深度伪造与检测技术

程 健 研究员

jcheng@nlpr.ia.ac.cn

中国科学院自动化研究所

2022.10.31

大 纲

- 深度伪造介绍
- 基于生成模型的深度伪造技术
- 反伪造检测技术
- 总结与展望

大 纲

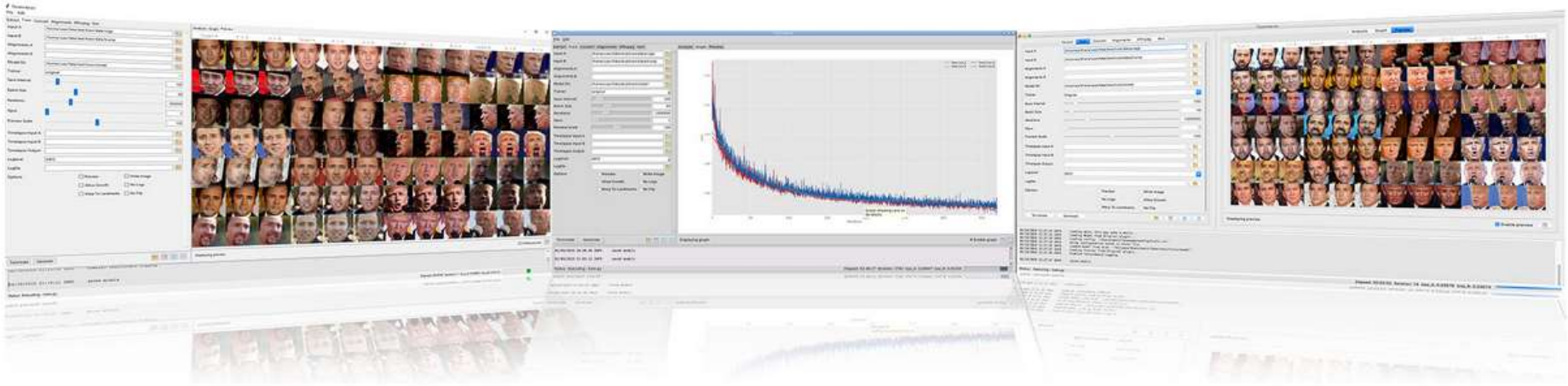
- **深度伪造介绍**
- 基于生成模型的深度伪造技术
- 反伪造检测技术
- 总结与展望

深度伪造技术

- 深度伪造（Deepfakes）通过自动化的手段特别是使用人工智能技术，对数据进行智能生产、操纵、修改。
- Deepfakes = Deep Learning + Fake
- 常见应用如：AI换脸、视频合成、语言合成等。

faceswap

FaceSwap is a tool that utilizes deep learning to recognize and swap faces in pictures and videos.

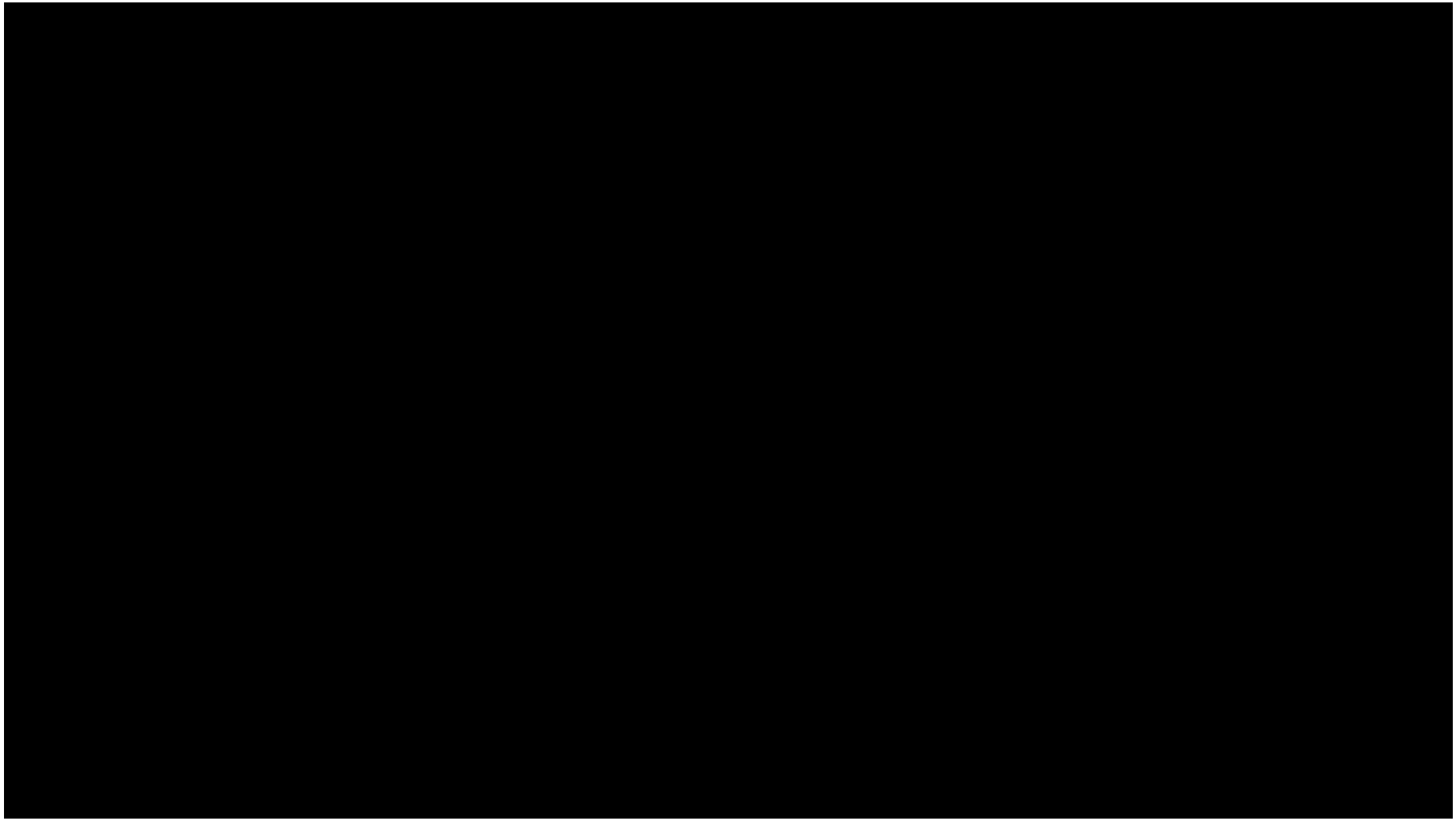


眼见为实？

眼见为实？

《阿甘正传》片段中肯尼迪总统会见阿甘

眼见为实？



Deepfakes

深度伪造的发展

Deepfakes技术（AI换脸）发展过程：

- 2014.06：提出生成对抗网络（GAN），在图片生成方面取得重大突破；
- 2017.07：提出一种使用LSTM学习口腔形状和声音之间关联性的方法，仅通过音频即可合成对应的口部特征；
- 2017.10：提出一种基于GAN的自动化实时换脸技术；
- 2018.04：出现美国前总统奥巴马的deepfakes演讲，使用了FakeApp；
- 2018.08：提出一种将源视频中的运动转移到另一个视频中目标人的方法，而不仅是换脸；
- 2019.03：提出一种控制图片生成器并能编辑造假图片各方面特性的方法，比如肤色、头发颜色和背景内容；
- 2019.05：提出一种真实头部说话型的少样本对抗学习，向其输入一张人物头像，可以生成人物头像开口说话的动图；
- 2020：主要进展在提高原生分辨率、提升deepfake制作效果方面。

