Data Tampering & Forgery

马兴军，复旦大学 计算机学院

Recap: week 10

- Membership Inference Attack
- Differential Privacy

This Week

- General Tampering
- Deepfake
- Deepfake Videos
- Detection

Significant Progress in Computer Vision

DaLL·E2

OpenAI

Text2Image,
Image Editing…

Imagen

Google

Text2Image,
Text2Vedio

Stable Diffusion

Stability AI

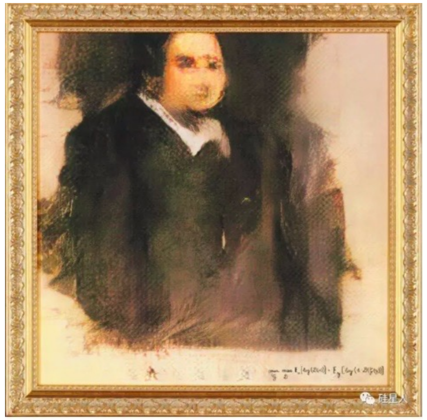
Text2Image,
Image Editing…

Significant Progress in Computer Vision

The resolution and fidelity of generated face images are constantly improving.

2019

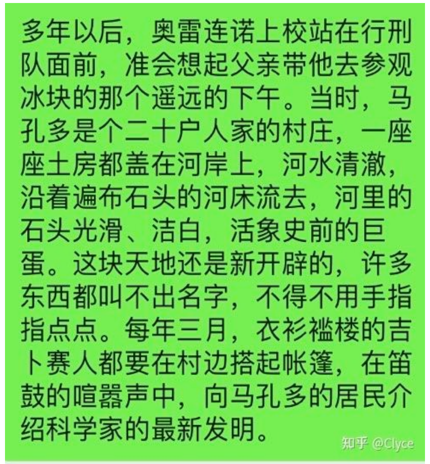
2021



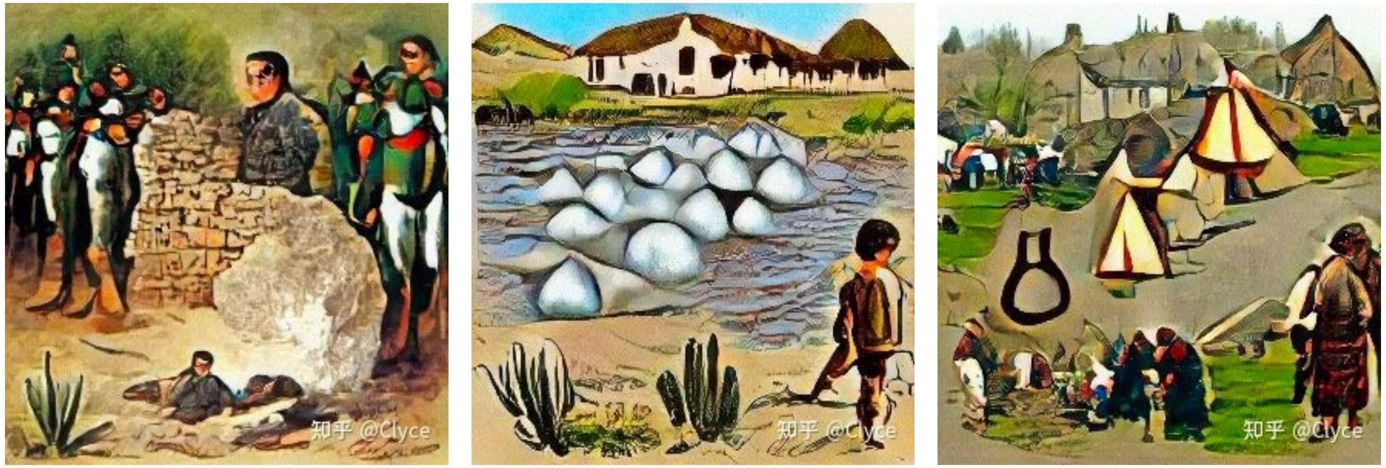
An AI-generated portrait sold for $432,000 at the Christie's (2018)

AI artwork won first prize in art competition.(2022)

Significant Progress in Computer Vision


多年以后，奥雷连诺上校站在行刑队面前，准会想起父亲带他去参观冰块的那个遥远的下午。当时，马孔多是个二十户人家的村庄，一座座土房都盖在河岸上，河水清澈，沿着遍布石头的河床流去，河里的石头光滑、洁白，活象史前的巨蛋。这块天地还是新开辟的，许多东西都叫不出名字，不得不用手指指点点。每年三月，衣衫褴褛的吉卜赛人都要在村边搭起帐篷，在笛鼓的喧嚣声中，向马孔多的居民介绍科学家的最新发明。

知乎 @Clyce



Generate an image using the first paragraph of "One Hundred Years of Solitude" (2021)

DaLL•E2 (2022)

Generate an image based on text: “I have always wanted to be a cool panda riding a skateboard in Santa Monica.”

“A photo of a sitting dog”

Imagic (2022)

Edit images with text.

Data Tampering and Forgery

- ***Definition*** **：** Tamper images and videos with variety of techniques, such as deepfakes.
- According to the content and type of the tampered data: *general tampering* & *face forgery*.

A fake image about Bush Jr. election

This Week

- **General Tampering**
- Deepfake
- Deepfake Videos
- Detection

General Tampering

- ***Definition:*** tamper the original image by adjusting the spatial position of objects, replacing the original content with forged content (style modification, texture transformation, image restoration…)
- ***Taxonomy***
  - Context-based
    - tamper **foreground objects**
    - tamper image **background**
  - Conditioned
    - Text-guided image tampering

General Tampering

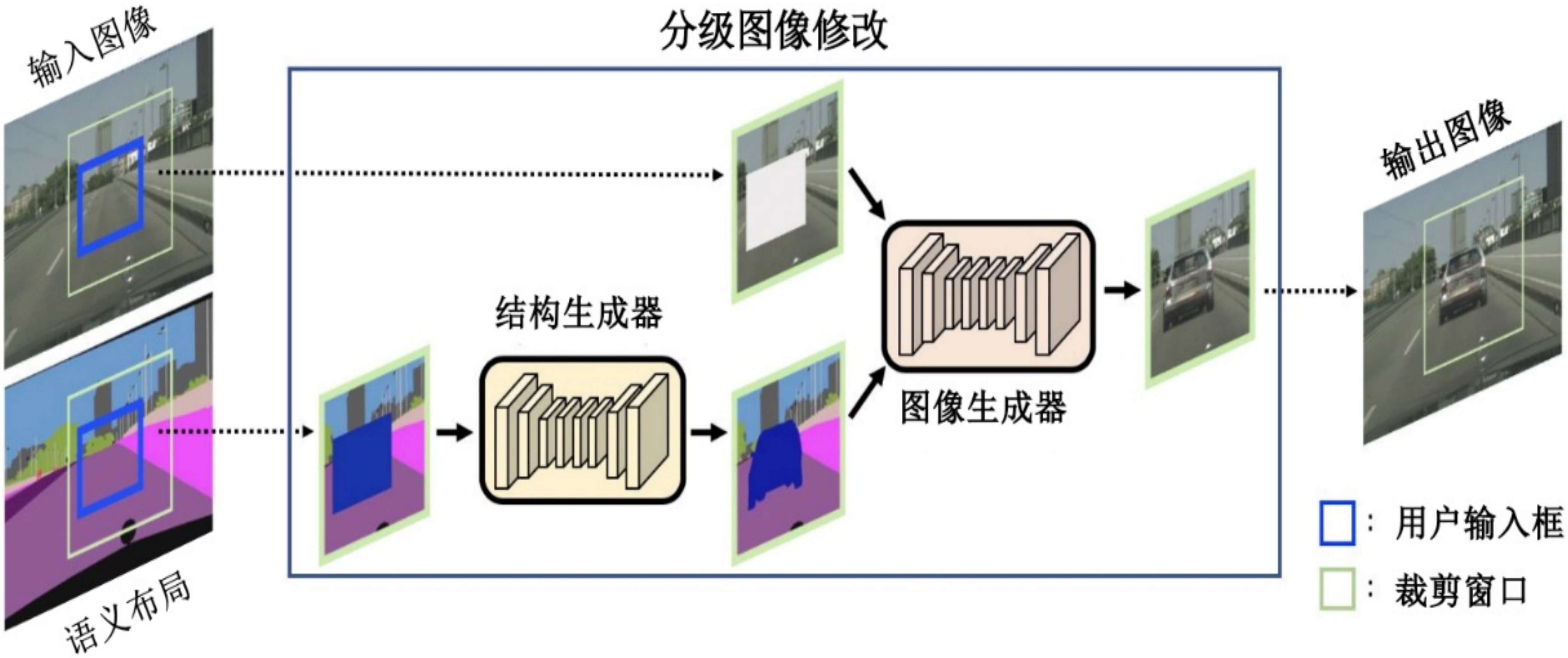
?

Core Problem **：** how to decouple different elements in an image?
(Foreground & Background, Texture & Structure, …)

- Model different elements in the image: the shape of objects, the interaction between objects and their relative positions, …

Foreground Tampering

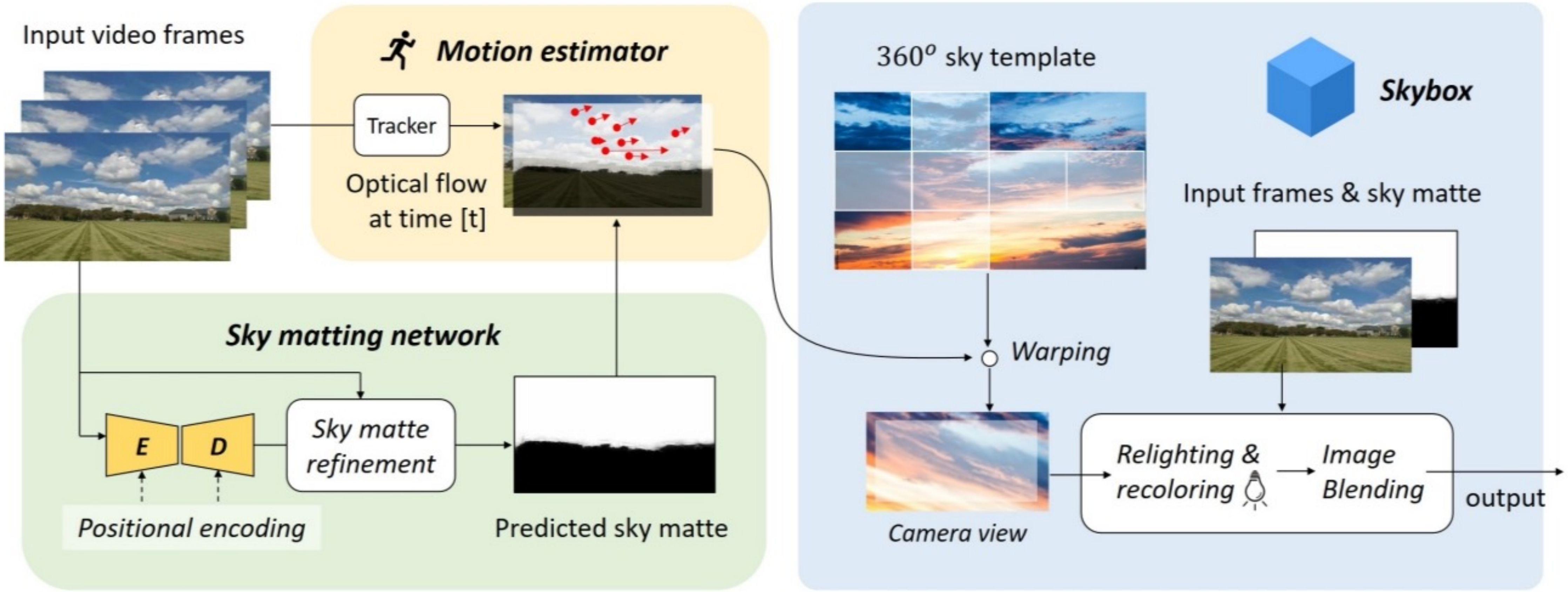
- Construct object-level semantic segmentation maps



Hong, S., Yan, X., Huang, T. S., & Lee, H. (2018). Learning hierarchical semantic image manipulation through structured representations. *Advances in Neural Information Processing Systems*, *31*.

