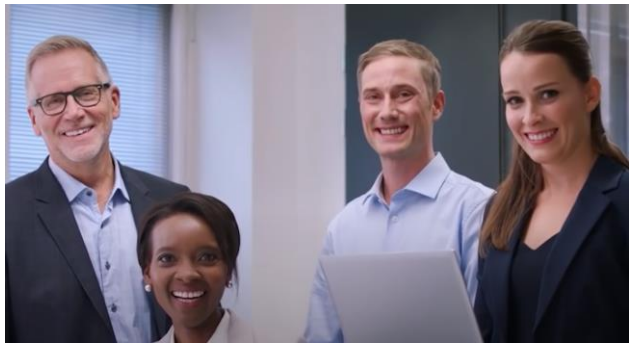
1. 如过往电影公司，电影重启多年后的续集时运用深伪造技术进行处理，让演员的年纪与当时一致。比如蜘蛛人无家日跟星际大战。

2. 或者由于演员因为拍摄过程中意外离世，电影公司运用其技术，拍摄后面的工作用此技术进行处理，将电影完成，让粉丝缅怀。

3. 营销公司根据模特儿产生不存在的脸孔，降低真人所产生的成本。

4. 让古老的相片成为栩栩如生的动画

5. 最后则是执法单位协助寻找失踪人口，运用其技术，让失散多年的家人重聚。



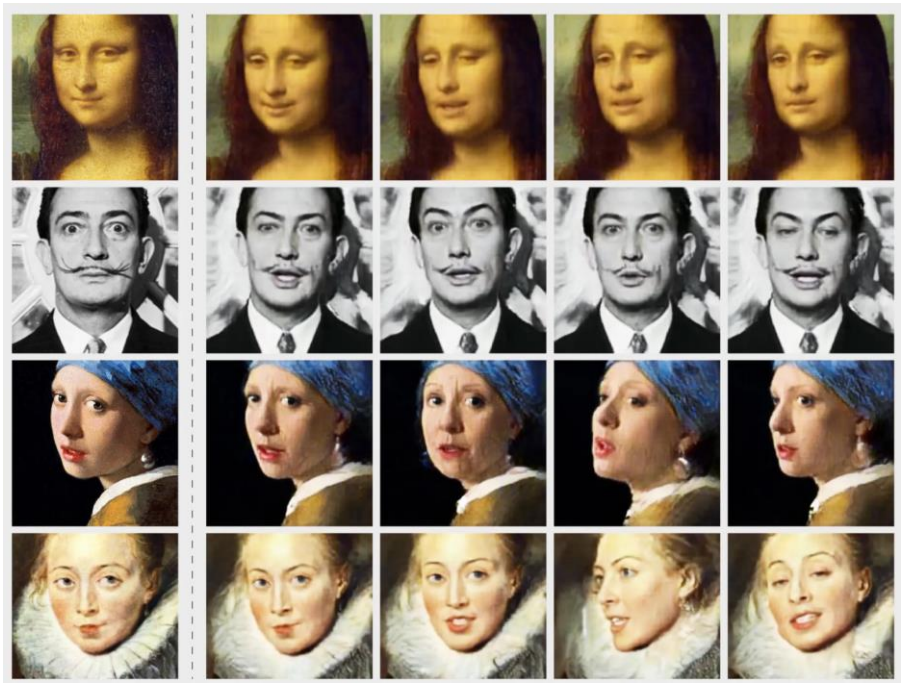
https://www.youtube.com/watch?v=WgpMz7PEZnc

1. 深度伪造概况

1. 未经过允许，经由其技术换脸到成人作品。而受害的当事人除了名誉与事业受到伤害之外，当事人向基层执法单位求助时，因为基层执法单位对该领域不理解，导致受害人在描述时遭到心理上二次伤害。而在台湾地区则正进行修法最高七年有期徒刑。

2. 政治上的宣传战，制作相关政治人物不正确的发言，从而导致冲突，用来打击对手选举。因此美国加州、德州修法，不得用此技术干扰选举。

3. 同时该技术也于近期的战争，并正式导入于战争中双方在社群媒体的宣传战。



https://www.youtube.com/watch?v=WgpMz7PEZnc

2. 深度伪造生成技术

深度伪造在生成技术上的分类

1. 根据图形学的方式进行伪造，找出人脸的关键点，通过 3D 模型进行处理。

2. 使用 GAN 又或者将其视为類似 CNN 卷積的风格迁移的方式进行处理。将想要的人脸替换到目标人物。

3. 运用文本到语音合成(text-to-speech synthesis , TTS）或者语音转换的方式（voice conversion），进行语音上的深度伪造。

4. 指不改变目标的人脸，而是运用传统图形学或者深度学习技术，抓取特征来改变目标人物的脸部表情。

2. 深度伪造生成技术

深度伪造在生成技术上的分类对应研究

Nirkin Y et al.
1704.06729

Garrido P et al.
10.1109/CVPR.2014.537

Garrido P et al.
10.1111/cgf.12552

Arik S et al.
1704.06729

Zakharov E et al.
1905.08233

2. 深度伪造生成技术

深度伪造从人类视觉上角度的分类

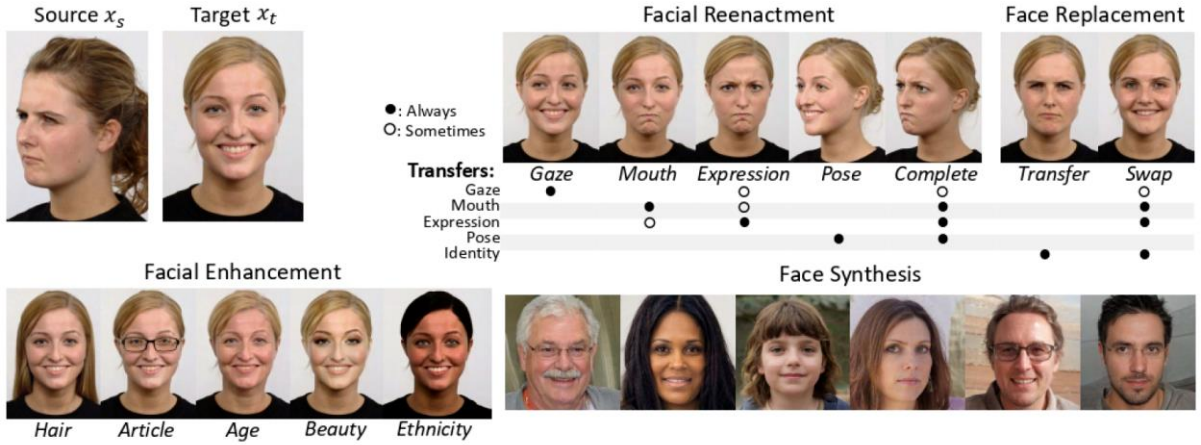


Fig. 2: Examples and illustrations of reenactment, replacement, editing, and synthesis deepfakes of the human face.