深度伪造的发展

- 依赖于深度学习的迅猛发展，深度伪造技术伪造出的数据愈加接近真实数据，并能够对不同模态的数据进行伪造。

名称	时间	功能
FakeApp	2018	采用3D 图形学进行图片和视频合成
Faceswap	2019	采用3D 图形学进行图片和视频合成
Deepfakes	2019	采用自动编码器进行图片和视频合成
DeepFaceLab	2019	对Faceswap 项目的模型扩充，对人脸模型进行扩充
Faceapp	2019	人脸编辑器, 可以换脸, 换表情, 编辑人脸属性
ZAO	2019	指定的影视模板换脸, 只需一张目标人脸即可换脸
RealTalk	2019	实时语音合成
Deep-voicev-conversion	2020	只需要目标说话者的音波素材, 即可转换成特定目标人物的声音
MelNet	2020	基于频谱图的端到端语音生成

深度伪造技术的影响

- **消极影响：被用于误导舆论、扰乱社会秩序，威胁国家安全和公共安全、引发社会忧虑和信任危机等，已成为当前新型网络攻击形式。**
  - 虚假新闻：发布或歪曲知名政客的言论，愚弄公众等；
  - 语音诈骗：利用合成的语音进行金融诈骗；
  - 色情制作：2017年网络上出现AI换脸的色情视频；
  - 影像篡改：将个人面孔交换到电影明星身体插入影视剪辑中、移除图像中证据进行欺诈。

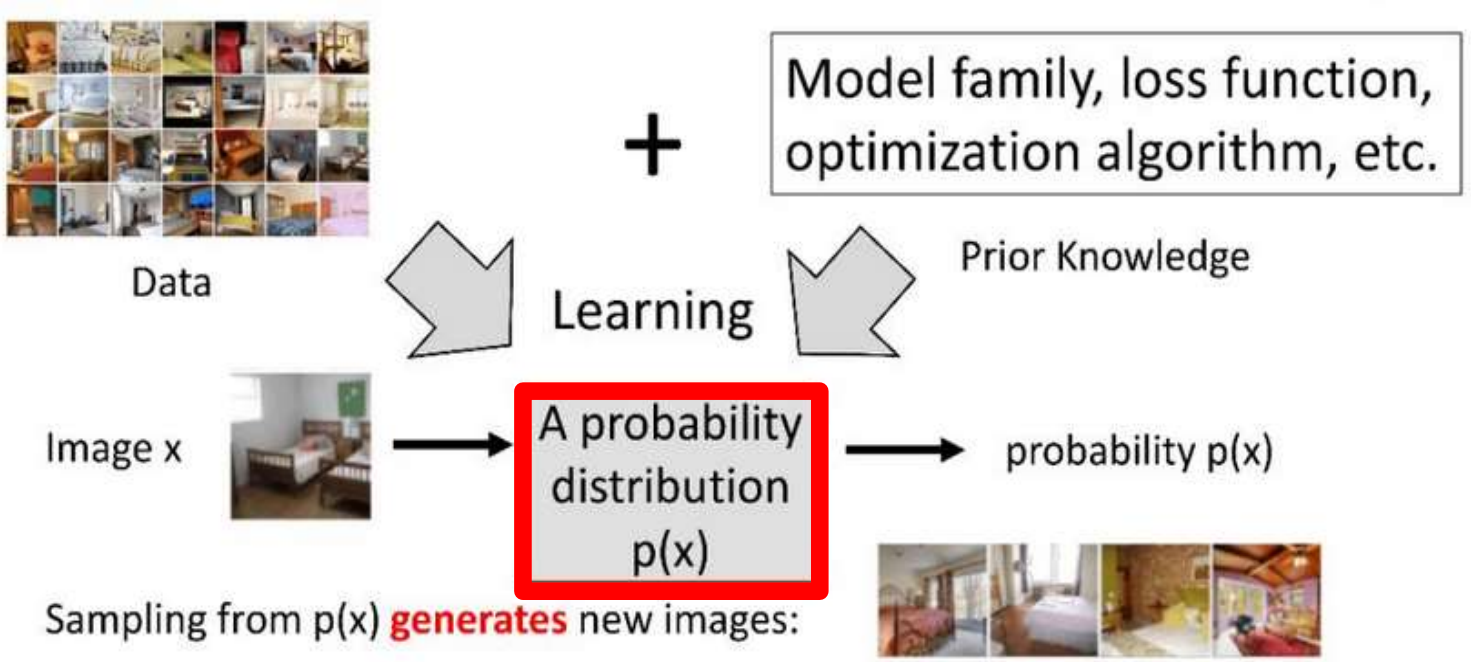
- **积极影响：推动娱乐与文化交流产业的新兴发展。**
  - 电影制作：电影制作中创建虚拟角色、视频渲染、声音模拟等；
  - 人物复活：“复活”历史人物，真实地还原历史人物的原貌。

大 纲

- 深度伪造介绍
- **基于生成模型的深度伪造技术**
- 反伪造检测技术
- 总结与展望

生成模型

- **生成模型**（Generative Model）是能够随机生成观测数据，或将已知结构的简单分布转换成复杂的分布的模型。
- 在机器学习中，生成模型可以用来直接对数据建模，也可以用来建立变量间的条件概率分布。