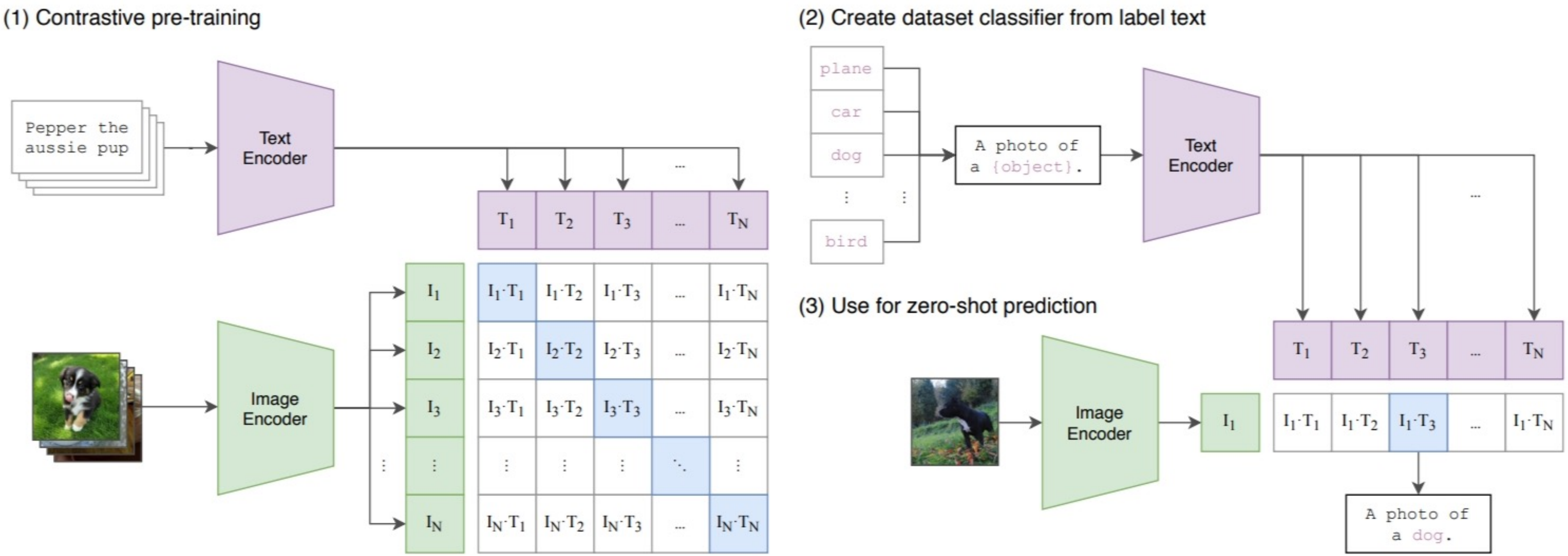
Background Tampering

- the background can be viewed as a larger object



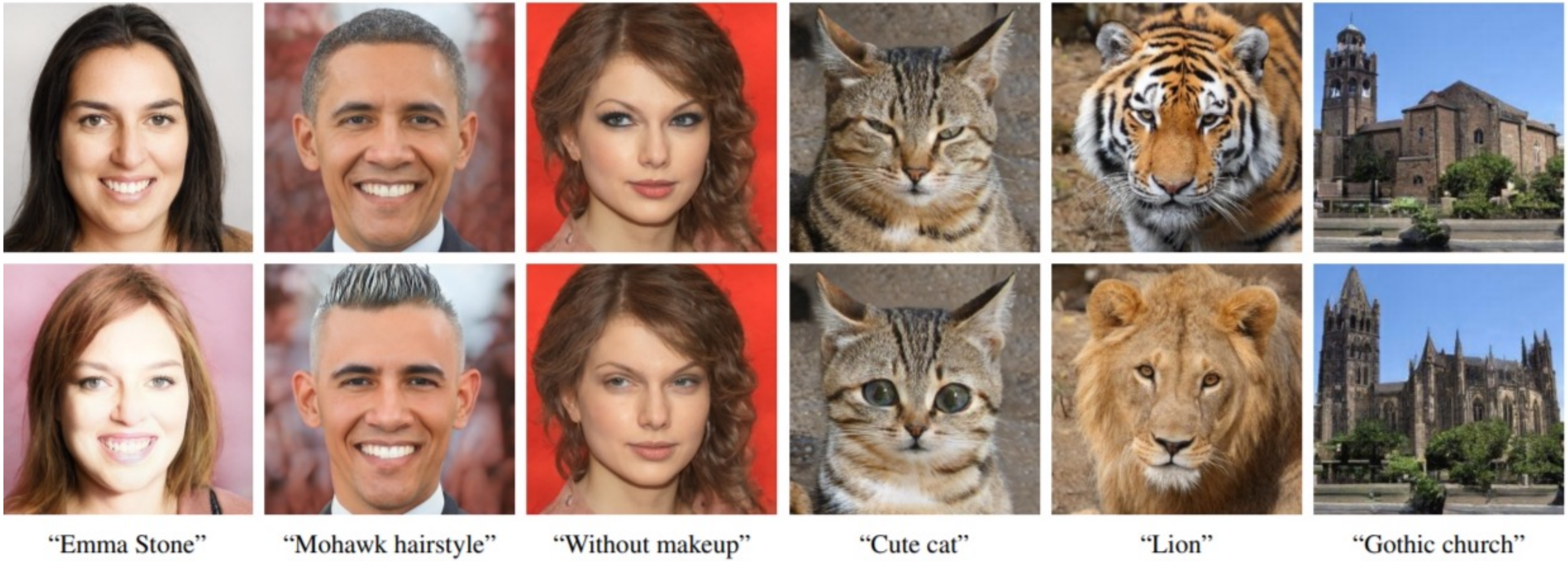
Zou, Z., Zhao, R., Shi, T., Qiu, S., & Shi, Z. (2022). Castle in the sky: dynamic sky replacement and harmonization in videos. *IEEE Transactions on Image Processing*.

Text-guided Tampering | CLIP



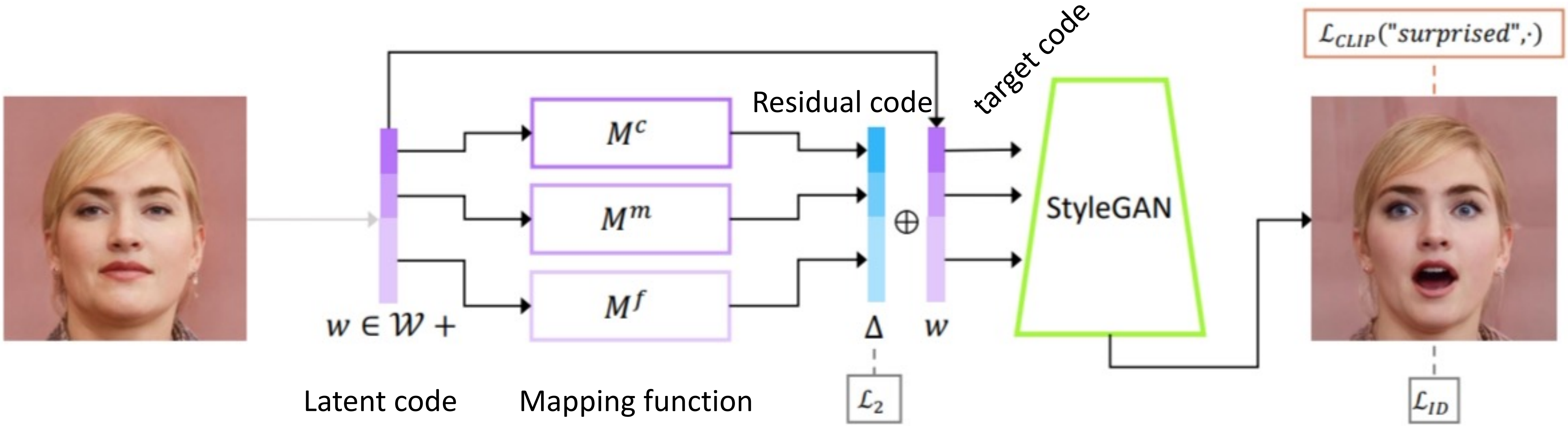
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … others. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning* (pp. 8748–8763).

Text-guided Tampering | CLIP + StyleGAN



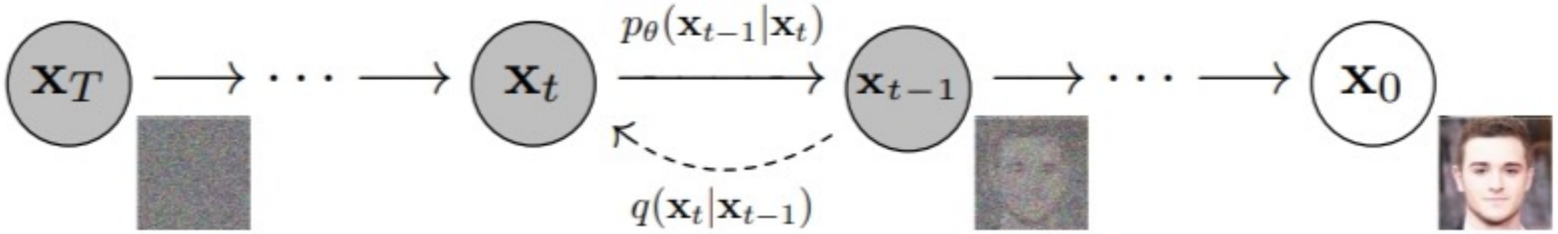
Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: text-driven manipulation of stylegan imagery. *IEEE/CVF International Conference on Computer Vision* (pp. 2085–2094).

Text-guided Tampering | StyleGAN

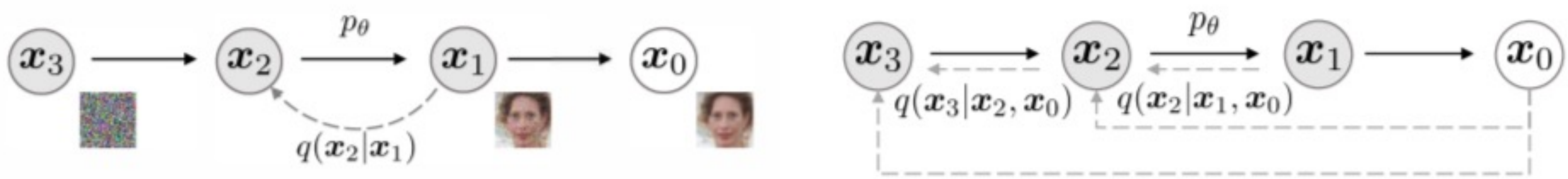


Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: text-driven manipulation of stylegan imagery. *IEEE/CVF International Conference on Computer Vision* (pp. 2085–2094).

Text-guided Tampering | Diffusion



The directed graphical model of DDPM



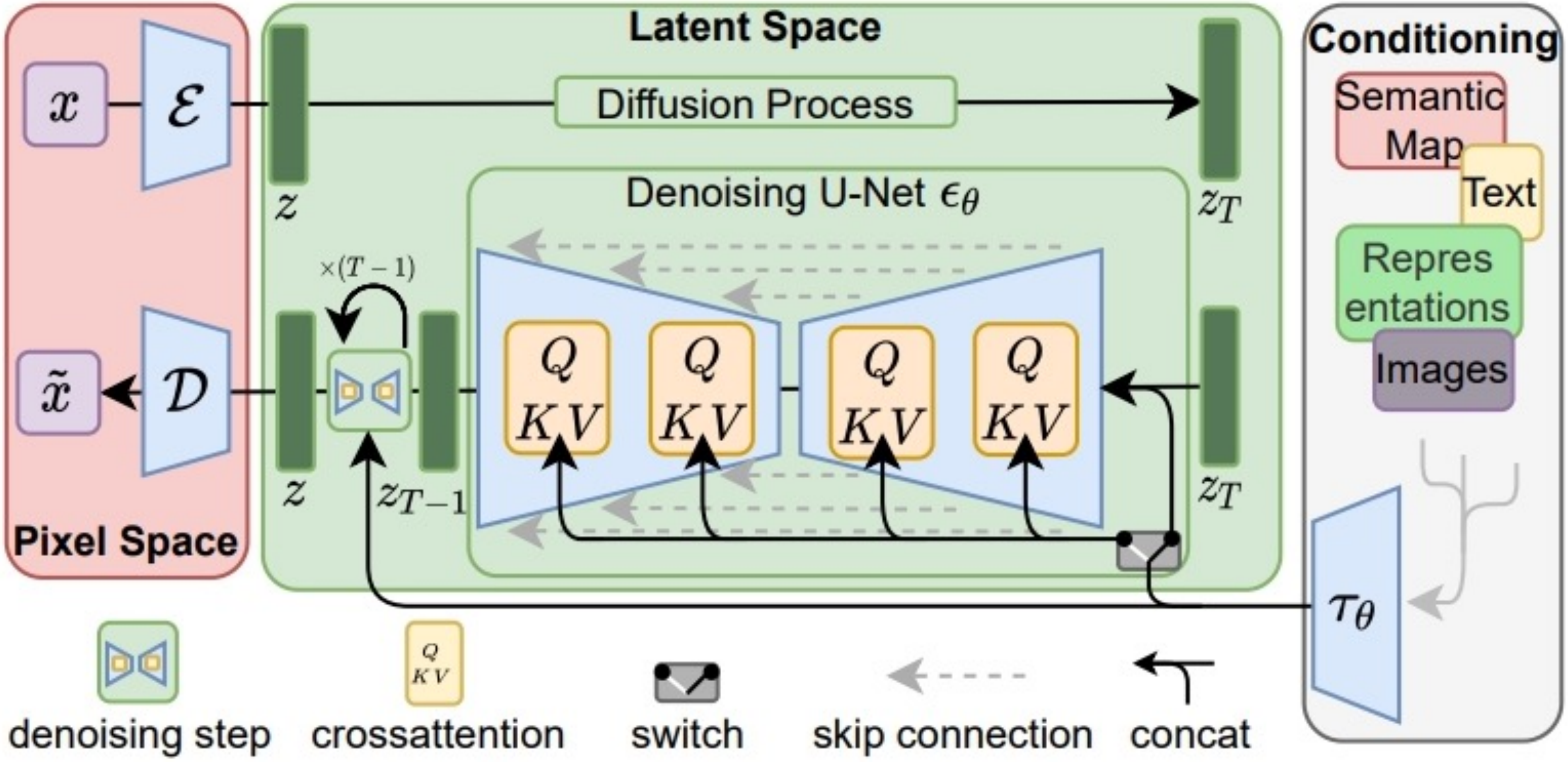
Graphical models for diffusion (left) and non-Markovian (right) inference models

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.

Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. International Conference on Learning Representations.

Text-guided Tampering | CLIP + Diffusion

- Stable Diffusion



Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.