Yisroel Mirsky et al.
2004.11138

用 S 来驱动 T

1. 重演：deepfake 是 s 用于驱动 t 的表情、嘴巴、凝视、姿势或身体的地方。

2. 替换：deepfake 是将 t 的内容替换为 s 的内容，保留 s 的身份。比如模特儿的更衣。

3. 增强: t 的属性所在被添加、更改或删除。比如头发等。

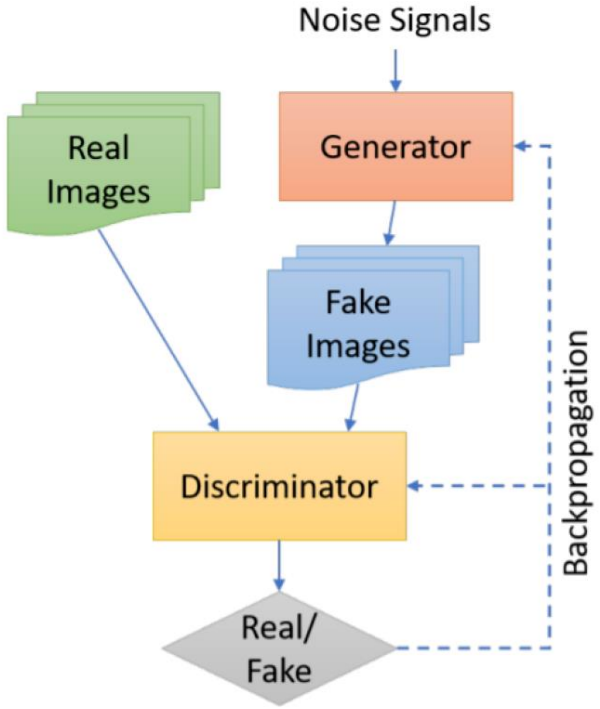
4. 合成：没有目标作为基础的情况下 deepfake 。人脸和身体合成技术，可以创造没有版权的虚拟人物。

2. 深度伪造生成技术

深伪在生成的 Generative Adversarial Networks

北京大学 PEKING UNIVERSITY

Goodfellow I et al.
1406.2661



Yisroel Mirsky et al.
2004.11138

Thanh Thi Nguyen et al.
1909.11573

生成網路盡可能地想辦法去欺騙判別網路，讓兩者對抗，不斷調整細節。

https://www.youtube.com/watch?v=3vHvOyZ0GbY

2. 深度伪造生成技术

深伪在生成的 Generative Adversarial Networks

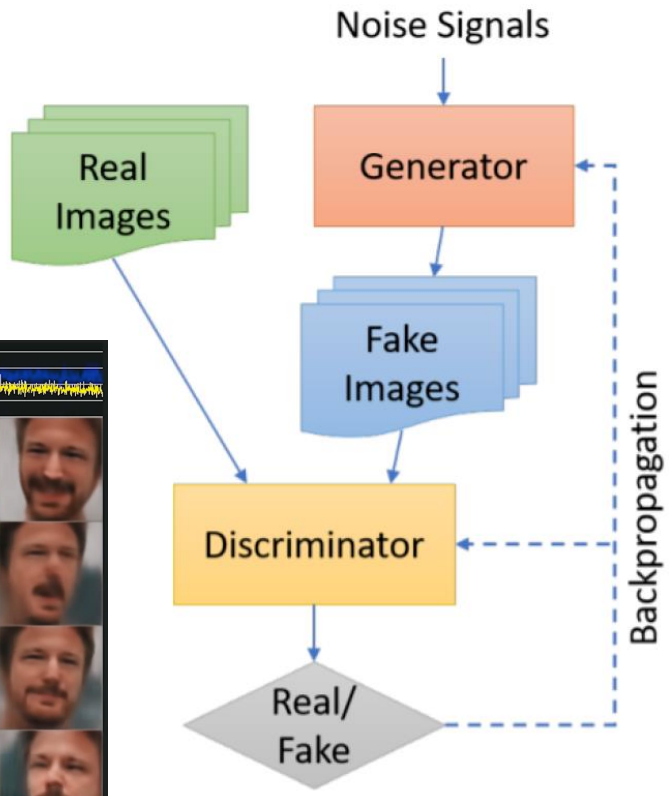


Fig. 3. The GAN architecture consisting of a generator and a discriminator, and each can be implemented by a neural network. The entire system can be trained with backpropagation that allows both networks to improve their capabilities.