生成模型

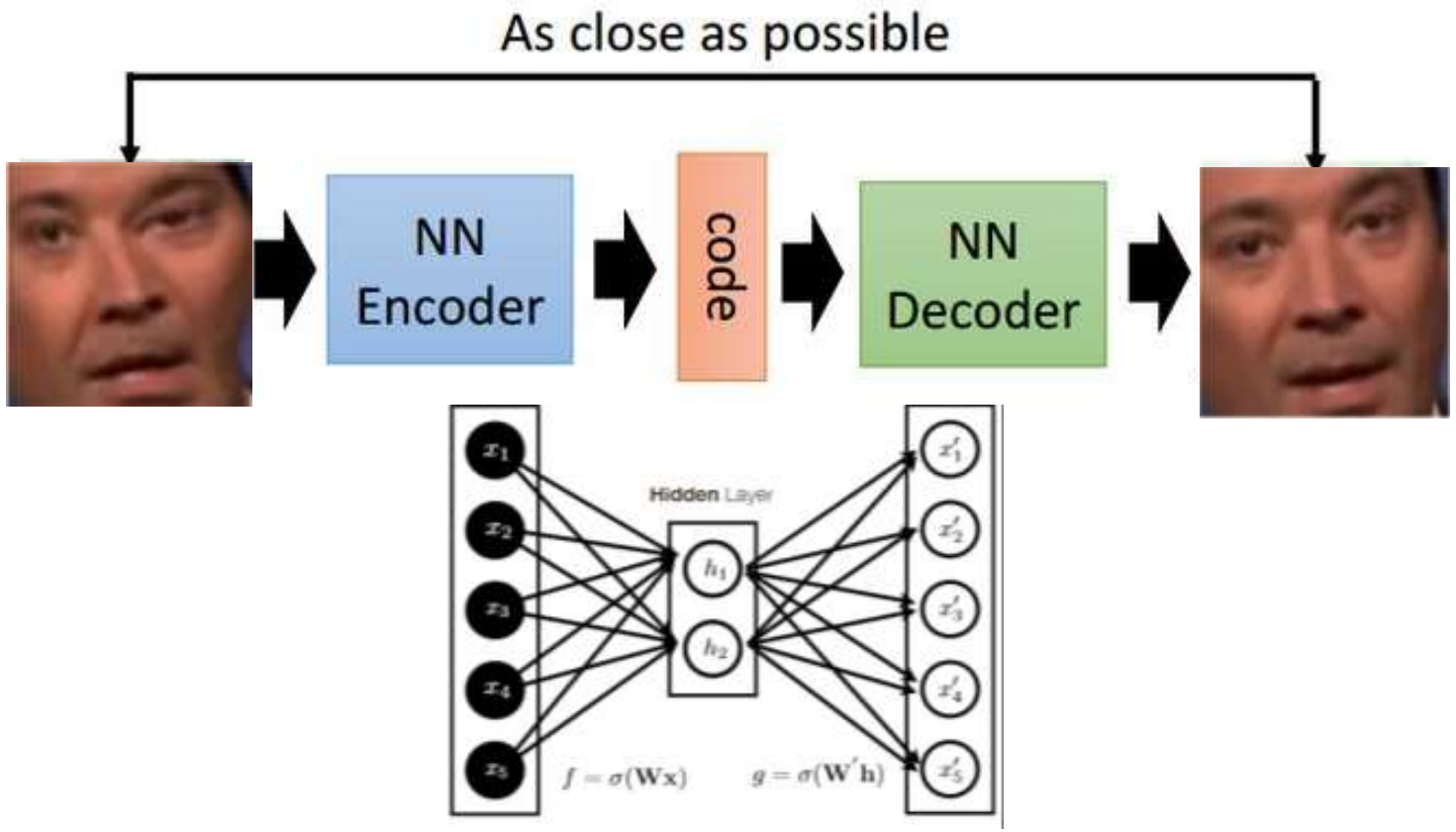
- 典型的生成模型
  - 自编码器（Autoencoder, AE）
  - 生成式对抗模型（Generative Adversarial Model, GAN）
  - 高斯混合模型（Gaussian Mixture Model, GMM）
  - 隐马尔可夫模型（Hidden Markov Model, HMM）
  - 朴素贝叶斯模型（Naive Bayesian Model, NBM）
  - 潜在狄利克雷分配模型（Latent Dirichlet Allocation Model, LDA）
  - 受限玻尔兹曼机（Restricted Boltzmann Machine, RBM）
  - ……
- 现有的Deepfakes方法基于AE或GAN实现。

基于AE的Deepfakes

- Recap

Autoencoders

Self (i.e. self-encoding)



基于AE的Deepfakes

- Recap: Autoencoder

— Feed forward network intended to reproduce the input

— Encoder/Decoder architecture

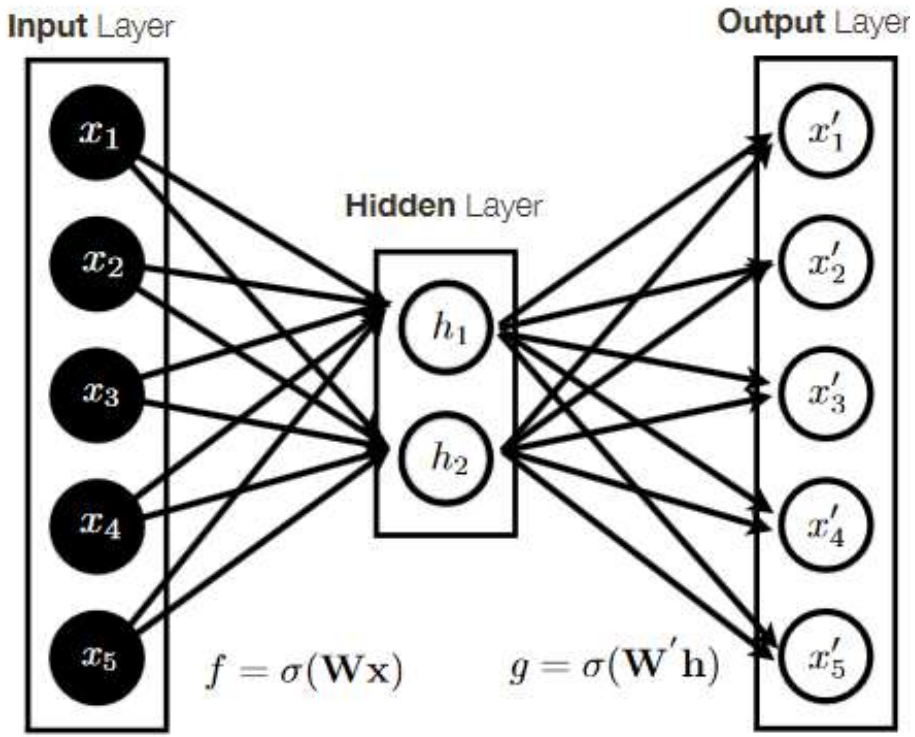
Encoder: $f = \sigma(\mathbf{W}\mathbf{x})$

Decoder: $g = \sigma(\mathbf{W}'\mathbf{h})$

— Score function

$$\mathbf{x}' = f(g(\mathbf{x}))$$

$$\mathcal{L}(\mathbf{x}', \mathbf{x})$$



基于AE的Deepfakes

- Deepfakes换脸原理
  - 分别使用A和B的人脸图片训练两个AE模型，训练过程中共享两个模型encoder而不共享decoder，达到encoder能同时抓取两人人脸关键信息，decoder能分别还原出两个人的脸的目的。