This Week

- ☐ General Tampering
- ☐ **Deepfake**
- ☐ Deepfake Videos
- ☐ Detection

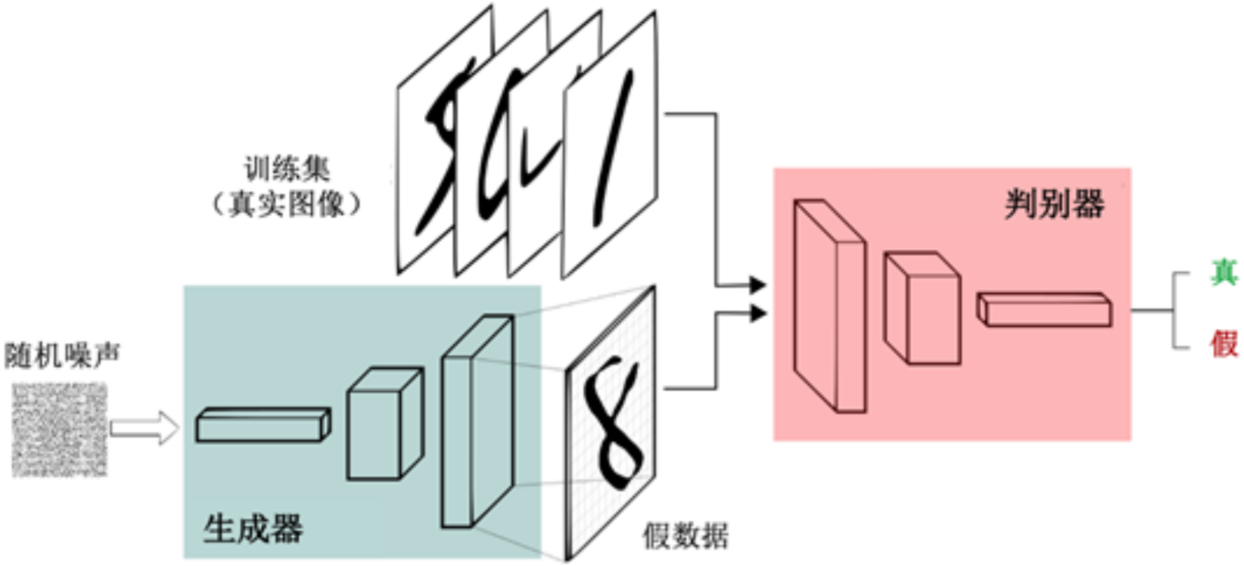
Deepfake

Deep learning + fake

- ***Definition:*** believable media generated by a deep neural network
- ***Form:*** *generation* & *manipulation* of human imagery

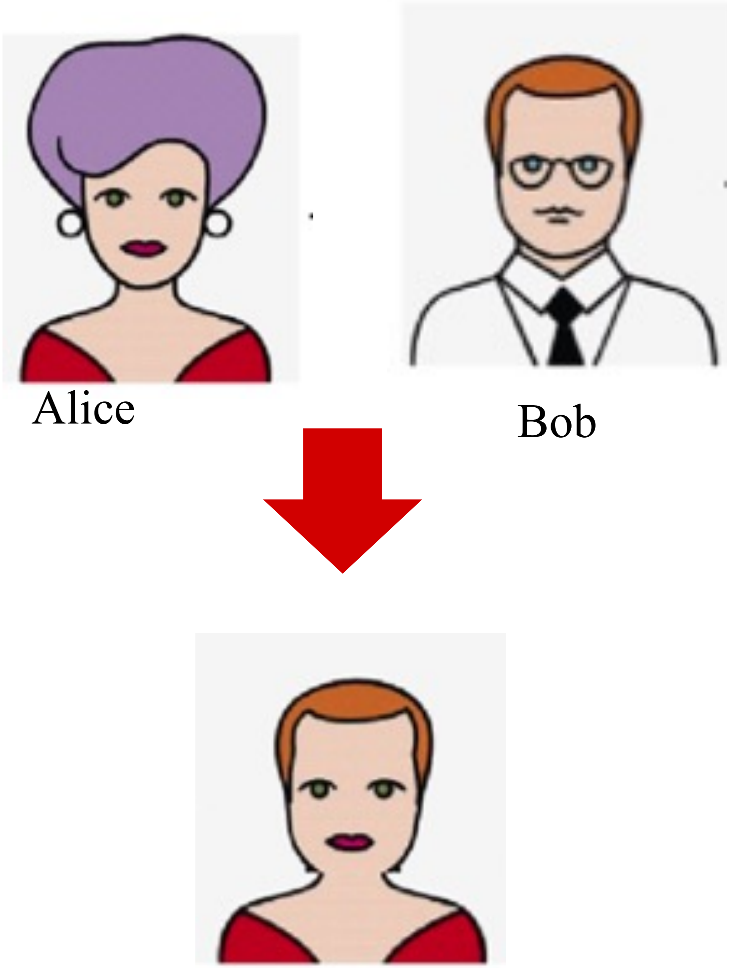
GANs (Generative Adversarial Networks)

- Derives from the “zero-sum game” in game theory.
- Learn the distribution of data through a Generator and a Discriminator



$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

Face Forgery



Alice

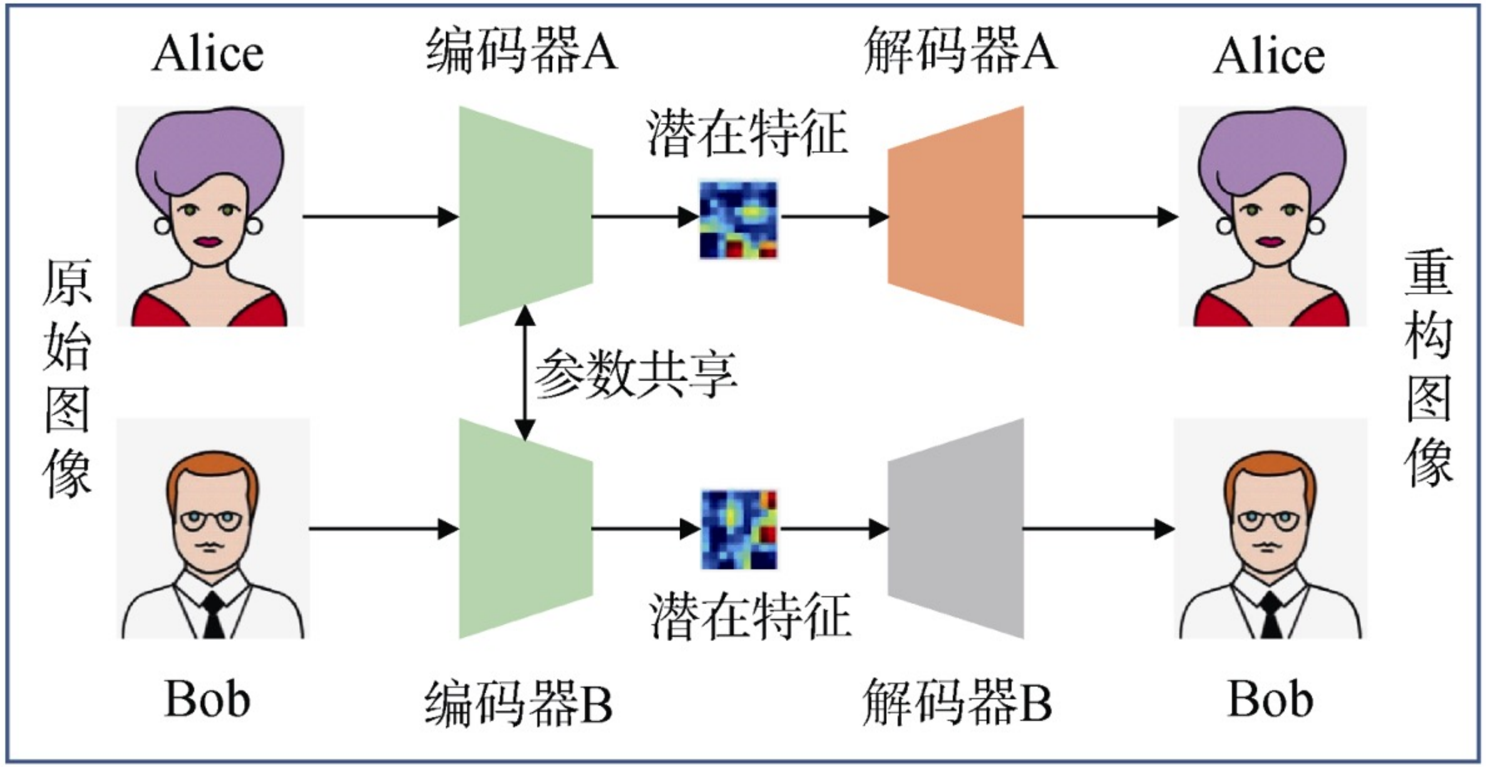
Bob

Alice’s body with Bob’s face

- **Data collection**
- Model training
- Deepfake face forgery

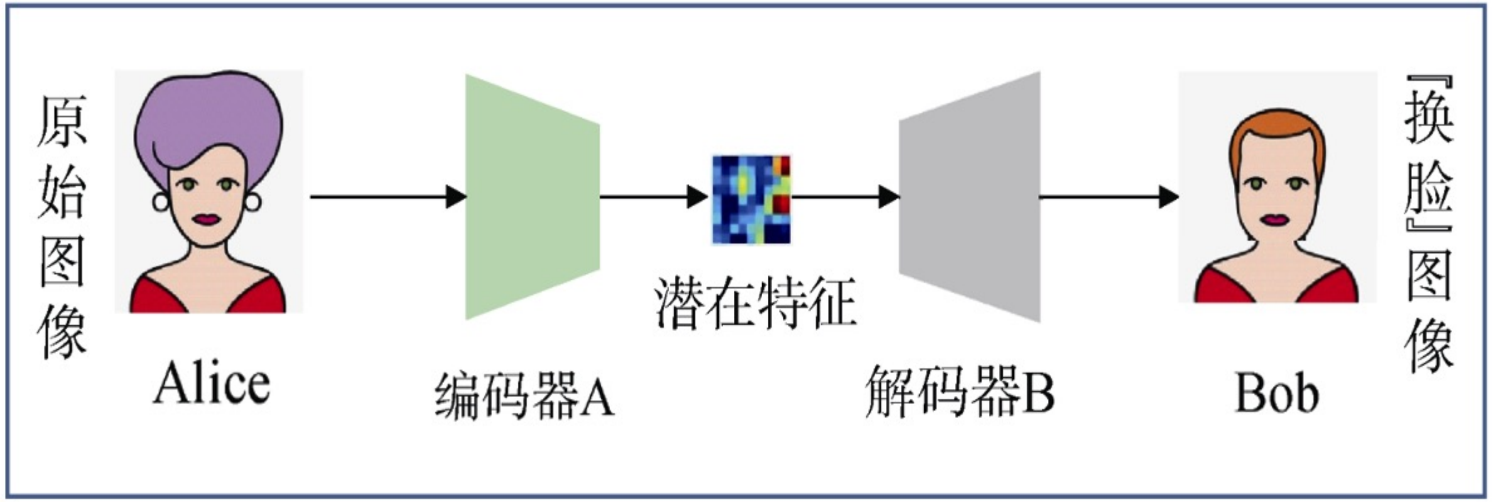
Face Forgery

- Data collection
- **Model training**
- Deepfake face forgery



Face Forgery

- Data collection
- Model training
- **Deepfake face forgery**



Face Forgery

- Reenactment (人脸重演)
- Replacement (人脸互换)
- ~~Editing (人脸编辑)~~
- ~~Synthesis (人脸合成)~~

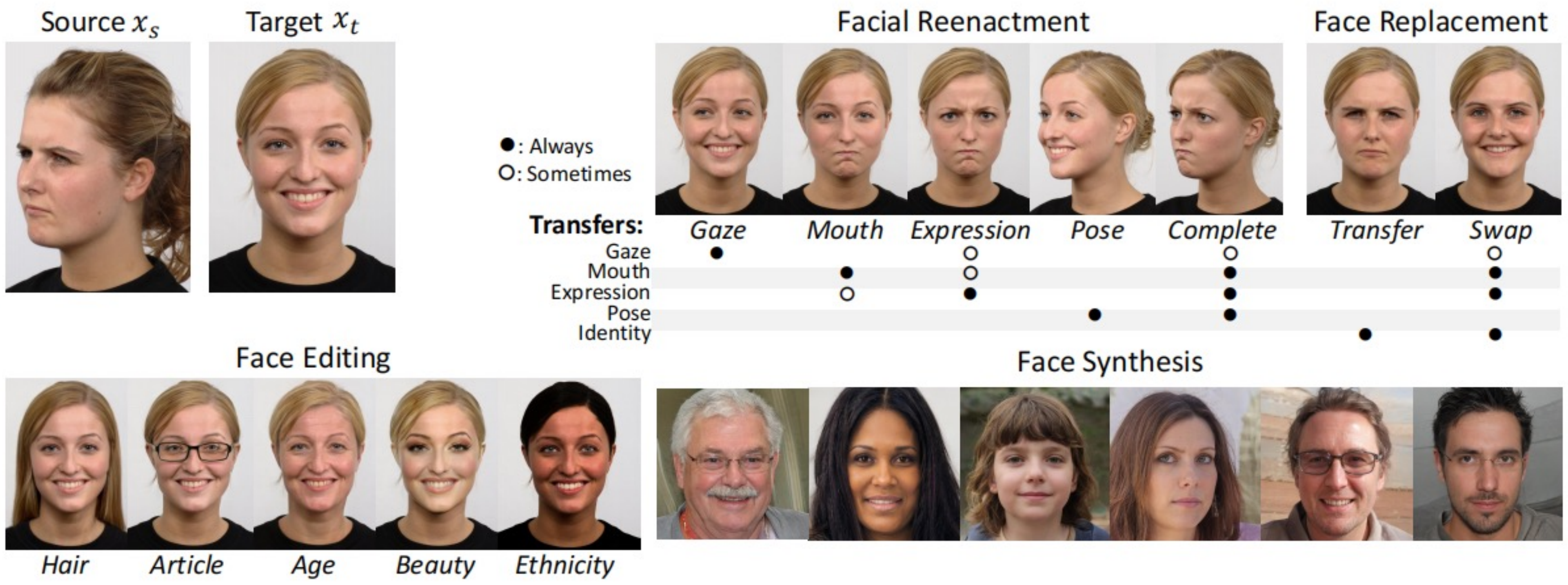


Fig. 3. Examples of reenactment, replacement, editing, and synthesis deepfakes of the human face.

s : the source identities

t : the target identities

x_t: images of the target identities

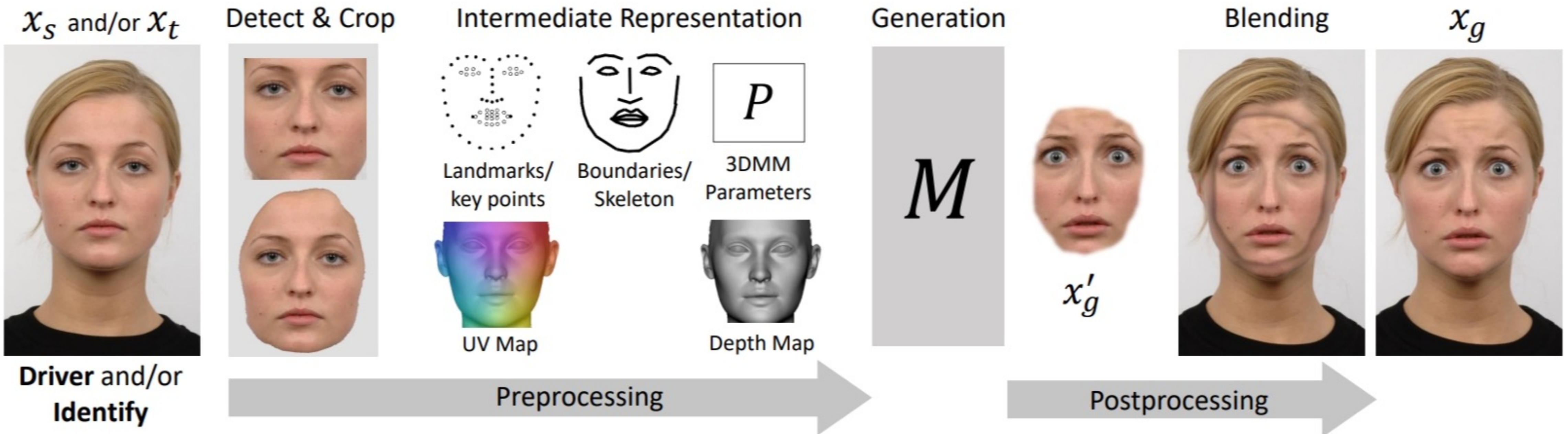
x_s: images of the source identities

x_g: the deepfake generated from s and t

Mirsky Y, Lee W. The creation and detection of deepfakes: A survey[J]. ACM Computing Surveys (CSUR), 2021, 54(1): 1-41.