生成

判别

https://www.youtube.com/watch?v=3vHvOyZ0GbY

2. 深度伪造生成技术

深伪在生成的 Convolutional Neural Networks

北京大学 PEKING UNIVERSITY

Iryna Korshunova et al.
1611.09577

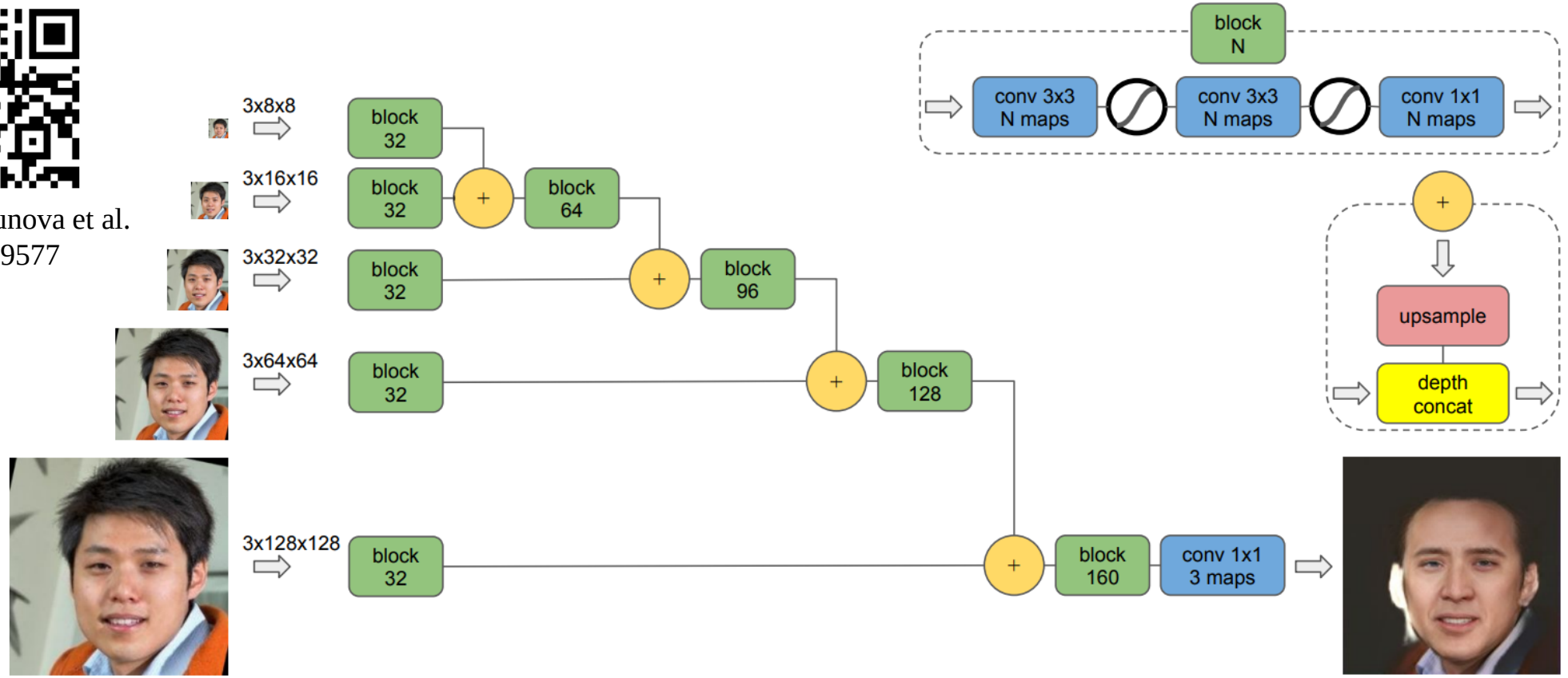


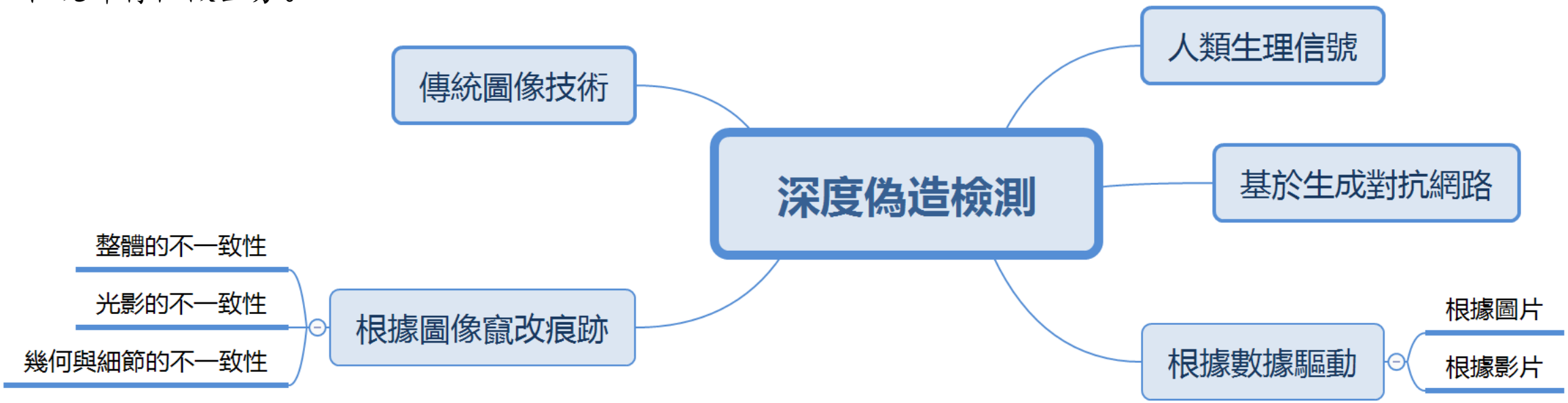
Figure 3: Following Ulyanov *et al.* [31], our transformation network has a multi-scale architecture with inputs at different resolutions.

Korshunova I, Shi W, Dambre J, Theis L. Fast face-swap using convolutional neural networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 3677−3685.

深度伪造检测的分类

1. 使用传统信号处理的方式，利用图像的频域特征和统计特征做区分。

2. 根据人类的生理信号特征。



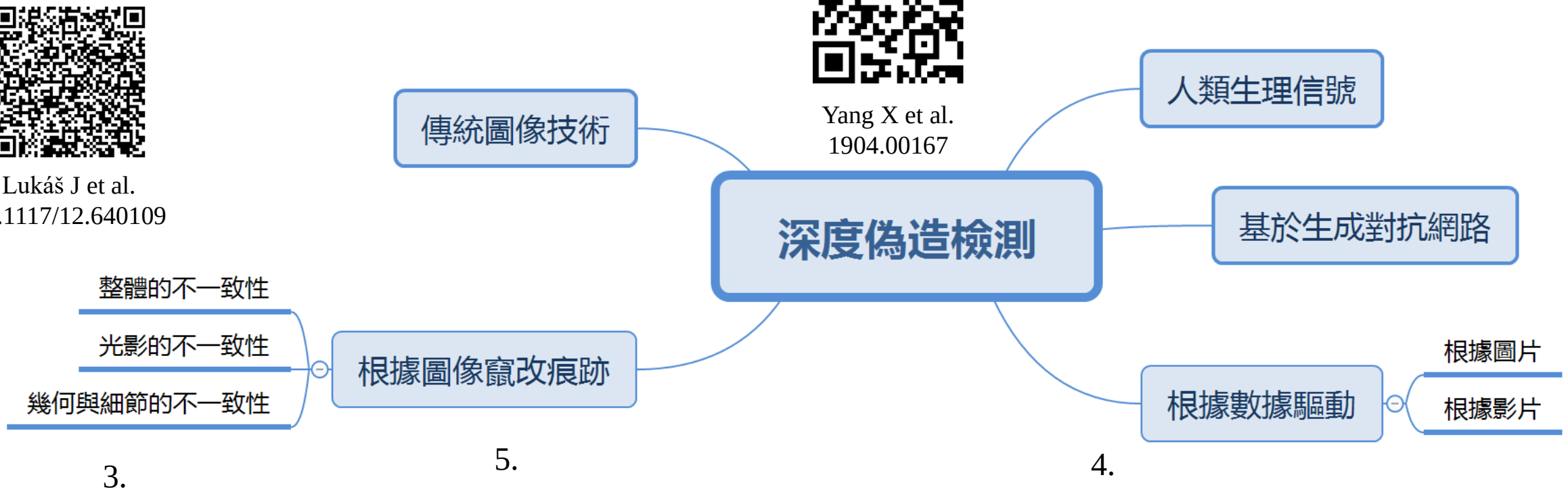
3. 根据模型窜改的痕迹来判断，可以从整体性、光影、牙齿等几何细节上不一致。

4. 目前深度伪造的生成多数是用生成对抗网络，那研究其技术特点就很重要。

5. 針對目前當下模型資料規模逐漸增加的過程下，分為圖片和影片。

3. 深度伪造检测技术

深度伪造检测的分类



1. Lukáš J et al. 10.1117/12.640109

2. Yang X et al. 1904.00167

3. Rössler A et al. 1803.09179

4. Yu N et al. 1811.08180

5. Li X et al. 10.1145/3366424.3382711

Wu Z et al.