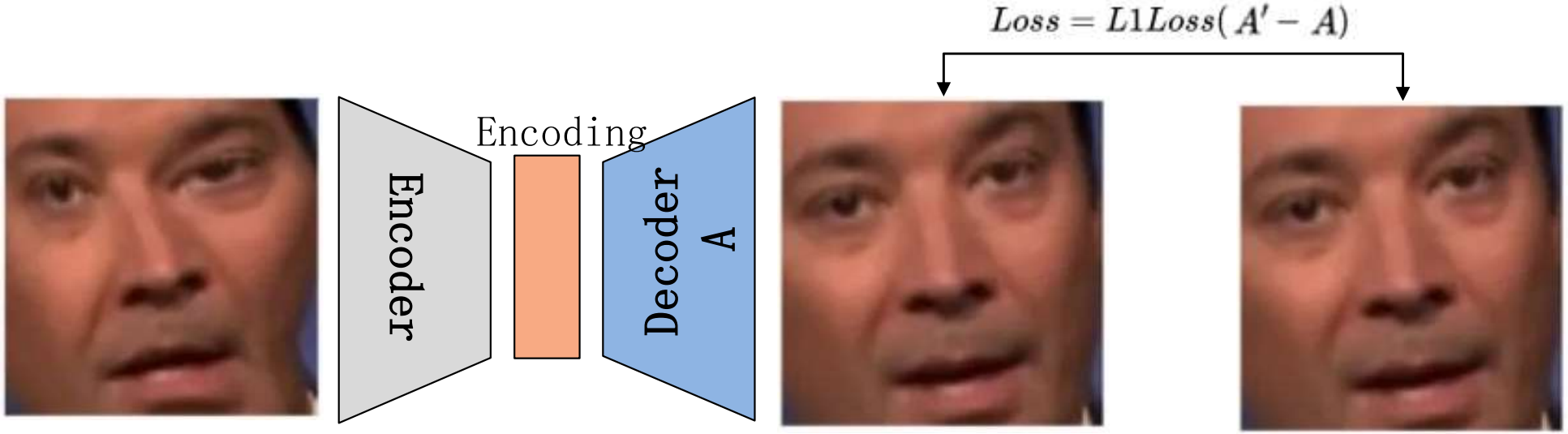
$$A' = Decoder_A(Encoder(AW)) \quad Loss = L1Loss(A' - A)$$

$$B' = Decoder_B(Encoder(BW)) \quad Loss = L1Loss(B' - B)$$

基于AE的Deepfakes

- Deepfakes换脸学习过程

1. 使用A的人脸训练模型A。

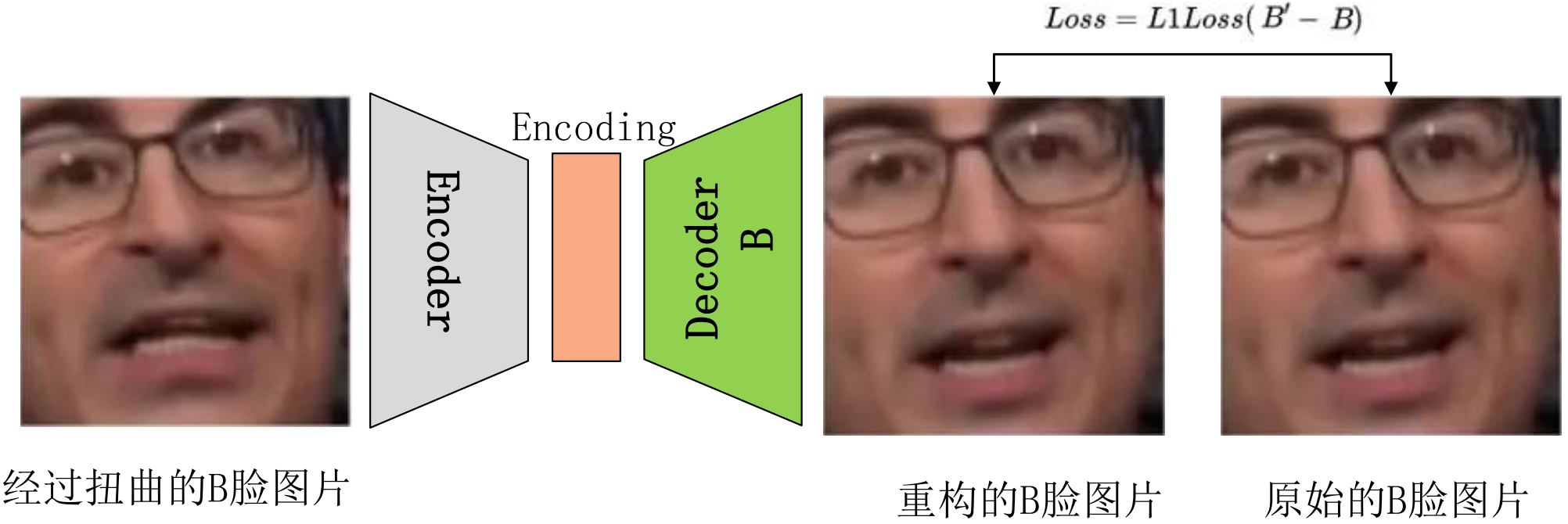


经过扭曲的A脸图片　　重构的A脸图片　　原始的A脸图片

基于AE的Deepfakes

- Deepfakes换脸学习过程

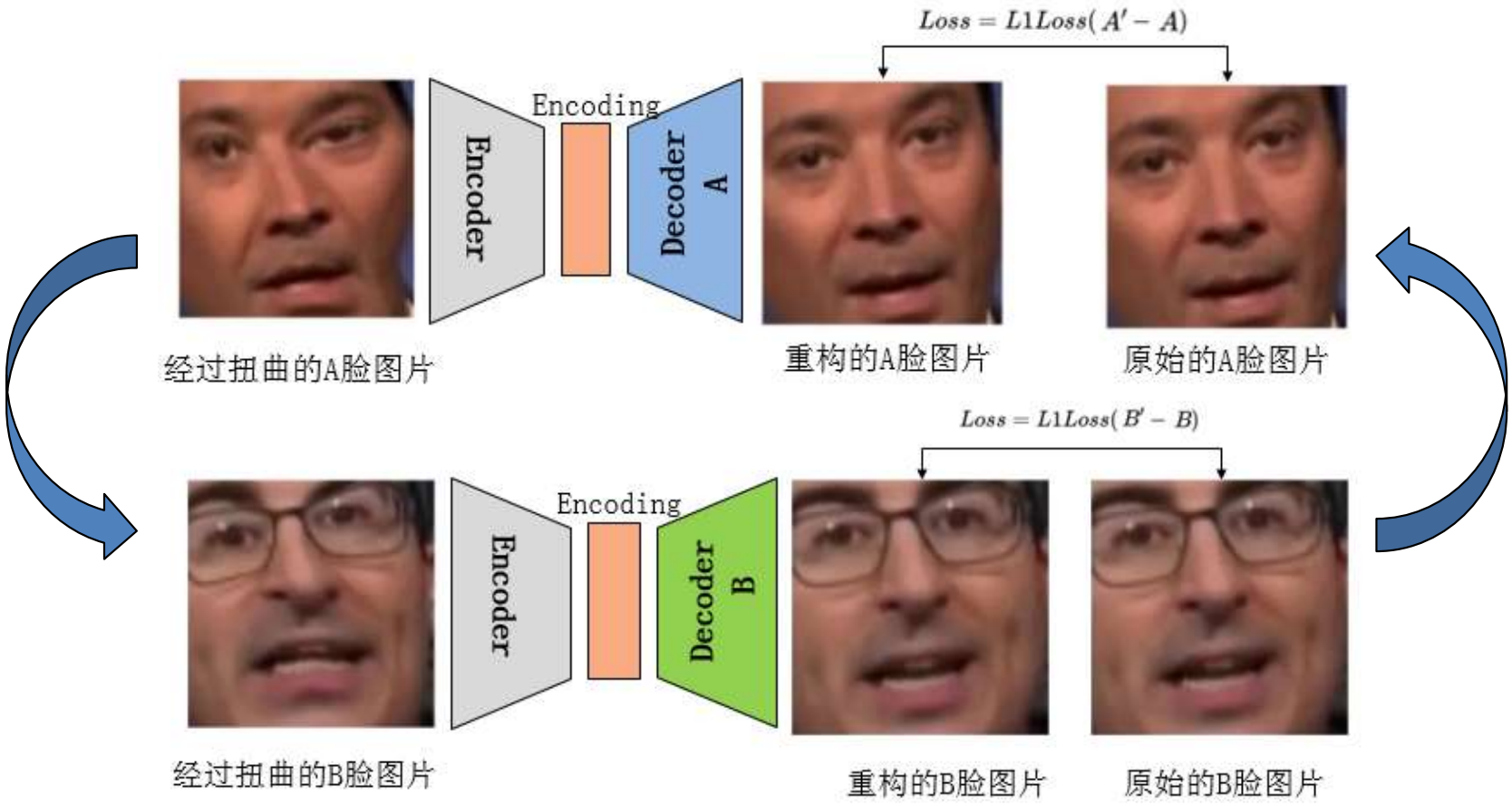
2. 使用B的人脸训练模型B。模型B与模型A共享encoder但不共享decoder。