Face Forgery

- STEPS：
  1. Detects and crops the face
  2. Extracts intermediate representations
  3. Generates a new face based on some driving signal
  4. Blends the generated face back into the target frame

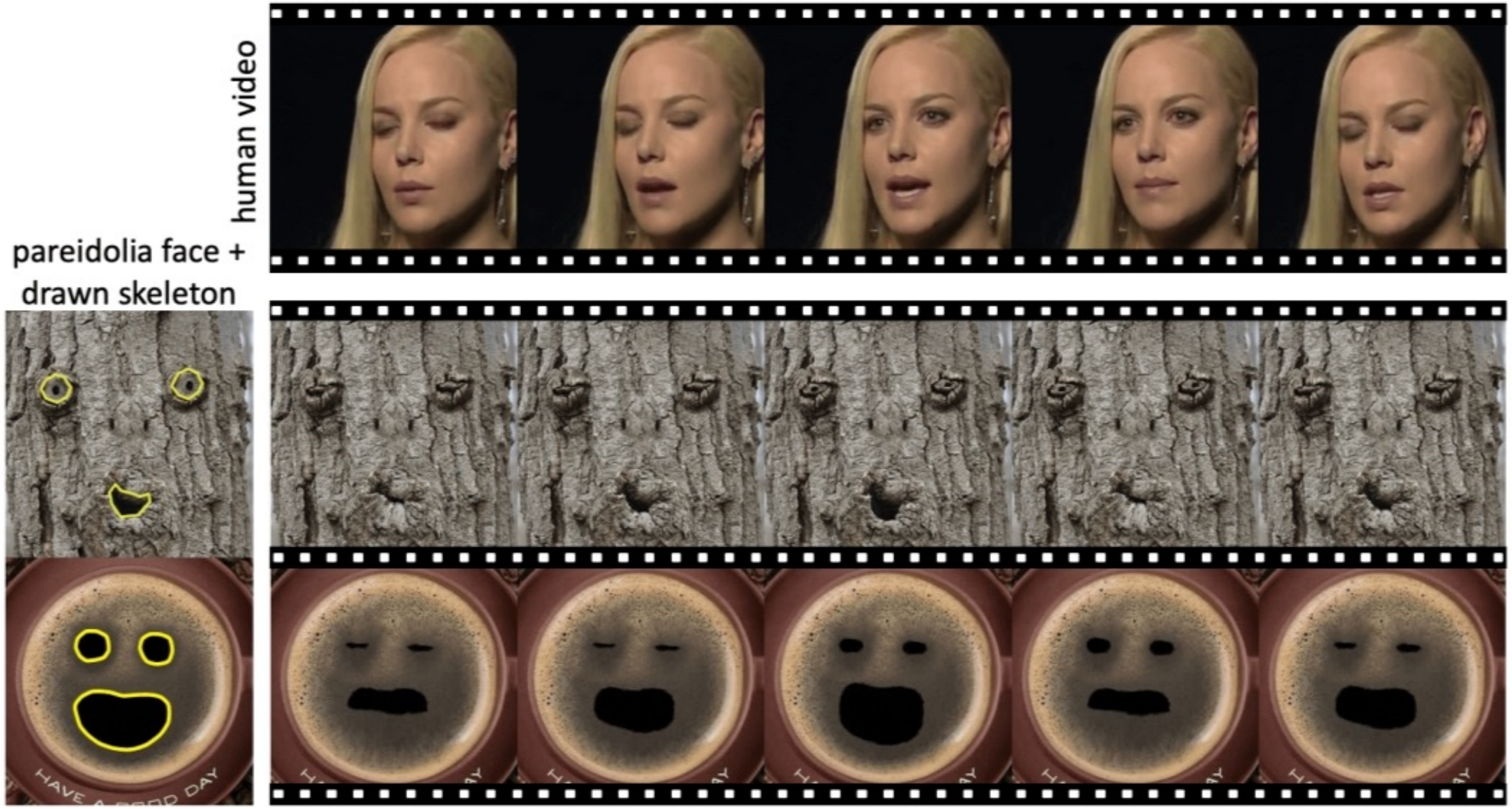


Face Reenactment

- STEPS in general:
  1. face tracking (面部追踪)
  2. face matching (面部匹配)
  3. face transfer (面部迁移)

Pareidolia Face Reenactment

(a) Examples ofparediolia faces



(b) Parediolia face reenactment

Song, L., Wu, W., Fu, C., Qian, C., Loy, C. C., & He, R. (2021). Everything's talkin': pareidolia face reenactment. IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Pareidolia Face Reenactment

- Challenges

The target faces are not human faces

1 Shape variance

e.g. square mouth

2 Texture variance

e.g. wood texture

Song, L., Wu, W., Fu, C., Qian, C., Loy, C. C., & He, R. (2021). Everything's talkin': pareidolia face reenactment. IEEE/CVF Conference on Computer Vision and Pattern Recognition.

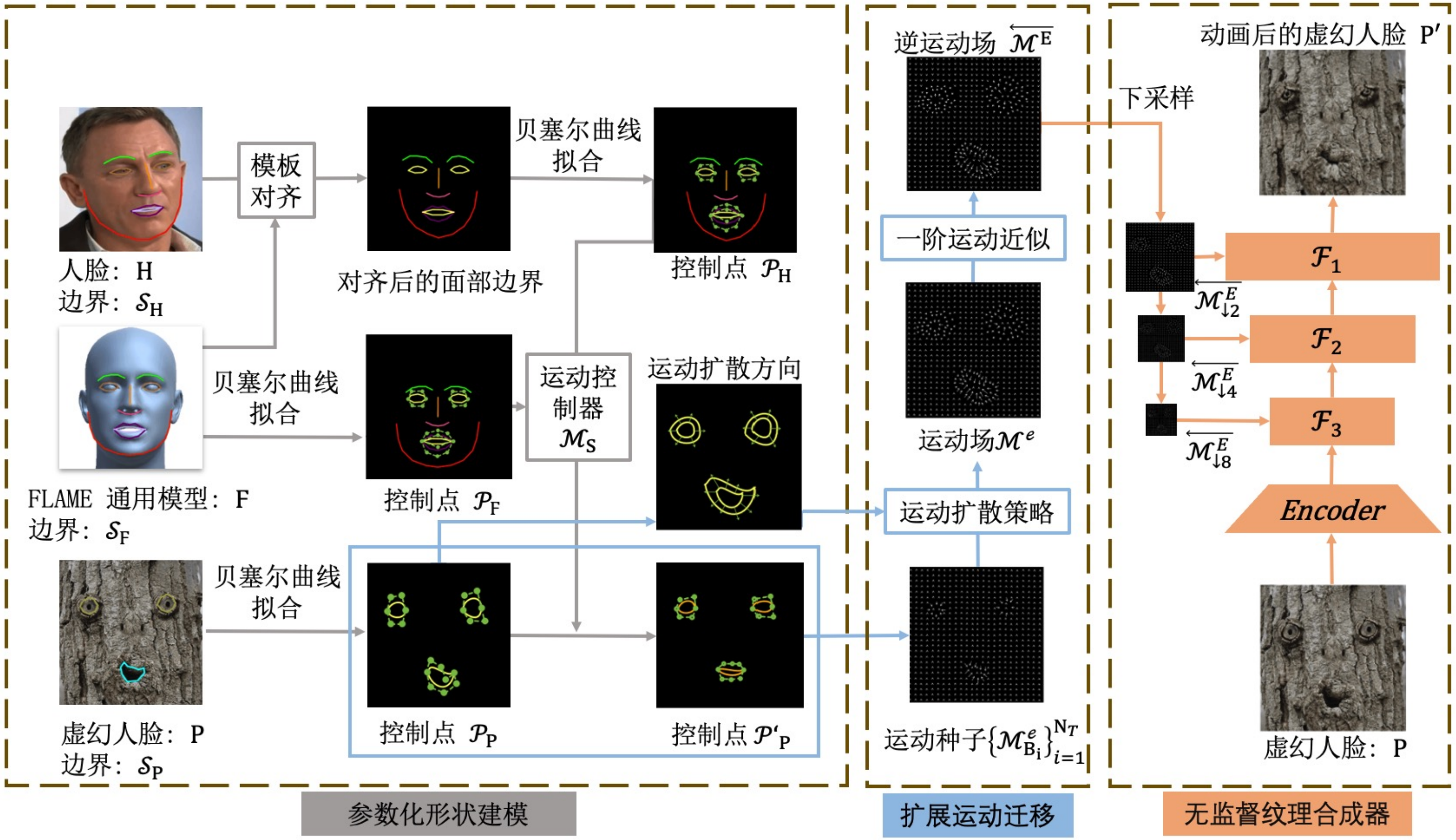
PURA

- Parametric Unsupervised Reenactment Algorithm
  - Parametric Shape Modeling (PSM，参数化形状建模)
  - Expansionary Motion Transfer (EMT，扩展运动迁移)
  - Unsupervised Texture Synthesizer (UTS，无监督纹理合成器)

Song, L., Wu, W., Fu, C., Qian, C., Loy, C. C., & He, R. (2021). Everything's talkin': pareidolia face reenactment. IEEE/CVF Conference on Computer Vision and Pattern Recognition.

PURA

- Parametric Unsupervised Reenactment Algorithm



Song, L., Wu, W., Fu, C., Qian, C., Loy, C. C., & He, R. (2021). Everything's talkin': pareidolia face reenactment. IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Face Replacement | Simswap

- High Fidelity Face Swapping

Source Target Result

❌lack the ability to ***generalize to arbitrary identity***
❌fail to ***preserve attributes*** like facial expression and gaze direction

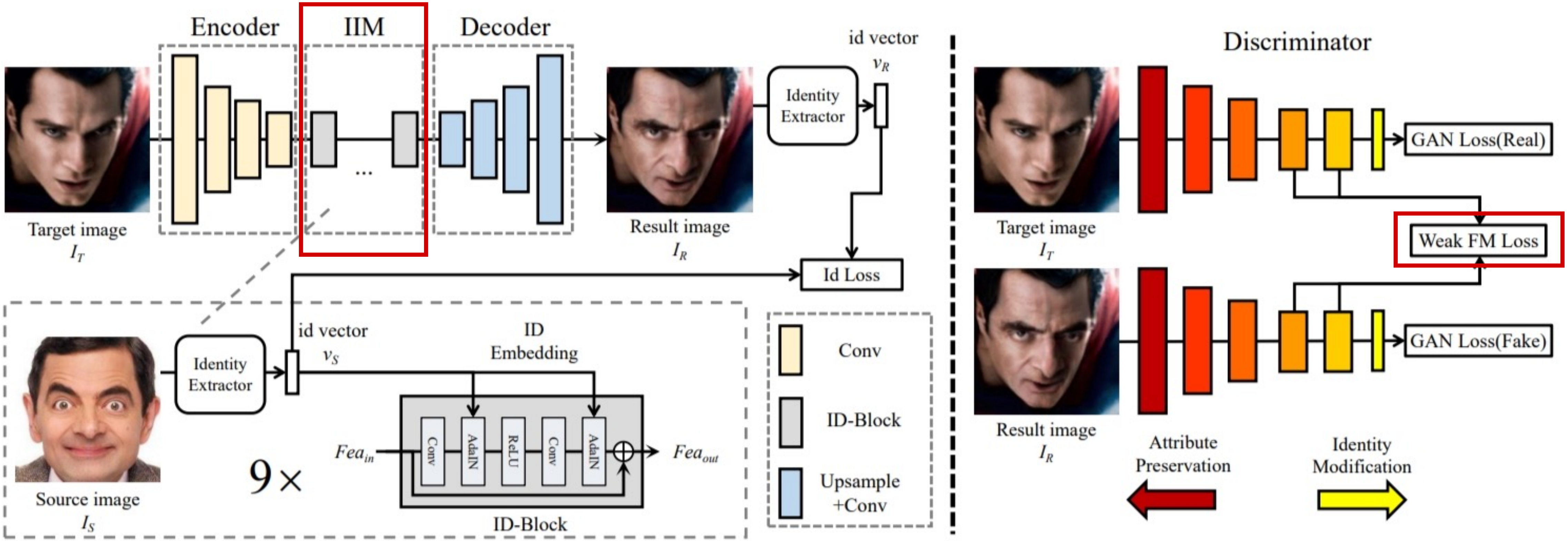
ID Injection Module (IIM)
(身份注入模块)

Weak Feature Matching Loss
(弱特征匹配损失)

Chen, R., Chen, X., Ni, B., & Ge, Y. (2020). Simswap: an efficient framework for high fidelity face swapping. ACM International Conference on Multimedia (pp. 2003–2011).