深度伪造检测的分类

Thanh Thi Nguyen et al.
1909.11573

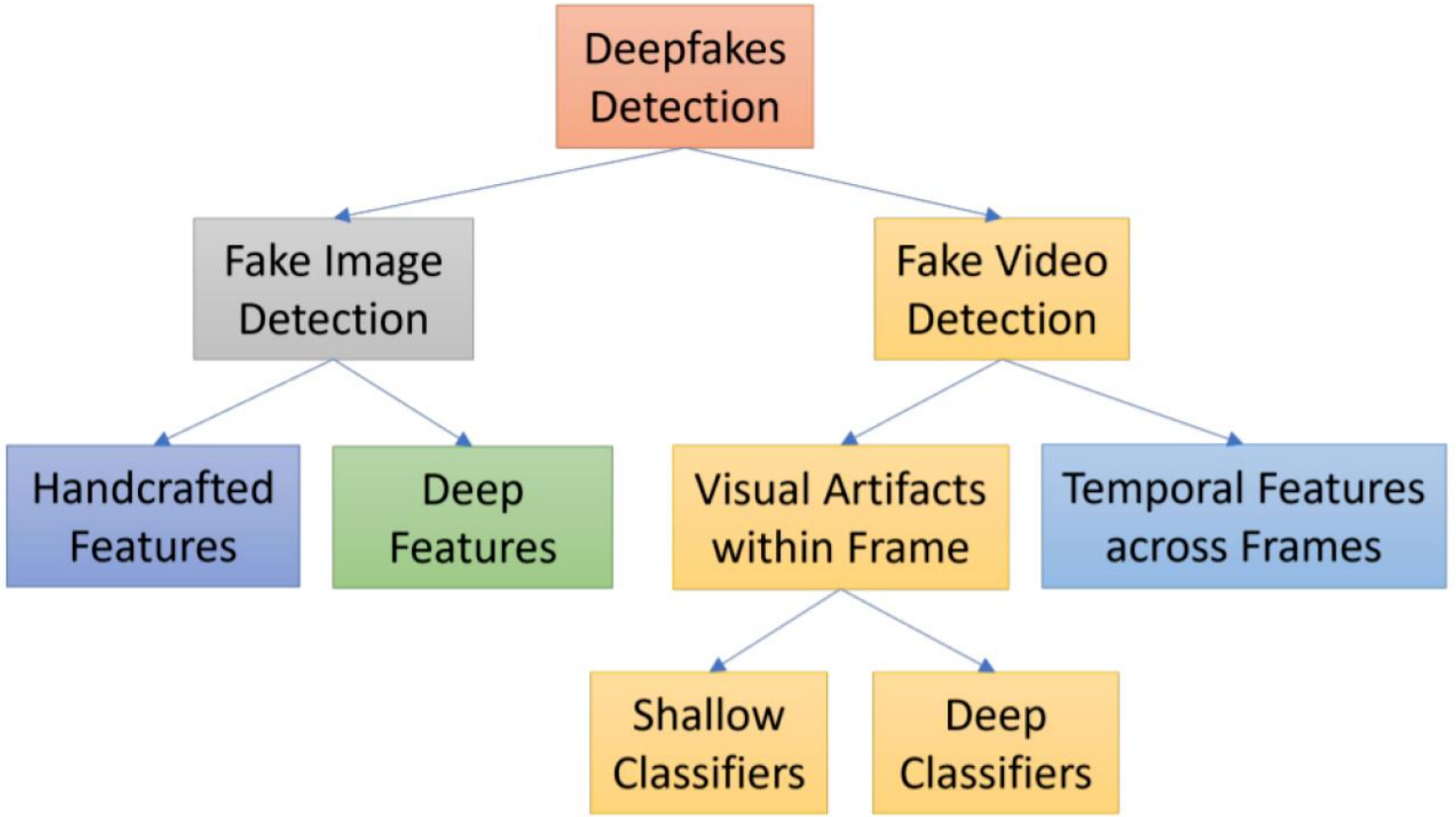
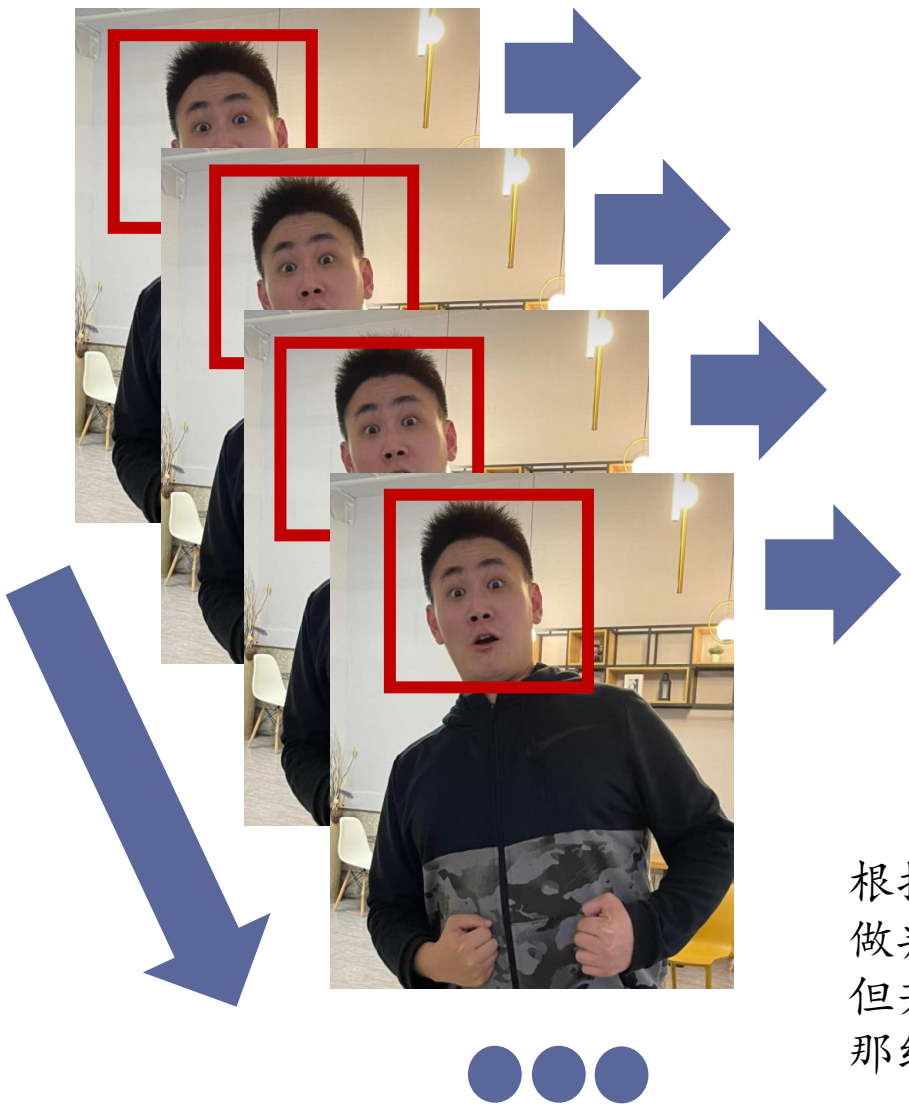


Fig. 6. Categories of reviewed papers relevant to deepfake detection methods where we divide papers into two major groups, i.e., fake image detection and face video detection.