基于AE的Deepfakes

- Deepfakes换脸学习过程

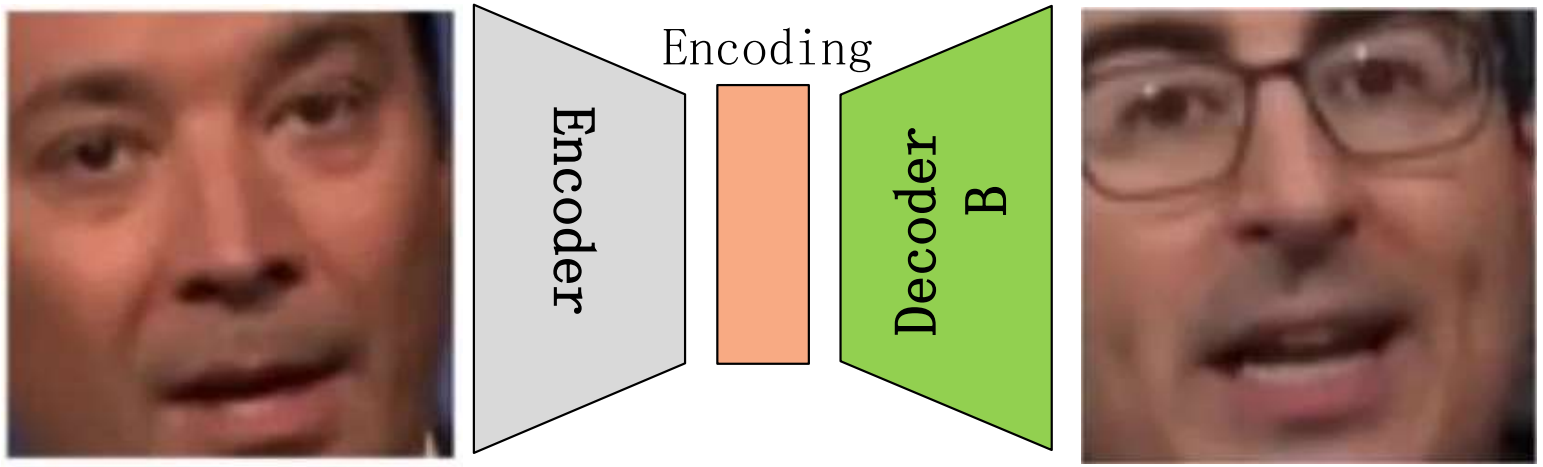
3. 重复以上两步直至收敛。



基于AE的Deepfakes

- Deepfakes换脸学习过程

4. 利用训练好的模型换脸。



原始A脸图片

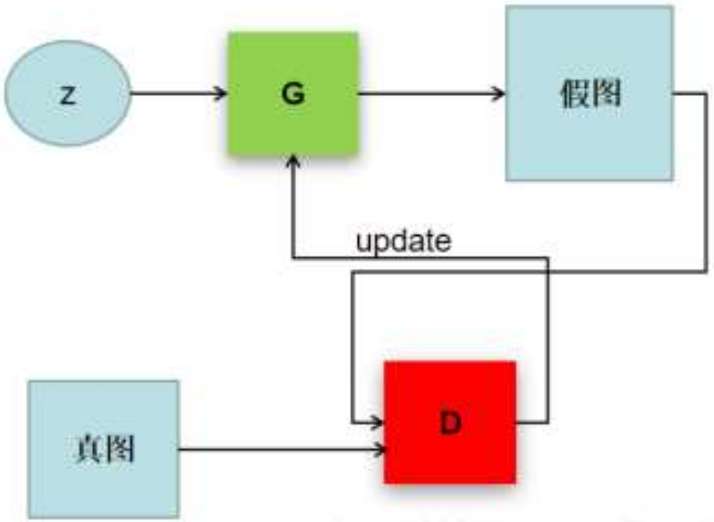
原始的B脸图片

基于AE的Deepfakes

- 基于AE的换脸缺点
  - 消耗大量计算资源。
  - 该方法只有在拥有大量目标人物图片和视频素材（300到2000张）作为训练数据的前提下才能达到相对理想的效果。