Face Replacement | Simswap

- High Fidelity Face Swapping



Chen, R., Chen, X., Ni, B., & Ge, Y. (2020). Simswap: an efficient framework for high fidelity face swapping. ACM International Conference on Multimedia (pp. 2003–2011).

Face Replacement | Simswap

- Identity Loss

$$L_{Id} = 1 - \frac{v_R \cdot v_S}{\|v_R\|_2 \|v_S\|_2}$$

- Weak Feature Matching Loss

$$L_{wFM}(D) = \sum_{i=m}^{M} \frac{1}{N_i} \|D^{(i)}(I_R) - D^{(i)}(I_T)\|_1$$

$$L_{wFM\_sum} = \sum_{i=1}^{2} L_{wFM}(D_i)$$

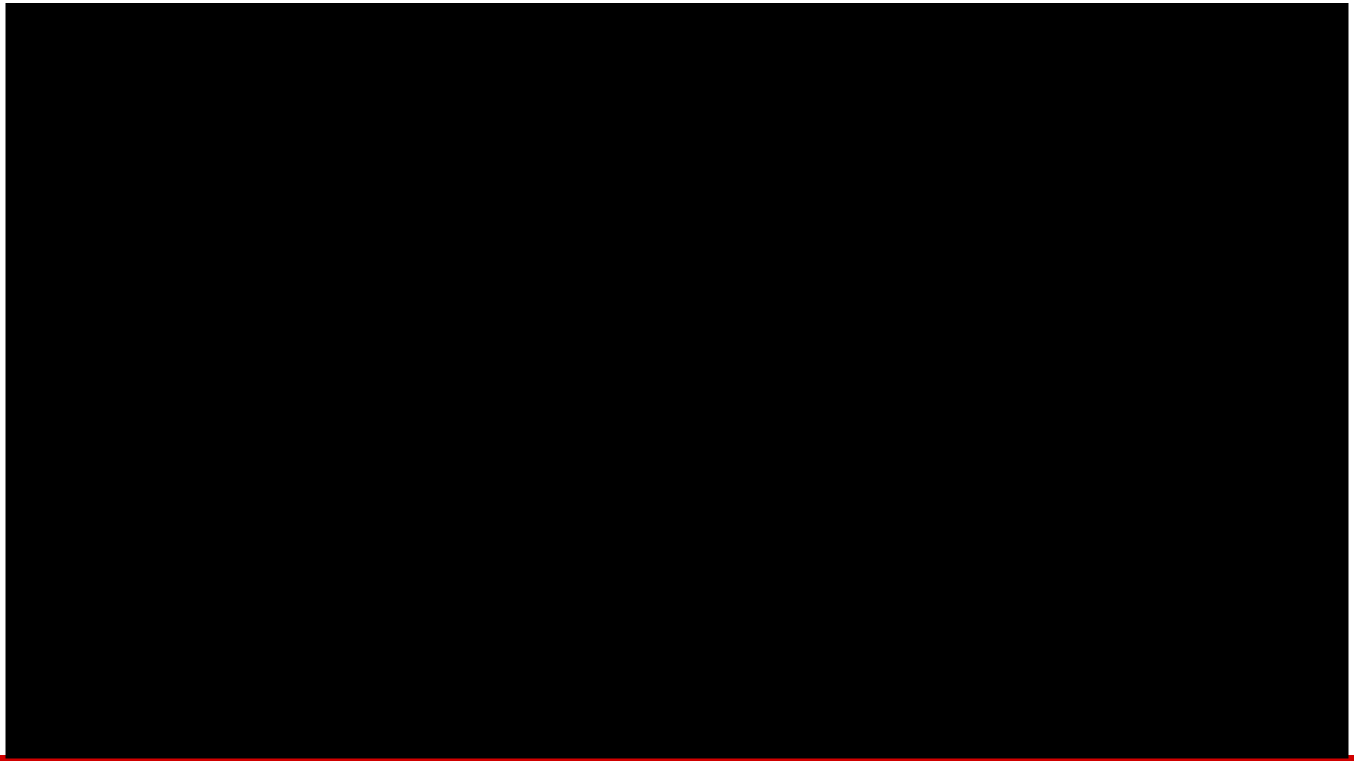
Chen, R., Chen, X., Ni, B., & Ge, Y. (2020). Simswap: an efficient framework for high fidelity face swapping. ACM International Conference on Multimedia (pp. 2003–2011).

This Week

- General Tampering
- Deepfake
- **Deepfake Videos**
- Detection

Deepfake Videos

- More dimensions:
  - Timing information
  - The relative position of different subjects and objects
  - Audio fakes

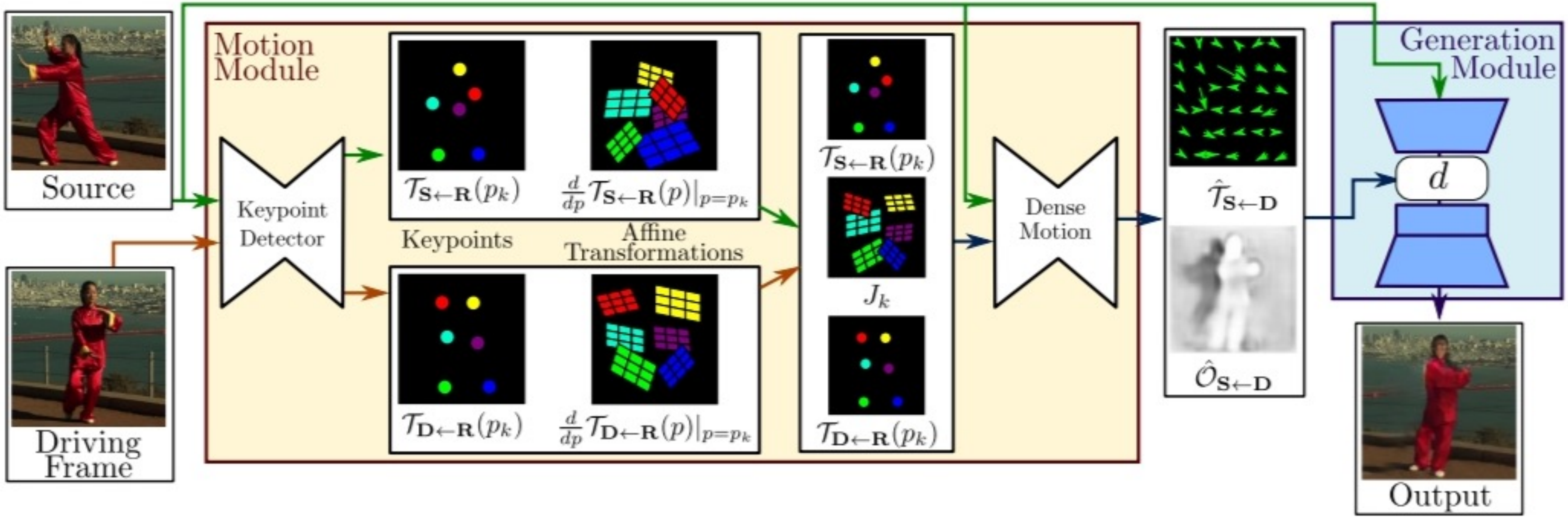


Deepfake Videos

- Challenges
  - How to generate reasonable gestures
  - How to generate a fake video in high resolution
  - How to generate high-quality long videos

Reasonable Gestures

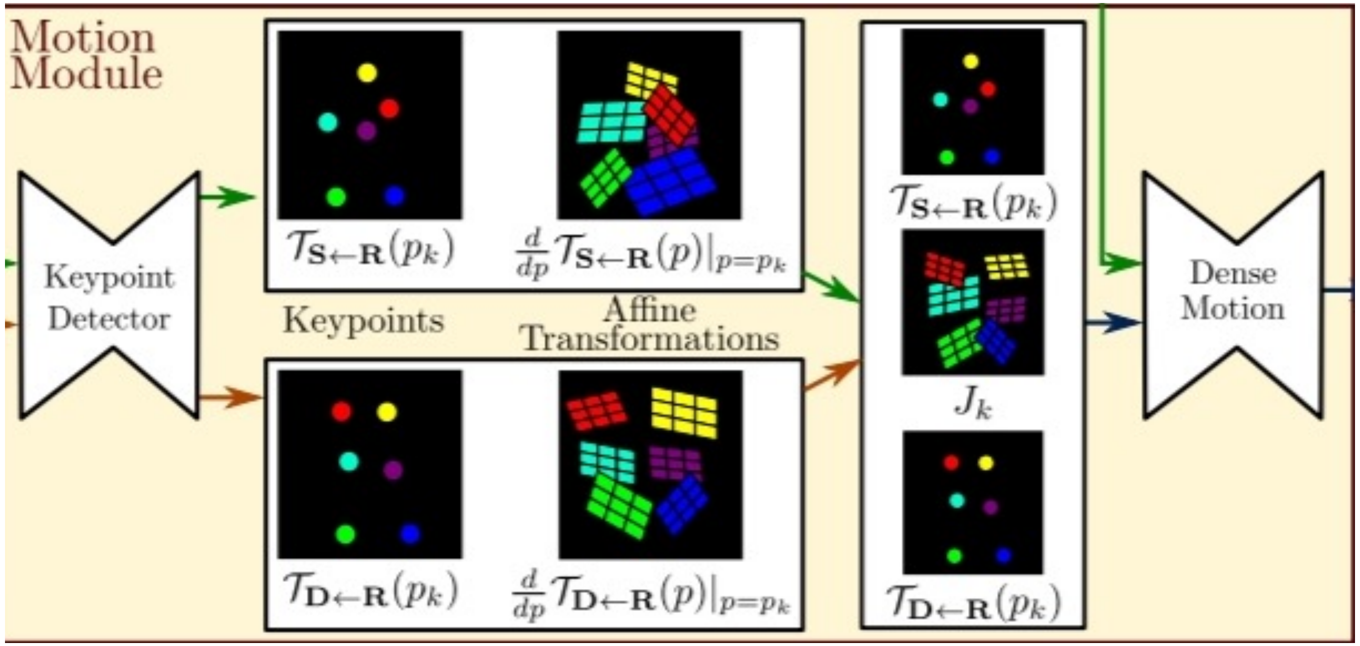
- First-order-motion Model



Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. *Advances in Neural Information Processing Systems*, *32*.

Reasonable Gestures

- Motion Estimation Module

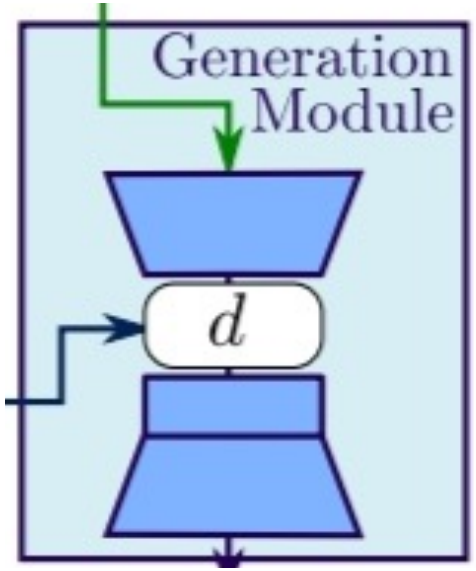


Use a set of ***learned key points*** and their ***affine transformations*** to predict dense motion

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. *Advances in Neural Information Processing Systems*, *32*.

Reasonable Gestures

- Generation Module

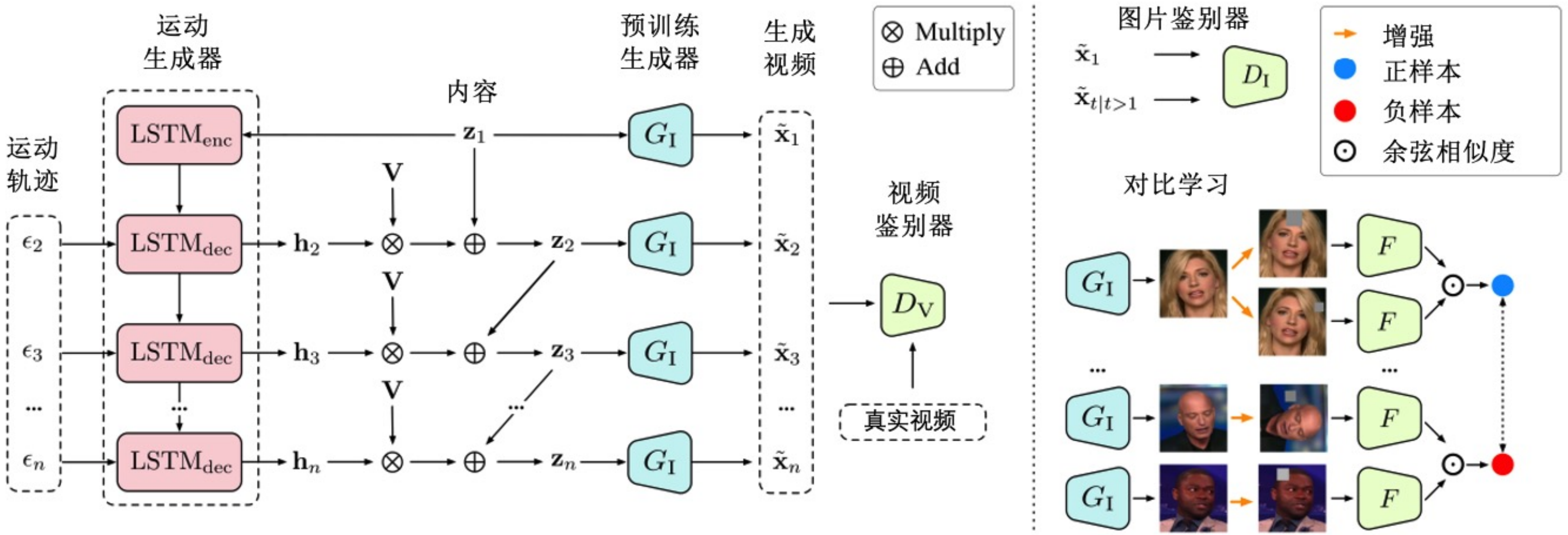


- ☐ Warp the source image according to $\hat{\mathcal{T}}_{\mathbf{S}\leftarrow\mathbf{D}}$
- ☐ Inpaint the image parts that are occluded in the source image.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. *Advances in Neural Information Processing Systems, 32*.