根据图片与影片的深度伪造检测

根据将影片处理成一帧帧的图像每个图片的伪造去做判断，可以判断单帧的图像真假，最后给出机率。但若面对未知伪造方式，或者窜改人脸部分很少，那结果就不理想。

根据影片的时间先后，去判断整个影片是否是伪造，可以找出少数人脸窜改的结果。但面对压缩过后的影像，或者是单帧的真假，就有可能不理想。

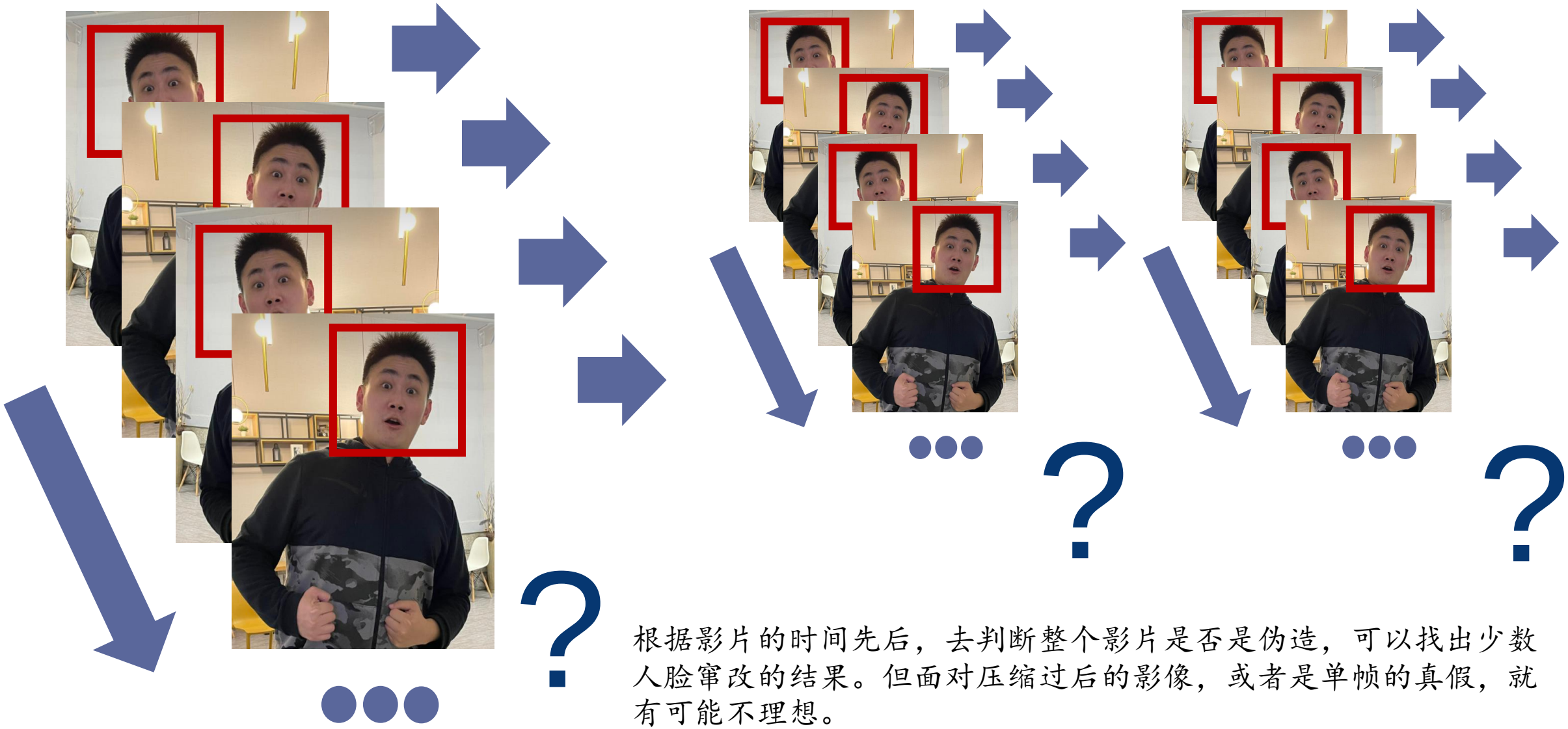


真

假

?? %

根据将影片处理成一帧帧的图像每个图片的伪造去做判断，可以判断单帧的图像真假，最后给出机率。但若面对未知伪造方式，或者窜改人脸部分很少，那结果就不理想。



根据影片的时间先后，去判断整个影片是否是伪造，可以找出少数人脸窜改的结果。但面对压缩过后的影像，或者是单帧的真假，就有可能不理想。

Zeinali H et al.
1907.12908

Table 3: *Physical access detailed results based on min-tDCF for different conditions. The first section shows the baseline results and the second section shows the primary and single best results of the best-performing systems, both from team T28.*

	Development set									Evaluation set								
System	**AA**	**AB**	**AC**	**BA**	**BB**	**BC**	**CA**	**CB**	**CC**	**AA**	**AB**	**AC**	**BA**	**BB**	**BC**	**CA**	**CB**	**CC**
CQCC-GMM	0.4928	0.0539	0.0213	0.3999	0.0360	0.0197	0.4338	0.0414	0.0149	0.4975	0.1751	0.0529	0.4658	0.1483	0.0433	0.5025	0.1360	0.0461
T28 Primary	0.0132	0.0030	0.0009	0.0073	0.0017	0.0009	0.0065	0.0023	0.0008	0.0190	0.0079	0.0034	0.0113	0.0083	0.0022	0.0127	0.0075	0.0024
T28 Single	0.0185	0.0044	0.0013	0.0146	0.0043	0.0014	0.0146	0.0081	0.0024	0.0251	0.0107	0.0055	0.0152	0.0114	0.0058	0.0183	0.0111	0.0063
Primary	0.0389	0.0062	0.0039	0.0243	0.0049	0.0048	0.0233	0.0073	0.0028	0.0776	0.0217	0.0091	0.0586	0.0223	0.0088	0.0557	0.0256	0.0110
Single best	0.0611	0.0046	0.0040	0.0404	0.0052	0.0053	0.0402	0.0085	0.0039	0.1061	0.0267	0.0117	0.0901	0.0277	0.0115	0.0843	0.0330	0.0128
Contrastive1	0.0523	0.0245	0.0151	0.0256	0.0156	0.0130	0.0280	0.0229	0.0135	0.0695	0.0383	0.0148	0.0493	0.0383	0.0141	0.0437	0.0394	0.0192
Contrastive2	0.0726	0.0323	0.0170	0.0562	0.0283	0.0153	0.0633	0.0353	0.0167	0.0969	0.0547	0.0187	0.0843	0.0519	0.0193	0.0842	0.0532	0.0229