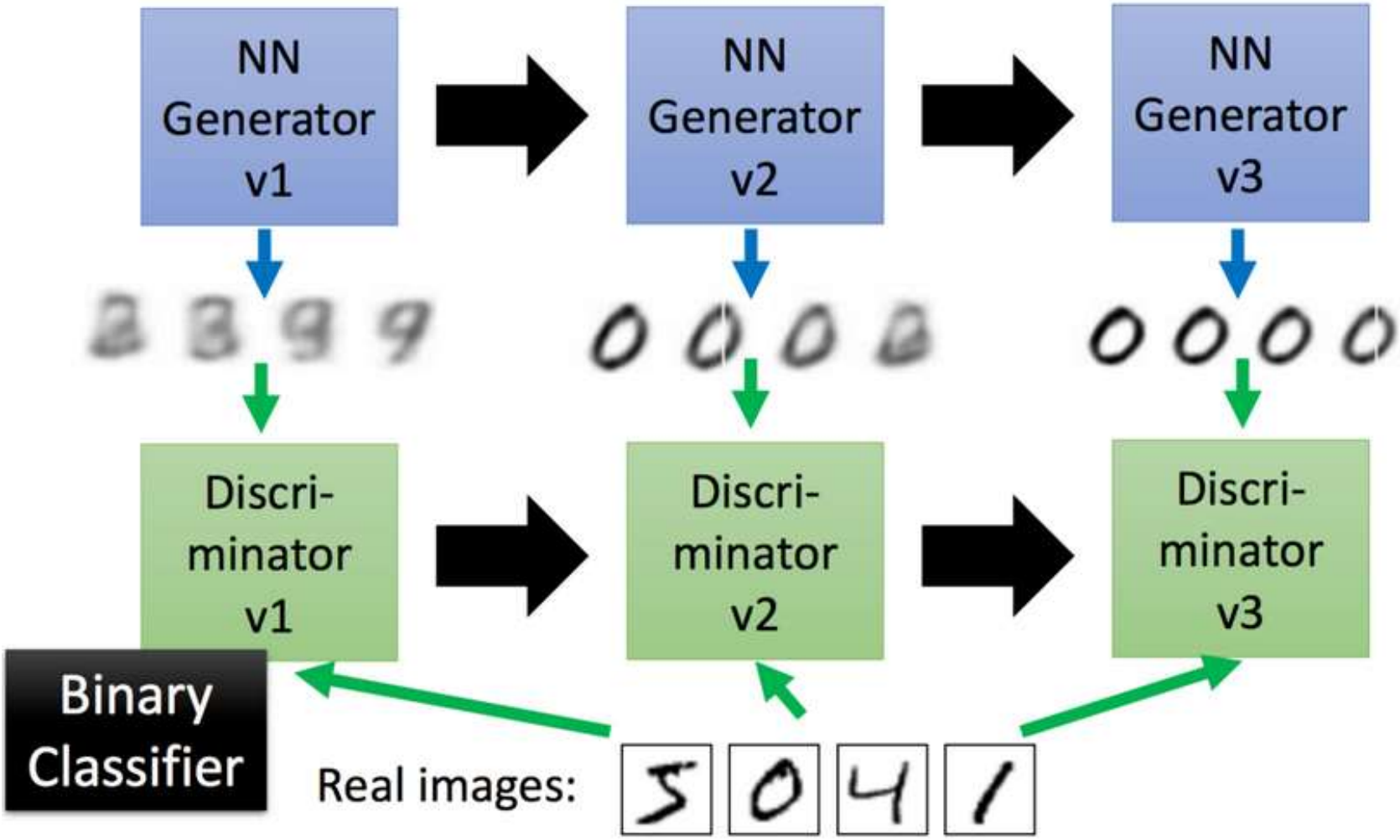
基于GAN的Deepfakes

- 生成式对抗模型（Generative Adversarial Network，GAN）基本概念
  - GAN由Ian Goodfellow等于2014年提出[NeurIPS2014]。
  - GAN模型包括两部分：生成器generator与判别器discriminator。生成器和判别器不断博弈对抗，最终达到动态均衡：生成器生成的图像接近于真实图像分布，而判别器识别不出真假图像，对于给定图像的预测为真的概率基本接近0.5。

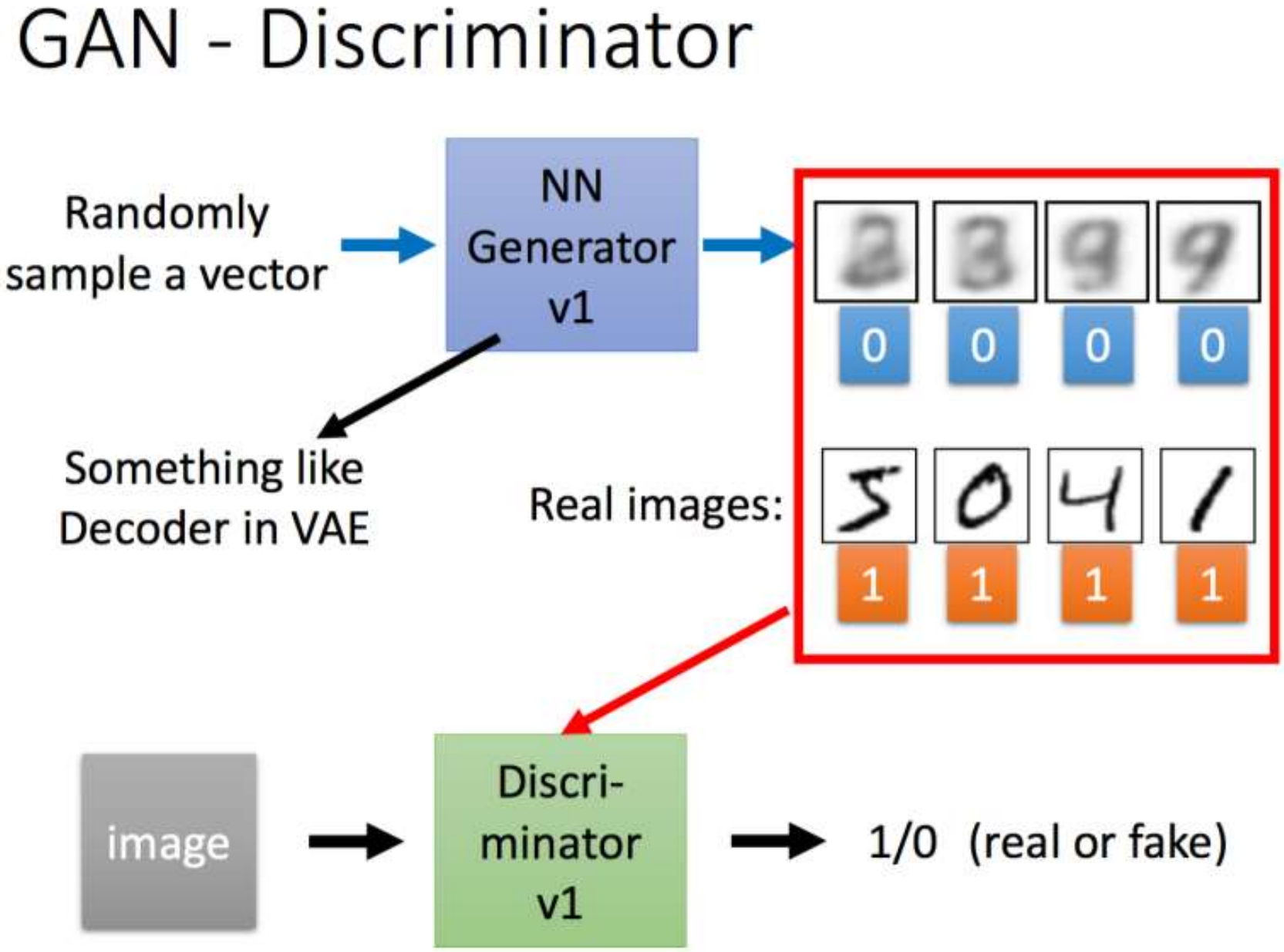


GAN基本概念

- GAN的学习过程



GAN基本概念


GAN - Discriminator
Randomly sample a vector
NN Generator v1
Something like Decoder in VAE
Real images:
0
0
0
0
1
1
1
1
image
Discri-minator v1
1/0 (real or fake)