High Resolution

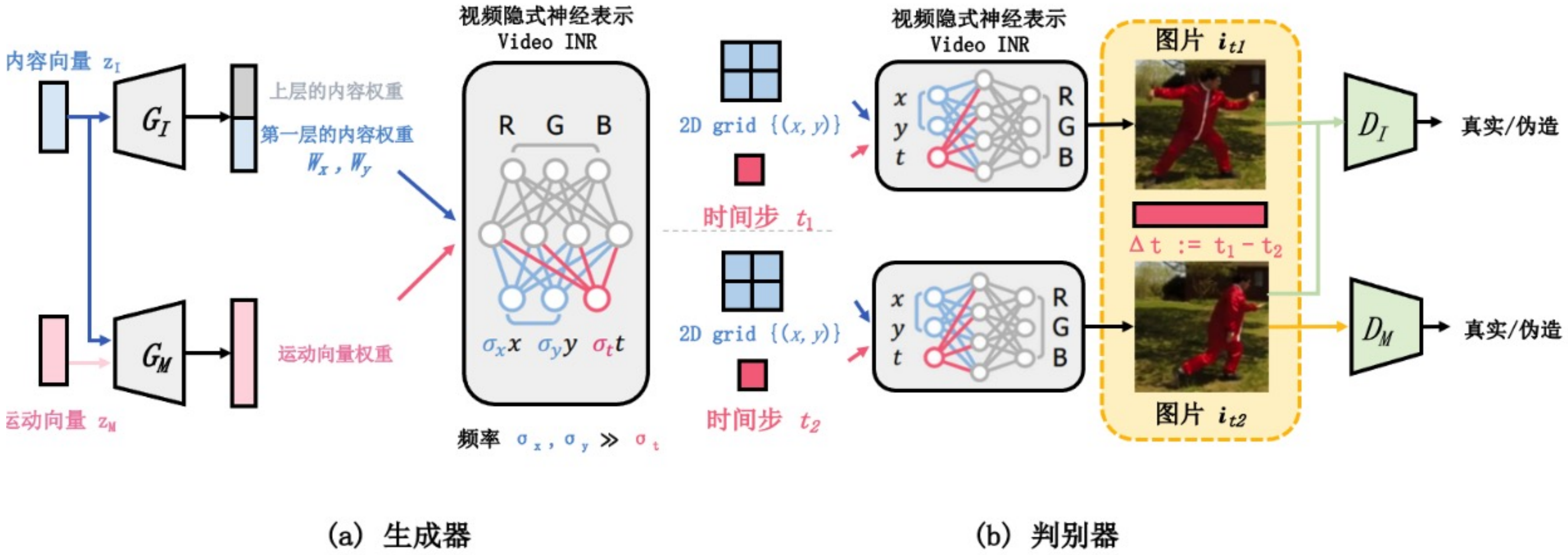
- MoCoGAN-HD



Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D. N., & Tulyakov, S. (2021). A good image generator is what you need for high-resolution video synthesis. International Conference on Learning Representations.

High-quality Long Videos

- DIGAN



Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.-W., & Shin, J. (2022). Generating videos with dynamics-aware implicit generative adversarial networks. arXiv preprint arXiv:2202.10571.

This Week

- General Tampering
- Deepfake
- Deepfake Videos
- **Detection**

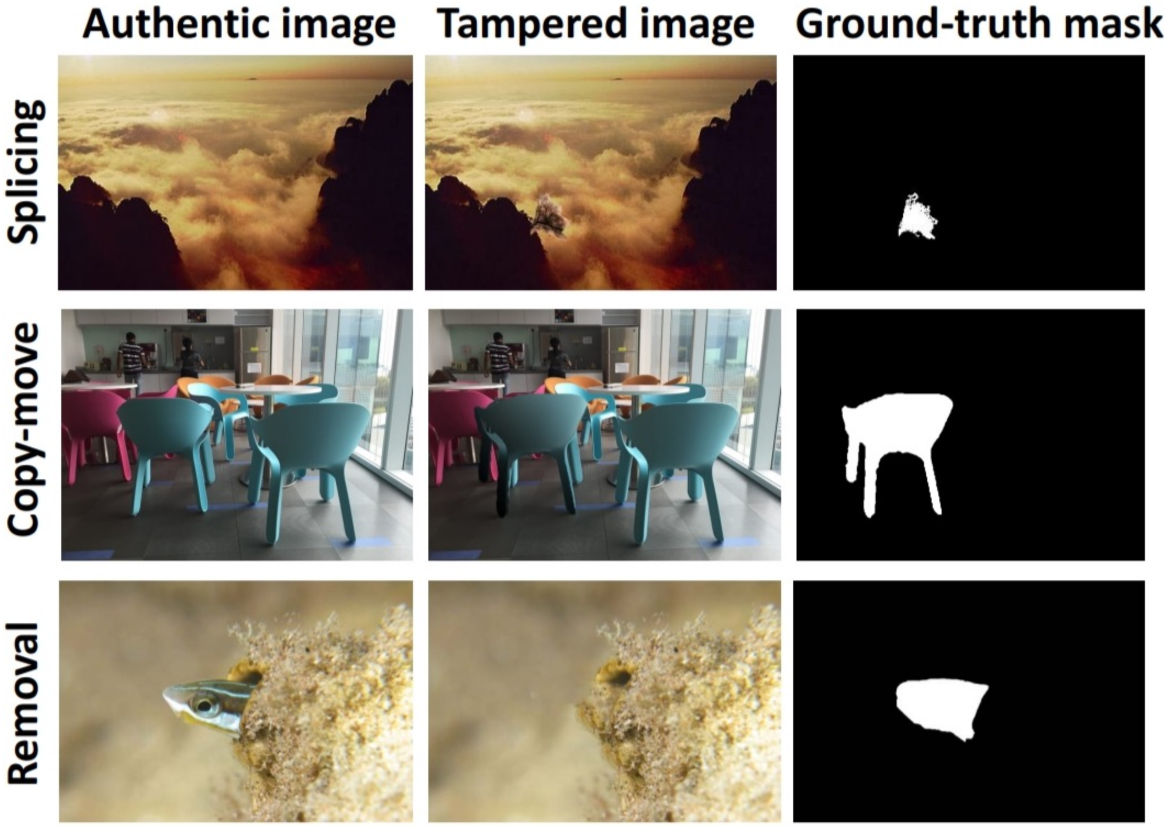
Tampering Detection

Features & Semantics

- Taxonomy:
  - **General Tampering Detection**——whether an ***ordinary object*** in an image has been tampered with
  - **Deepfake Detection**——whether the part of the ***face*** in the image has been tampered with

General Tampering Detection

- Existing general tampering detection methods mainly focus on ***splicing***, ***copy-move*** and ***removal***



General Tampering Detection

- Early detection methods

Image Tampering

The correlation between pixels introduced during camera imaging (LCA, …)

The frequency-domain or statistical features of the image and the noise it contains (PRNU)

General Tampering Detection

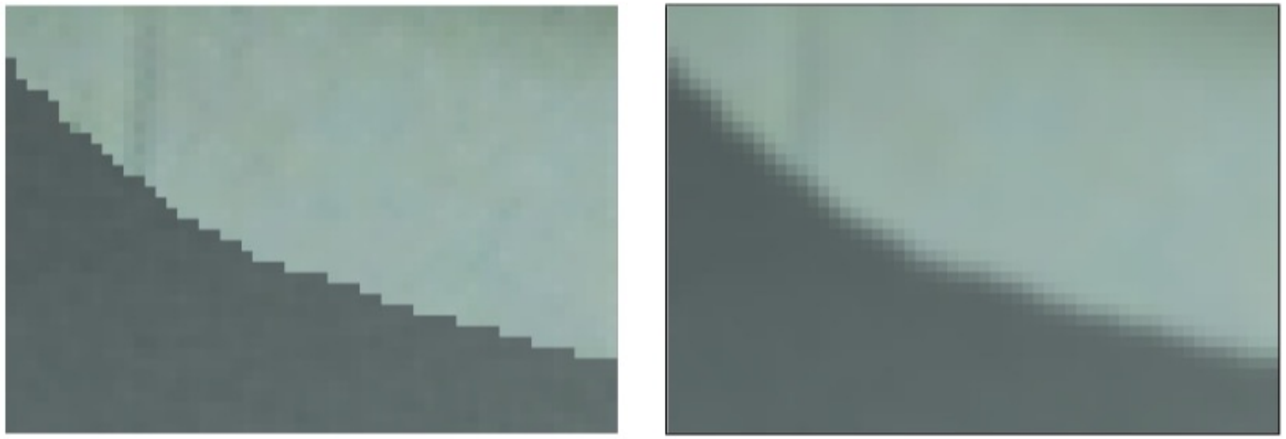
- Copy-move Detection Methods
  - Block-based region duplication

    Divide an image into many equal-size blocks, and if duplicated regions exist in the image, there should be duplicated blocks as well. Compare the blocks. (Pixel values, Statistical measures, Frequency coefficients, Moment invariants, …)
  - Keypoint-based region duplication

    Concentrate on a few keypoints within an image so the computation cost can be significantly reduced. (SIFT, SURF)

General Tampering Detection

- Splicing Detection Methods
  - ☐ Edge anomaly



(a) Edge pattern without blur (b) Edge pattern with Gaussian blur

  - ☐ Region anomaly: JPEG compression
  - ☐ Region anomaly: lighting inconsistency
  - ☐ Region anomaly: inconsistences of camera traces

General Tampering Detection

- Removal Detection Methods
  - ☐ Blurring artifacts by diffusion-based tampering
  - ☐ Block duplication by exemplar-based tampering

General Tampering Detection

- Later detection methods (DL)
  - ☐ Median filtering forensics + CNN (Chen et al., 2015)
  - ☐ RGB-N (Zhou et al., 2018)
  - ☐ SPAN, spatial pyramid attention network (Hu et al., 2020)
  - ☐ Mantra-Net (Wu et al., 2019)
  - ☐ PSCC-Net, progressive spatio-channel correlation network (Liu et al., 2022)

Countermeasures

- Detection
- Prevention

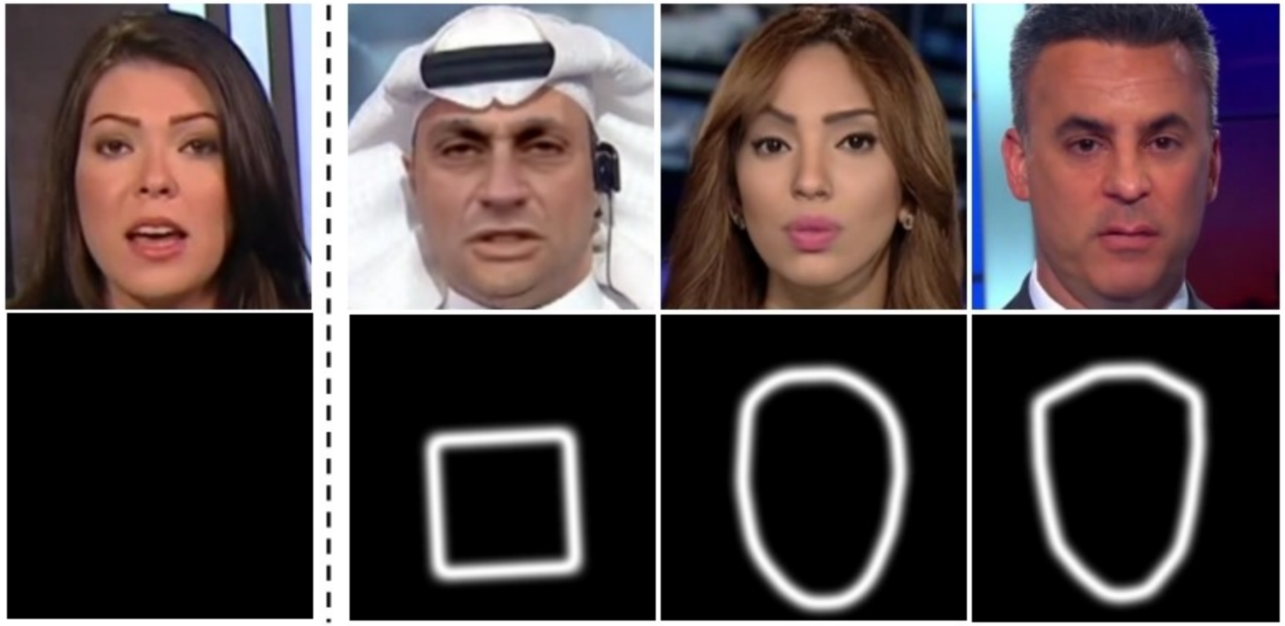
Detection | Artifact-specific

- Deepfakes often generate ***artifacts*** which may be subtle to humans, but can be easily detected using machine learning and forensic analysis.
    - ☐ Blending (spatial)
    - ☐ Environment (spatial)
    - ☐ Forensics (spatial)
    - ☐ Behavior (temporal)
    - ☐ Physiology (temporal)
    - ☐ Synchronization (temporal)
    - ☐ Coherence (temporal)

Mirsky Y, Lee W. The creation and detection of deepfakes: A survey[J]. ACM Computing Surveys (CSUR), 2021, 54(1): 1-41.