Table 4: *Logical access detailed results based on min-tDCF for different conditions. The first section shows the baseline results and the second section shows the primary system results of the best performing team (T05) as well as the overall best single system results (team T45). The bold numbers show conditions where our single system performs better or the same as the best single system.*

	Development set						Evaluation set												
System	**A01**	**A02**	**A03**	**A04**	**A05**	**A06**	**A07**	**A08**	**A09**	**A10**	**A11**	**A12**	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**A19**
CQCC-GMM	0.0000	0.0000	0.0020	0.0000	0.0261	0.0011	0.0000	0.0007	0.0060	0.4149	0.0020	0.1160	0.6729	0.2629	0.0344	0.0000	0.9820	0.2818	0.0014
T05 Primary	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0009	0.0014	0.0000	0.0077	0.0055	0.0045	0.0028	0.0035	0.0050	0.0015	0.0341	0.0276	0.0020
T45 Single	0.0027	0.0000	0.0000	0.0036	0.0068	0.0085	0.0034	0.0308	0.0000	0.0130	0.0017	0.0058	0.0034	0.0042	0.0065	0.0071	0.9833	0.1171	0.0895
Primary	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0029	0.5672	0.0425	0.0425	0.1098	0.0005	0.5525	0.0000	0.3775	0.6473	0.0000
Single best	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0002**	**0.0004**	0.1393	0.9423	0.0426	1.0000	0.3693	**0.0000**	1.0000	**0.0004**	**0.4764**	0.6731	**0.0000**
Contrastive1	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000	0.0654	0.2004	0.1663	0.5031	0.0002	0.9297	0.8583	0.0000	0.0002	0.0007	0.0263	0.5749	0.3217
Contrastive2	0.0000	0.0000	0.0000	0.0010	0.0000	0.0000	0.0017	0.0026	0.1505	0.9992	0.0253	1.0000	0.4737	0.0000	1.0000	0.0022	0.4131	0.9420	0.0009

Table 5: *Physical access results of different submissions*

	Development set		Evaluation set	
System	**EER[%]**	**min-tDCF**	**EER [%]**	**min-tDCF**
CQCC-GMM	9.87	0.1953	11.04	0.2454
Primary	0.66	0.0170	1.51	0.0372
Single best	1.02	0.0254	2.11	0.0527
Contrastive1	1.07	0.0253	1.49	0.0401
Contrastive2	1.59	0.0401	2.31	0.0591

Table 6: *Logical access results of different submissions*

	Development set		Evaluation set	
System	**EER[%]**	**min-tDCF**	**EER [%]**	**min-tDCF**
CQCC-GMM	0.43	0.0123	9.57	0.2366
Primary	0.00	0.0000	8.01	0.2080
Single best	0.00	0.0000	20.11	0.3563
Contrastive1	0.00	0.0000	10.52	0.2790
Contrastive2	0.03	0.0003	22.99	0.3811

4. 深度伪造对抗性

根据发现对人脸进行加躁、修改、变形等干扰行为，可以欺骗机器。

Song Q et al. 1811.12026

在訓練時對訓練的資料進行類似 JPEG 壓縮、模糊等操作能提升泛化性能。

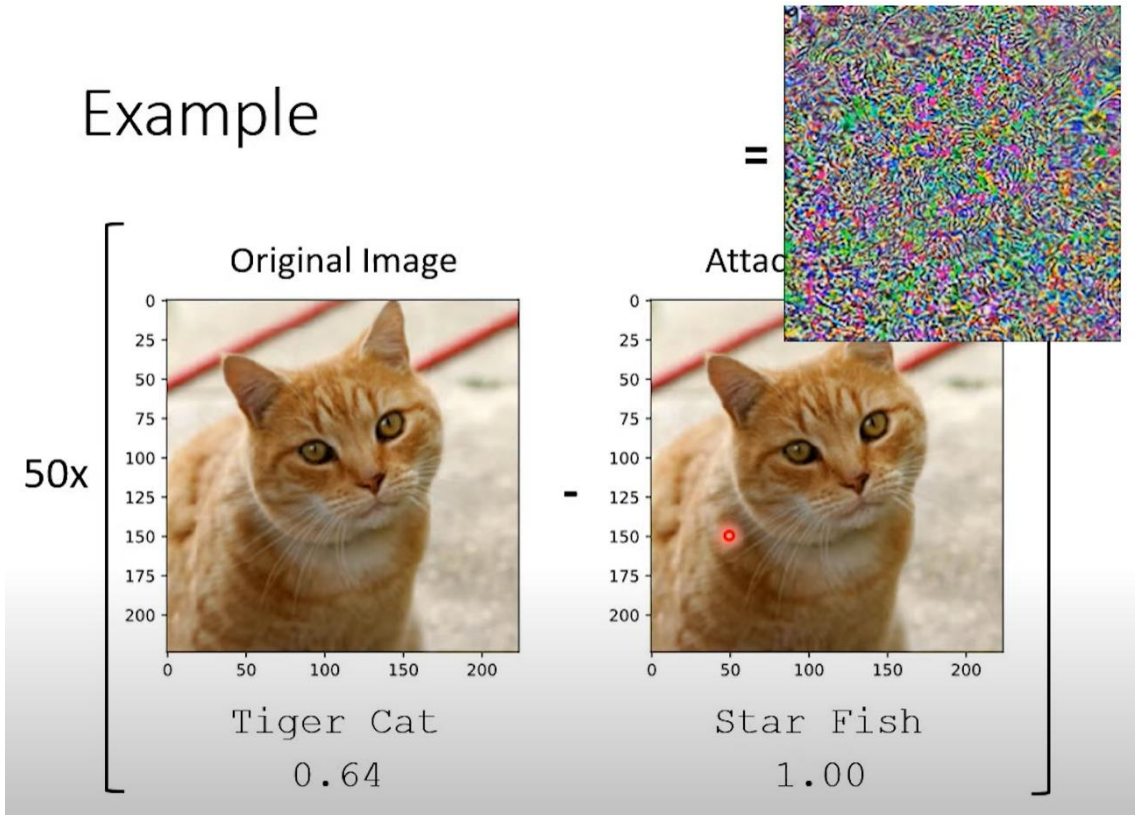
Marra F et al.