GAN基本概念

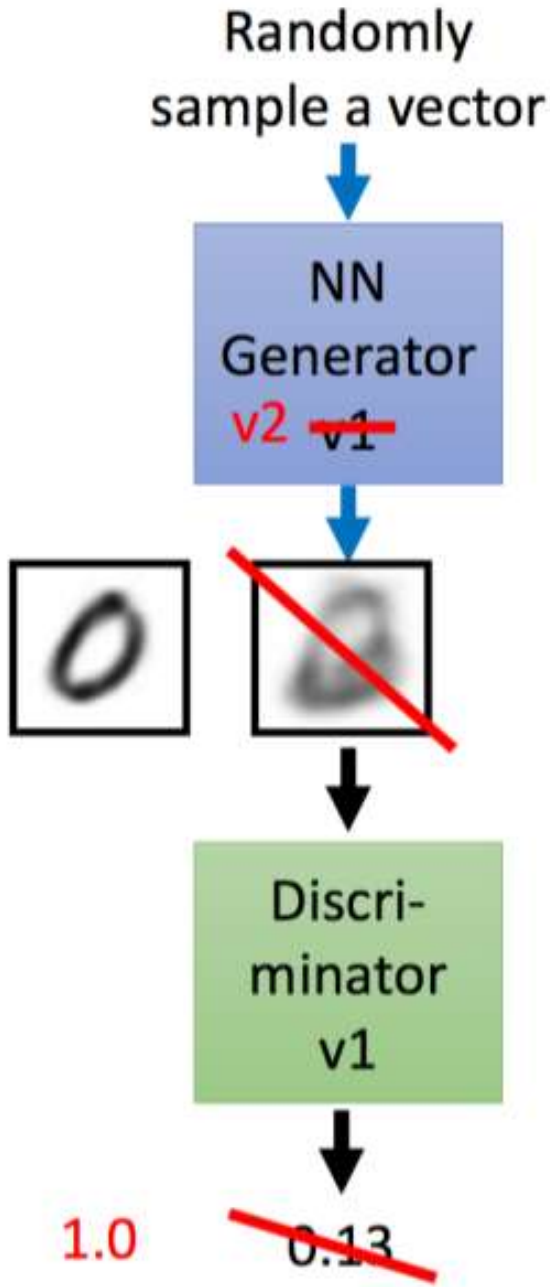
GAN - Generator

Updating the parameters of generator

The output be classified as "real" (as close to 1 as possible)

Generator + Discriminator = a network

Using gradient descent to update the parameters in the generator, but fix the discriminator



GAN原理

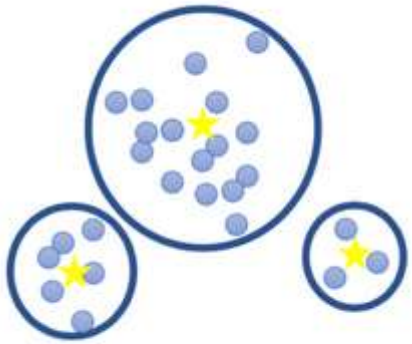
- **Recap: Maximum Likelihood Estimation**
  - Given a data distribution $P_{data}(x)$
  - We have a distribution $P_G(x;\theta)$ parameterized by θ
    - E.g. $P_G(x;\theta)$ is a Gaussian Mixture Model, θ are means and variances of the Gaussians
    - We want to find θ such that $P_G(x;\theta)$ close to $P_{data}(x)$

Sample $\{x^1, x^2, \dots, x^m\}$ from $P_{data}(x)$

We can compute $P_G(x^i;\theta)$

Likelihood of generating the samples

$$L = \prod_{i=1}^{m} P_G(x^i;\theta)$$



Find θ^* maximizing the likelihood

GAN原理

- **Recap: Maximum Likelihood Estimation**

$$\theta^* = arg\max_\theta \prod_{i=1}^{m} P_G(x^i;\theta) = arg\max_\theta log \prod_{i=1}^{m} P_G(x^i;\theta)$$

$$= arg\max_\theta \sum_{i=1}^{m} logP_G(x^i;\theta)$$ $\{x^1, x^2, \ldots, x^m\}$ from $P_{data}(x)$

$$\approx arg\max_\theta E_{x\sim P_{data}}[logP_G(x;\theta)]$$

$$= arg\max_\theta \int_x P_{data}(x)logP_G(x;\theta)dx - \int_x P_{data}(x)logP_{data}(x)dx$$

$$= arg\min_\theta KL(P_{data}(x)||P_G(x;\theta))$$

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

Now $P_G(x;\theta)$ is a NN

$$P_G(x) = \int_z P_{prior}(z)I_{[G(z)=x]}dz$$

It is difficult to compute the likelihood.

GAN原理

- Generator G — Hard to learn by maximum likelihood
  - G is a function, input z, output x
  - Given a prior distribution $P_{prior}(z)$, a probability distribution $P_G(x)$ is defined by function G
- Discriminator D
  - D is a function, input x, output scalar
  - Evaluate the "difference" between $P_G(x)$ and $P_{data}(x)$
- There is a function V(G,D).

求解的最终目标 $G^* = arg \min_G \max_D V(G,D)$

GAN原理

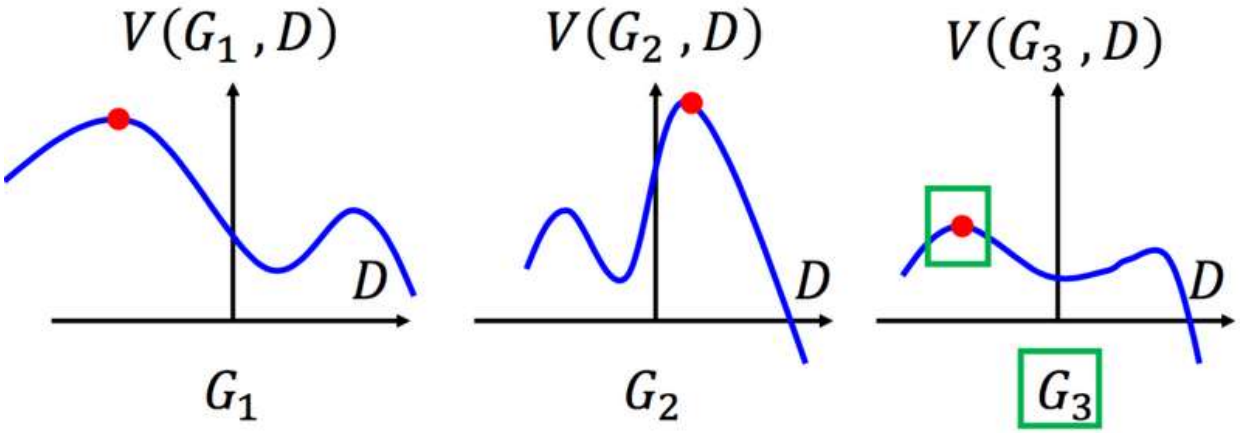
$$G^* = arg \min_G \max_D V(G,D)$$

G希望生成相似真实
min V(G,D)
D希望区别二者
max V(G,D)

$$V = E_{x\sim P_{data}}[logD(x)] + E_{x\sim P_G}[log(1-D(x))]$$

Given a generator G, $\max_D V(G,D)$ evaluate the "difference" between P_G and P_{data}

Pick the G defining P_G most similar to P_{data}



GAN原理

How to define?