Blending

- Trained a CNN network to predict an image's blending boundary and a label (real or fake)



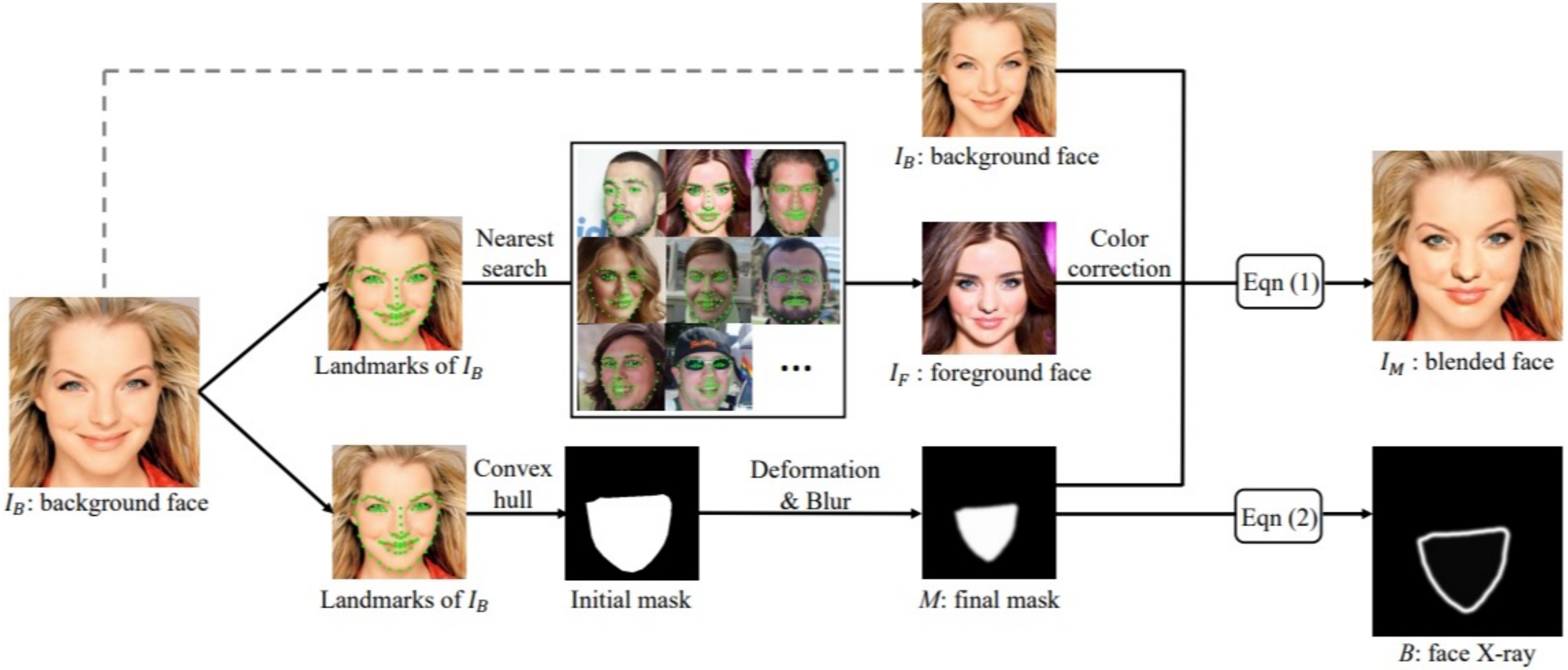
(a) (b)

Figure 1. Face X-ray reveals the blending boundaries in forged face images and returns a blank image for real images. (a) a real image and its face X-ray, (b) fake images and their face X-rays.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5001–5010.

Blending

- Splice similar faces found through facial landmark similarity to generate a dataset of face swaps.

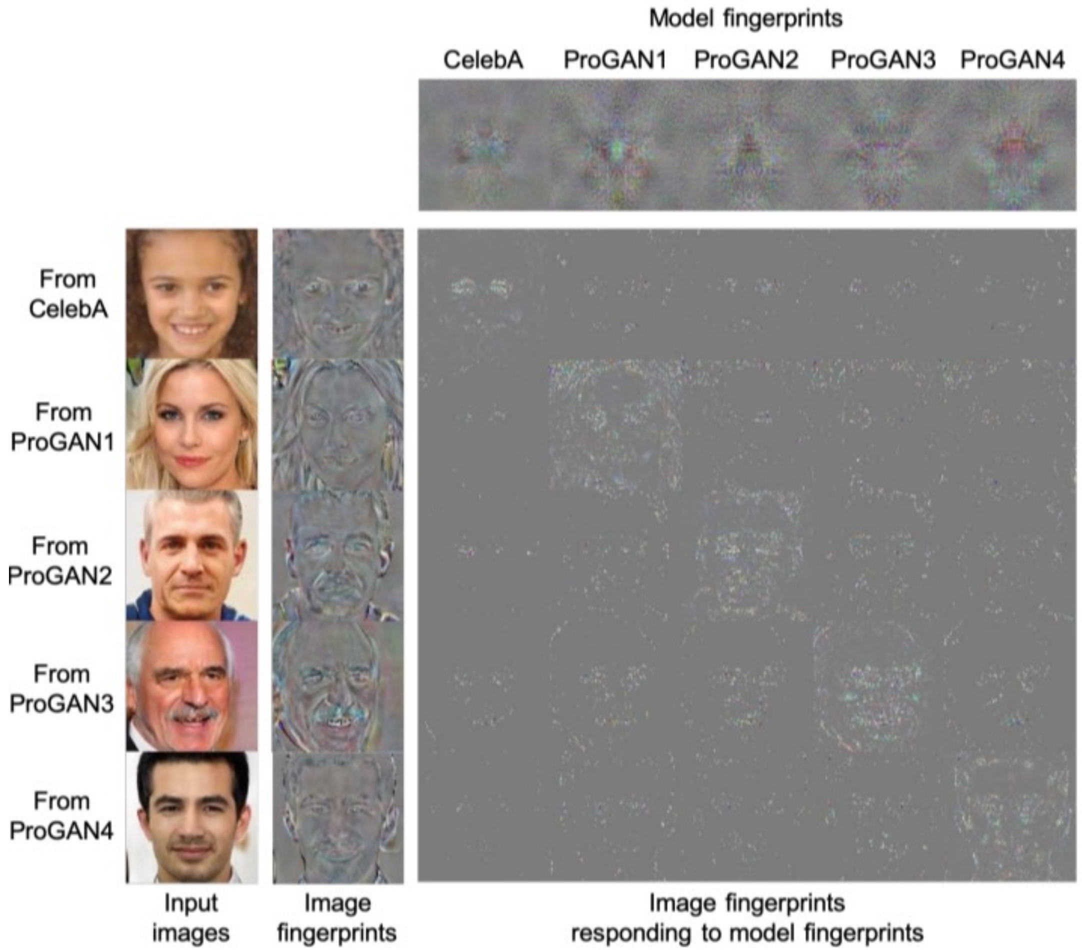


Overview of generating a training sample

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ern Recognition. 5001–5010.

Forensics

- Detect deepfakes by analyzing subtle features and patterns left by the model.
  - ☐GANs leave unique fingerprints
  - ☐It is possible to classify the generator given the content, even in the presence of compression and noise



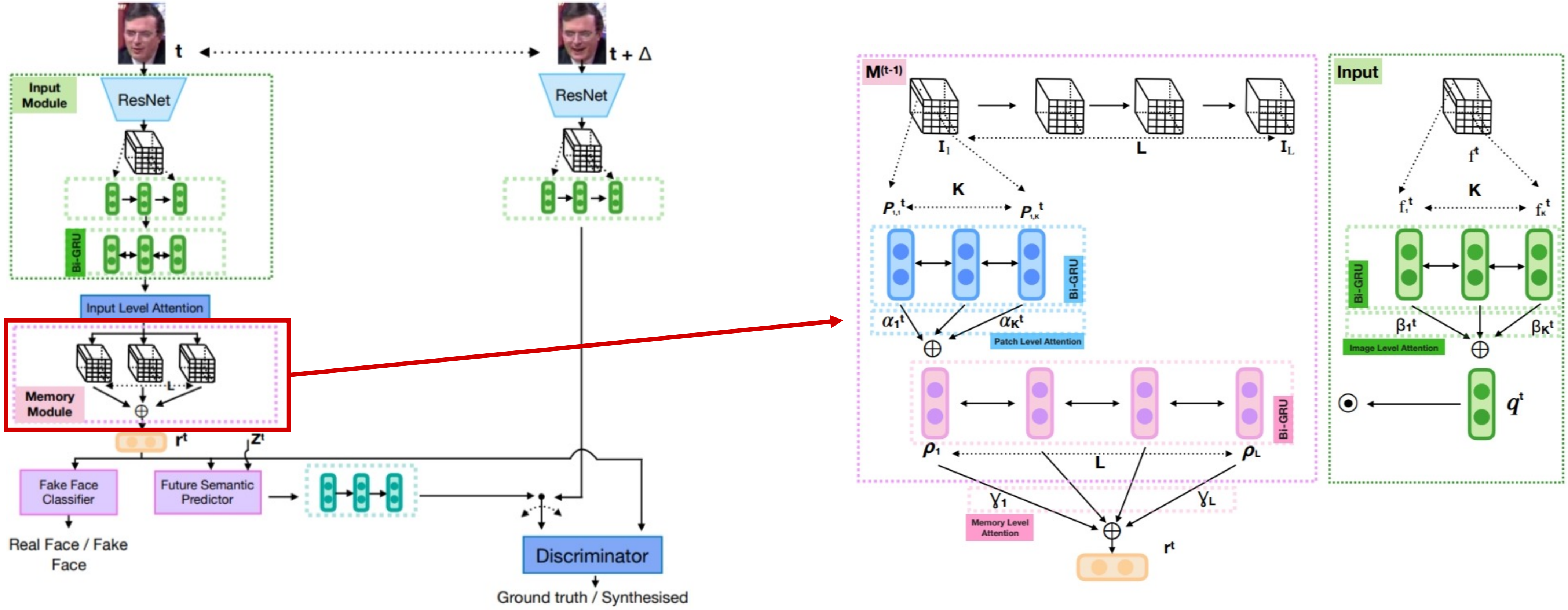
Ning Yu, Larry S Davis, and Mario Fritz. 2019. Attributing fake images to gans: Learning and analyzing gan ngerprints. In Proceedings of the IEEE International Conference on Computer Vision.

Detection | Undirected Approaches

- Train deep neural networks as generic ***classifiers***, and let the network decide which features to analyze.
  - ☐ Classification
  - ☐ Anomaly Detection

Classification

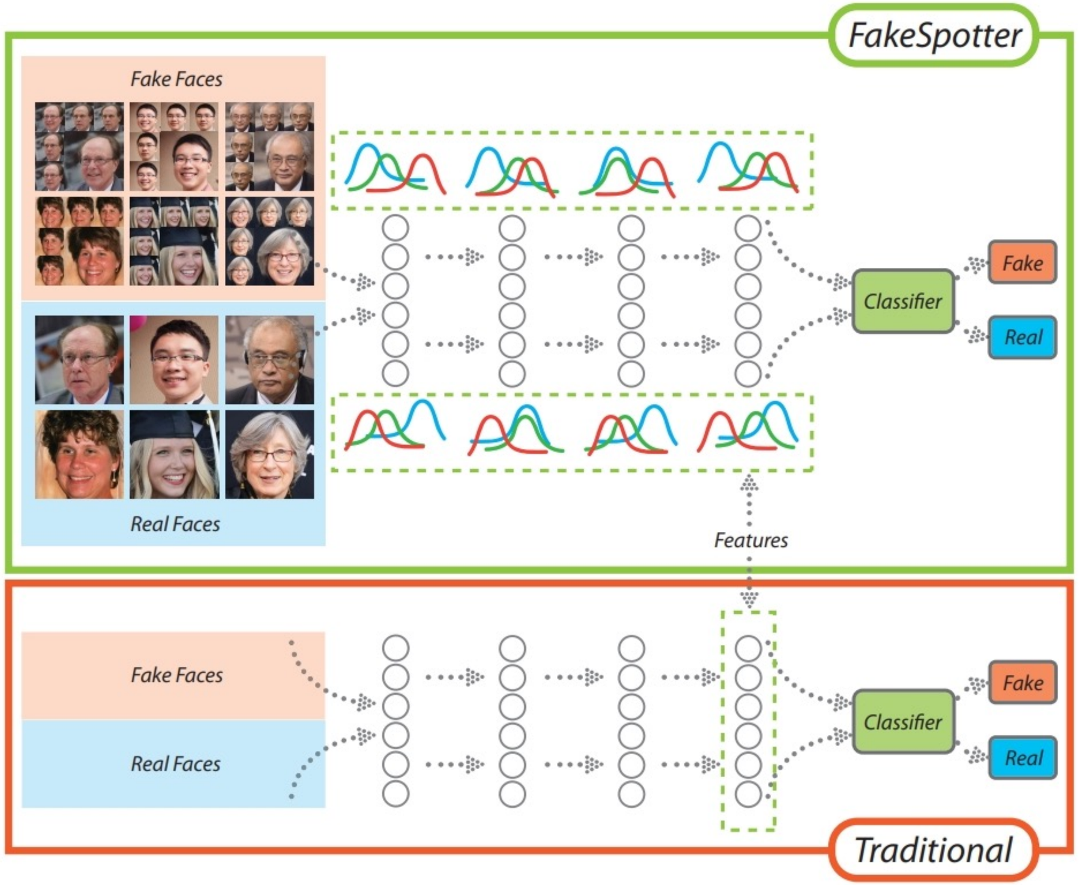
- Hierarchical Memory Network (HMN) architecture



Tharindu Fernando, Clinton Fookes, Simon Denman, and Sridha Sridharan. 2019. Exploiting Human Social Cognition for the Detection of Fake and Fraudulent Faces via Memory Networks. arXiv preprint arXiv:1911.07844 (2019).

Anomaly Detection

- anomaly detection models are trained on the normal data and then detect outliers during deployment.



- Monitor neuron behaviors(coverage) to spot AI-synthesized fake faces.
- Obtain a stronger signal from than just using the raw pixels.
- Is able to overcome noise and other distortions.

Run Wang, Lei Ma, Felix Juefei-Xu, Xiaofei Xie, Jian Wang, and Yang Liu. 2019. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. arXiv preprint arXiv:1909.06122 (2019).