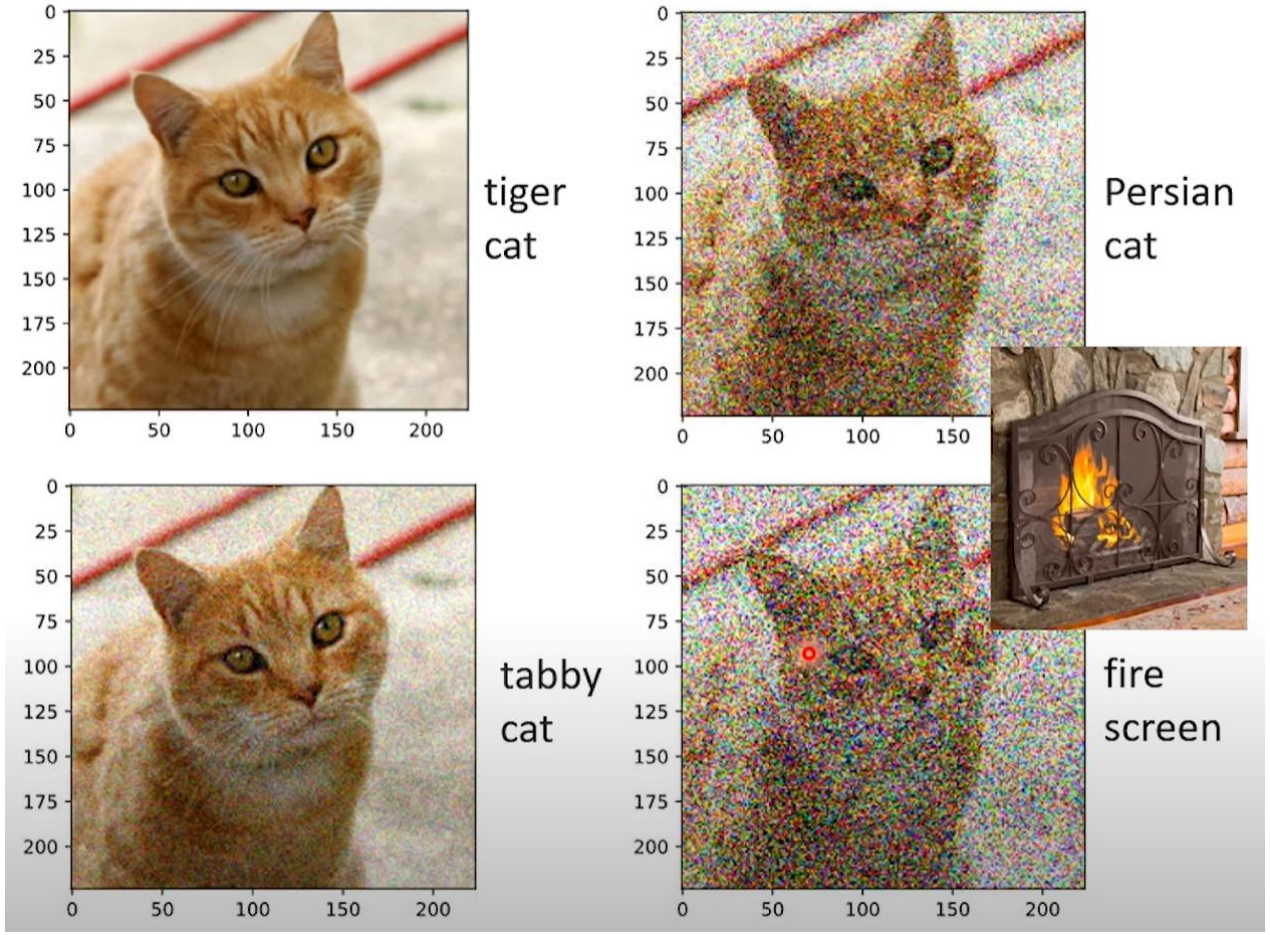
Tip：在若想使自己对照片进行加躁，使攻击者的生成无法顺利生成，又或者是对自己伪造后的影像进行干扰使他人的检测无法顺利执行!??

4. 深度伪造对抗性

北京大学 PEKING UNIVERSITY

[李宏毅教授演講] 今天的人工智慧 其實沒有你想的那麼厲害





https://www.youtube.com/watch?v=YjfzJdimPDA

4. 深度伪造对抗性

Huang, C. Y. et al.
2005.08781

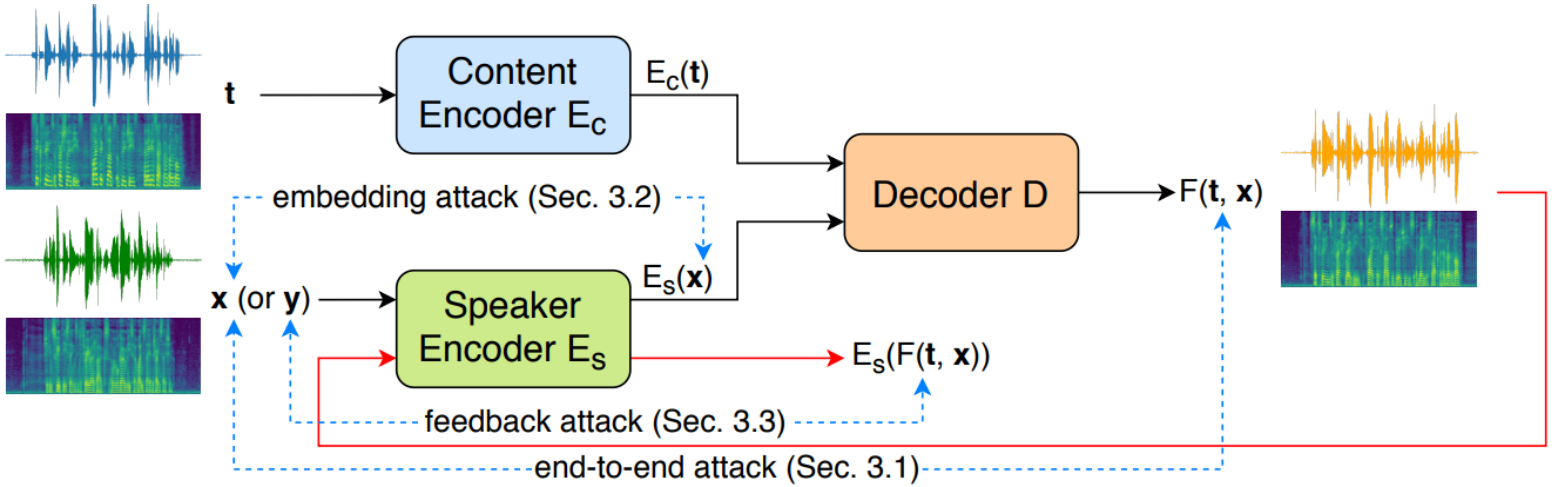


Fig. 1: The encoder-decoder based voice conversion model and the three proposed approaches. Perturbations are updated on the utterances providing speaker characteristics, as the blue dashed lines indicate.

够通过语音转换技术来防止一个人的语音被不当使用，报告对语音转换执行对抗性攻击的研究

Huang, C. Y., Lin, Y. Y., Lee, H. Y., & Lee, L. S. (2021, January). Defending your voice: Adversarial attack on voice conversion. In 2021 IEEE Spoken Language Technology Workshop (SLT) (pp. 552-559). IEEE.

5. 未来与近期研究

M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection

Video Transformer for Deepfake Detection with Incremental Learning

Wang, J et al.
2104.09770

Khan, S et al.
10.1145/3474085.3475332

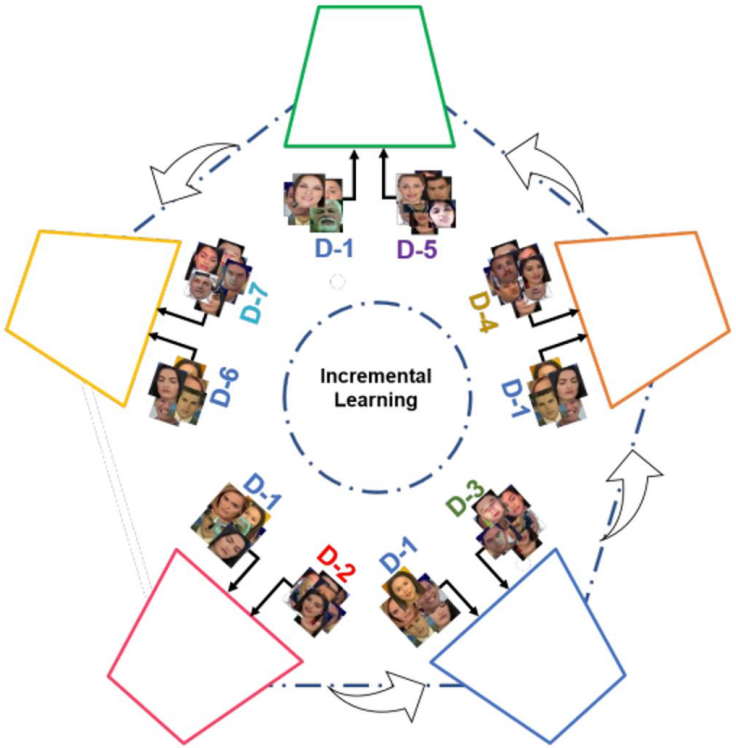


Figure 2: Illustration of the proposed incremental learning strategy. D1 represents the real data used to train the models. Whereas, D2 comprises of FaceSwap and Deepfakes datasets. D3 represents the Face2Face dataset and D4 represents Neural Textures dataset. D5 and D6 represents DFDC dataset and D7 represents DeepFake Detection (DFD) dataset.

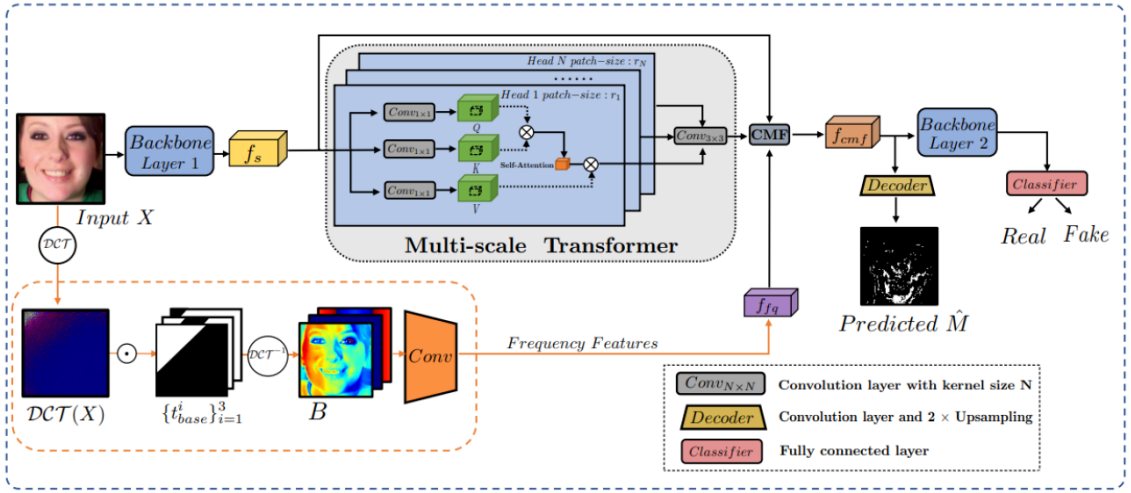


Figure 2: Overview of the proposed M2TR. The input is a suspicious face image (H x W x C), and the output includes both a forgery detection result and a predicted mask (H x W x 1), which locates the forgery regions.

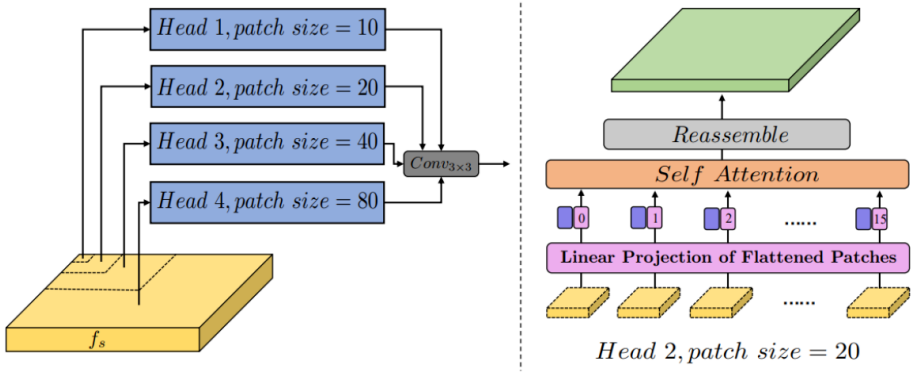


Figure 3: Illustration of the Multi-scale Transformer.

Reference

Li XR et al.
2104.09770

Thanh Thi Nguyen et al.
1909.11573

Yisroel Mirsky et al.
2004.11138

END