Detection | Summary

			Type		Modality			Content				Method			Eval. Dataset								Performance*		
			Reenactment	Replacement	Image	Video	Audio	Feature	Body Part	Face	Image	Model	Indicates Affected Area	Input Resolution	DeepfakeTIMIT [86]	DFFD [149]	FaceForensics [130]	FaceForensics++ [131]	FFW [82]	Celeb-DF [101]	Other Deepfake DB	Custom	ACC	EER	AUC
Classic ML	[187]	2017		•	•	•				•		SVM-RBF		250x250								•	92.9		
	[4]	2017		•	•	•					•	SVM		*							•			18.2	
	[178]	2018	•	•	•	•				•		SVM		*							•				0.97
	[86]	2018	•	•		•				•		SVM		128x128	•									3.33	
	[42]	2019		•	•						•	SVM, Kmeans...		1024x1024								•	100		
	[8]	2019	•	•	•	•				•		SVM		*	•									13.33	
	[6]	2019	•	•		•				•		SVM		*								•			0.98
Statistics & Steganalysis	[85]	2018		•		•				•		PRNU		1280x720								•	*TPR=1*	*FPR=*	*0.03*
	[150]	2019		•		•					•	Statistics	•	-								•			
	[107]	2019	•	•	•						•	PRNU		*								•	90.3		

Detection | Summary

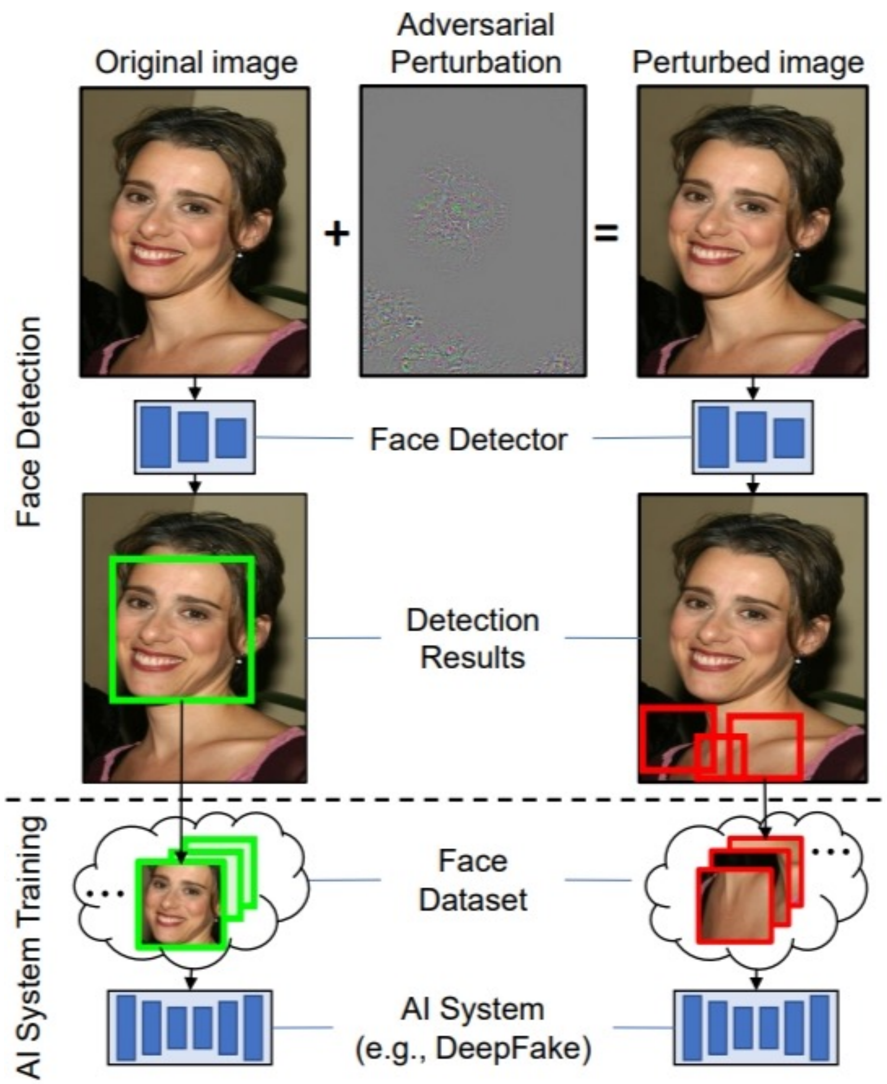
	[111]	2018	•	•	•	•				•		CNN		256x256								•	99.4		
	[97]	2018	•	•				•	•			LSTM-CNN		224x224								•			0.99
	[119]	2018	•	•	•	•				•		Capsule-CNN		128x128			•						99.3		
	[17]	2018		•	•	•				•		ED-GAN		128x128								•	92		
	[39]	2018	•	•	•					•		CNN		1024x1024							•				0.81
	[63]	2018	•	•		•				•		CNN-LSTM		299x299								•	97.1		
	[106]	2018	•	•	•						•	CNN		256256								•	94.4		
	[33]	2018	•	•	•	•					•	CNN AE		256x256			•					•	90.5		
	[3]	2018	•	•		•				•		CNN		256x256								•			0.99
	[132]	2019	•	•		•				•		CNN-LSTM		224x224				•					96.9		
	[118]	2019	•	•	•	•				•		CNN-DE	•	256x256			•	•					92.8	8.18	
	[38]	2019		•	•	•				•		CNN		-								•	98.5		
	[41]	2019	•	•	•	•					•	CNN AE GAN	•	256x256				•				•	99.2		
	[149]	2019	•	•	•	•				•		CNN+Attention	•	299x299		•								3.11	0.99
Deep Learning	[98]	2019	•	•	•	•				•		CNN		128x128							•	•			0.99
	[101]	2019		•		•				•		CNN		*						•					0.64
	[52]	2019	•	•	•	•				•		CNN+HMN		224x224			•	•	•				99.4		
	[92]	2019	•	•	•	•				•		FCN		256x256				•					98.1		
	[177]	2019	•	•	•	•					•	CNN		128x128								•	94.7		
	[161]	2019	•	•	•	•				•		CNN		224x224							•		86.4		
	[153]	2019	•	•	•					•		CNN		1024x1024								•			94
	[30]	2019	•	•		•				•		CNN		128x128			•					•	96		
	[99]	2019	•	•	•	•					•	CNN	•	224x224	•									93.2	
	[11]	2019	•	•		•				•		CNN		224x224				•					81.6		
	[?]	2019	•			•	•	•	•			LSTM		*								•		22	
	[47]	2019	•			•	•	•	•			LSTM-DNN		*								•		16.4	
	[25]	2019	•			•					•	CNN		256x256								•	97		
	[180]	2019	•	•	•			•			•	CNN		128x128			•		•			•	99.6	0.53	
	[166]	2019	•	•	•	•				•		SVM+VGGnet		224x224				•					85		
	[94]	2019	•	•	•	•		•				CNN	•	64x64		•		•		•	•				99.2
	[95]	2020	○	•	•					•	•	HRNet-FCN	•	64x64		•		•			•			20.86	0.86
	[96]	2020	•	•	•					•	•	PP-CNN		-	•		•				•				0.92
	[123]	2020	•	•	•					•	•	ED-CNN		299x299				•		•	•				0.99
	[108]	2020	•	•		•				•		ED-LSTM		224x224				•		•					
	[167]	2020	•	•	•						•	CNN ResNet		224x224				•				•	Avrg.	Prec.=	0.93
	[64]	2020	•	•	•					•		AREN-CNN		128x128			•					•	98.52		
	[110]	2020	•	•		•	•	•		•		ED-CNN		*	•	•					•				0.92
	[5]	2020	•	•		•	•		•			CNN		128x128							•		89.6		
	[10]	2020	•	•		•				•	•	LSTM		256x256				•					94.29		
	[69]	2020	•	•	•					•		Siamese CNN		64x64								•	TPR=0.91		
	[129]	2020	•	•	•					•		Ensemble		224x224				•					99.65		1.00
	[36]	2020	•	•		•				•		*		112x112						•			98.26		99.73
	[81]	2020	•	•	•					•		OC-VAE		100x100				•					TPR=0.89		
	[51]	2020	•	•	•					•		ABC-ResNet		224x224								•		?	

Prevention & Mitigation

- **Data provenance(数据溯源)**
  - Data provenance of multimedia should be tracked through distributed ledgers and blockchain networks.(Fraga-Lamas et al., 2019)
  - The content should be ranked by participants and AI.(Chen et al., 2019.)
  - The content should authenticated and managed as a global file system over Etherium start contracts.(Hasan et al., 2019)
- **Counter attacks(反击)**
  - Adversarial machine learning

Adversarial machine learning

- Use designed imperceptible adversarial perturbations to reduce the quality of the detected faces.



Yuezun Li, Xin Yang, Baoyuan Wu, and Siwei Lyu. 2019. Hiding Faces in Plain Sight: Disrupting AI Face Synthesis with Adversarial Perturbations. arXiv preprint arXiv:1906.09288 (2019).

Crux in Deepfake Detection

- Generalization of different dataset: Models trained on one dataset perform poorly on other datasets

Training Set	Testing Set	Xception [48]	Multi-task [40]	Capsule [41]	DSW-FPA [33]
FF++	FF++	99.7	76.3	96.6	93.0
	Celeb-DF	48.2	54.3	57.5	64.6
	SR-DF	37.9	38.7	41.3	44.0
SR-DF	SR-DF	88.2	85.7	81.5	86.6
	FF++	63.2	58.9	60.6	69.1
	Celeb-DF	59.4	51.7	52.1	62.9

Crux in Deepfake Detection

- Existing datasets: There are common problems such as ***obvious visual defects*** and ***insufficient diversity***

(a) FF++ [48] (b) DFD [9] (c) DFDC [13] (d) Celeb-DF [35]

(a) Celeb-DF (b) FF++ (c) SR-DF (ours)

The detection method can easily achieve high detection accuracy after training and testing on existing datasets, but its performance is poor on high-quality falsified data.

High-quality Deepfake Dataset

- SR-DF Dataset:

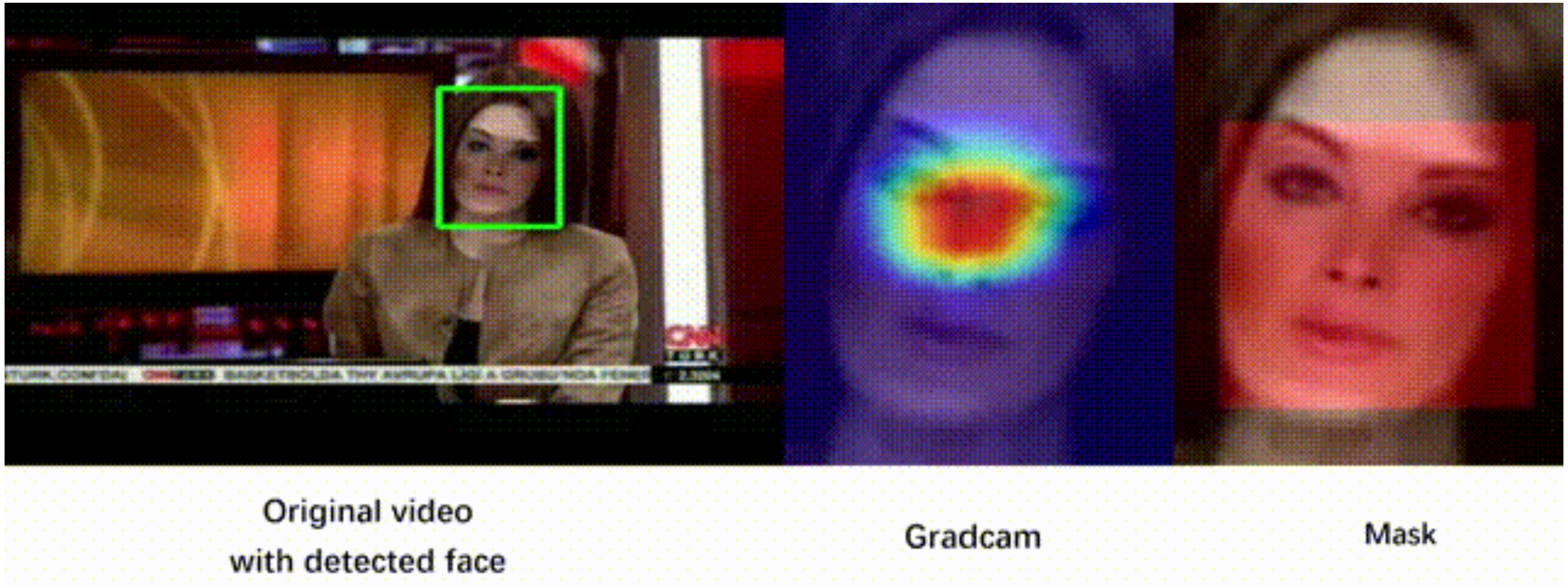
1. Use SOTA methods to generate deepfake images:

   (1) First-order-motion [Aliaksandr Siarohin, NIPS 2019] **(2) IcFace** [Soumya Tripathy, WACV 2020] **(3) FSGAN** [Yuval Nirkin, ICCV 2019] **(4). FaceShifter** [Lingzhi Li, Arxiv].

2. Post-processing： **DoveNet** [Cong Wenyan et al. CVPR 2020.]

Open-source Tools

PyDeepFakeDet



https://github.com/wangjk666/PyDeepFakeDet

Open-source Tools

PyDeepFakeDet

The baseline Models on Celeb-DF is also availabl

Method	Celeb-DF	Model
ResNet50	98.51	CelebDF
Xception	99.05	CelebDF
EfficientNet-b4	99.44	CelebDF
Meso4	73.04	CelebDF
MesoInception4	75.87	CelebDF
GramNet	98.67	CelebDF
F3Net	96.47	CelebDF
MAT	99.02	CelebDF
ViT	96.73	CelebDF
M2TR	99.76	CelebDF

Model Zoo and Baselines

The baseline Models on three versions of FF-DF dataset are provided.

Method	RAW	C23	C40	Model
ResNet50	97.61	94.87	84.95	RAW / C23 / C40
Xception	97.84	95.24	86.27	RAW / C23 / C40
EfficientNet-b4	97.89	95.61	87.12	RAW / C23 / C40
Meso4	85.14	77.14	60.13	RAW / C23 / C40
MesoInception4	95.45	84.13	71.31	RAW / C23 / C40
GramNet	97.65	95.16	86.21	RAW / C23 / C40
F3Net	99.95	97.52	90.43	RAW / C23 / C40
MAT	97.90	95.59	87.06	RAW / C23 / C40
ViT	96.72	93.45	82.97	RAW / C23 / C40
M2TR	99.50	97.93	92.89	RAW / C23 / C40

https://github.com/wangjk666/PyDeepFakeDet

谢谢！