Given a generator G, $\max_D V(G,D)$ evaluate the "difference" between P_G and P_{data}

Pick the G defining P_G most similar to P_{data}

$$V = E_{x\sim P_{data}}[logD(x)] + E_{x\sim P_G}[log(1-D(x))]$$

Why this definition?

- Given G, what is the optimal D* maximizing

$$V = E_{x\sim P_{data}}[logD(x)] + E_{x\sim P_G}[log(1-D(x))]$$

$$= \int_x P_{data}(x)logD(x)\,dx + \int_x P_G(x)log(1-D(x))\,dx$$

$$= \int_x [P_{data}(x)logD(x) + P_G(x)log(1-D(x))]\,dx$$

Assume that D(x) can have any value here

- Given x, the optimal D* maximizing

$\max_D V(G,D)$ $\quad G^* = arg\min_G \max_D V(G,D)$

GAN原理

- **Find the optimal D***
  - Given x, the optimal D* maximizing

$$P_{data}(x)logD(x) + P_G(x)log\big(1 - D(x)\big)$$

a D b D

  - Find D* maximizing: $f(D) = alog(D) + blog(1 - D)$

$$\frac{df(D)}{dD} = a \times \frac{1}{D} + b \times \frac{1}{1-D} \times (-1) = 0$$

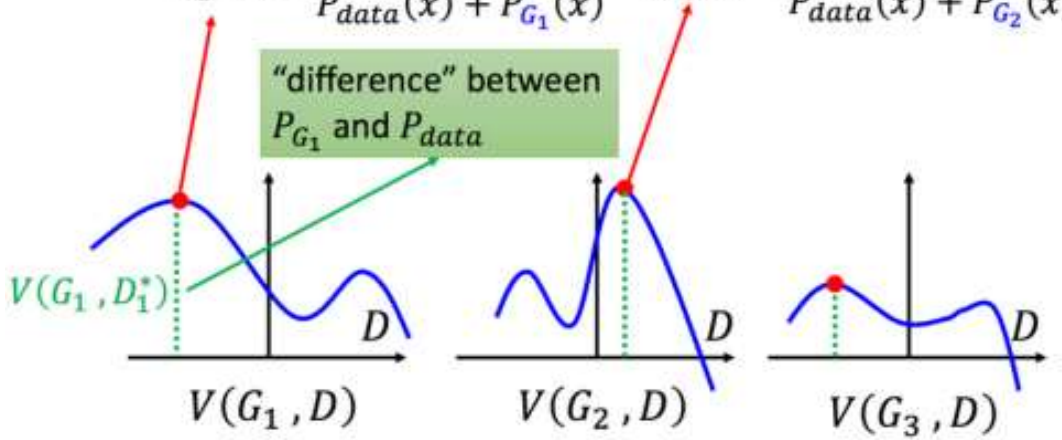
$$a \times \frac{1}{D^*} = b \times \frac{1}{1-D^*} \qquad a \times (1 - D^*) = b \times D^*$$

$$a - aD^* = bD^*$$

$$D^* = \frac{a}{a+b} \Rightarrow D^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_G(x)}$$

$0 <$ < 1

$$D_1^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_{G_1}(x)} \qquad D_2^*(x) = \frac{P_{data}(x)}{P_{data}(x) + P_{G_2}(x)}$$



GAN原理

- **Replace with optimal D***

$$\max_D V(G,D) = V(G,D^*) \qquad D^*(x) = \frac{P_{data}(x)}{P_{data}(x)+P_G(x)} \qquad V = E_{x\sim P_{data}}[logD(x)] + E_{x\sim P_G}[log(1-D(x))]$$

$$= E_{x\sim P_{data}}\left[log\frac{P_{data}(x)}{P_{data}(x)+P_G(x)}\right] + E_{x\sim P_G}\left[log\frac{P_G(x)}{P_{data}(x)+P_G(x)}\right]$$

$$= \int_x P_{data}(x) log\frac{\frac{1}{2}P_{data}(x)}{\frac{P_{data}(x)+P_G(x)}{2}}dx + \int_x P_G(x) log\frac{\frac{1}{2}P_G(x)}{\frac{P_{data}(x)+P_G(x)}{2}}dx$$

$$+2log\frac{1}{2} \quad -2log2$$

$$= -2log2 + \int_x P_{data}(x) log\frac{P_{data}(x)}{(P_{data}(x)+P_G(x))/2}dx + \int_x P_G(x) log\frac{P_G(x)}{(P_{data}(x)+P_G(x))/2}dx$$

$$= -2\log 2 + \mathrm{KL}\left(P_{data}(x)\,||\,\frac{P_{data}(x)+P_G(x)}{2}\right) + \mathrm{KL}\left(P_G(x)\,||\,\frac{P_{data}(x)+P_G(x)}{2}\right)$$

$$= -2log2 + 2JSD(P_{data}(x)||P_G(x))$$ Jensen-Shannon divergence

Objective function $L(G)$ for optimizing G with given D*

$$\mathrm{JSD}(P\,\|\,Q) = \frac{1}{2}D(P\,\|\,M) + \frac{1}{2}D(Q\,\|\,M) \quad M = \frac{1}{2}(P+Q)$$

GAN原理

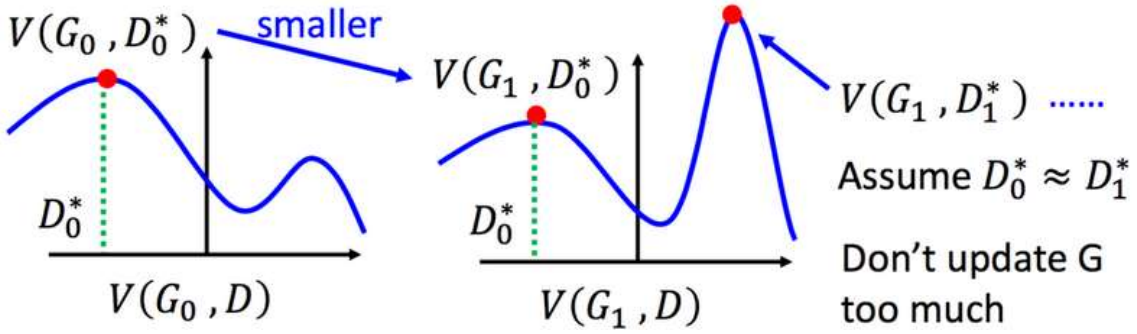
- **Find the optimal G* with gradient decent**
  - Given G_0
  - Find D_0^* maximizing $V(G_0, D)$

    $V(G_0, D_0^*)$ is the JS divergence between $P_{data}(x)$ and $P_{G_0}(x)$

  - $\theta_G \leftarrow \theta_G - \eta \, \partial V(G, D_0^*)/\partial \theta_G$ ➡ Obtain G_1 — Decrease JS divergence(?)
  - Find D_1^* maximizing $V(G_1, D)$

    $V(G_1, D_1^*)$ is the JS divergence between $P_{data}(x)$ and $P_{G_1}(x)$

  - $\theta_G \leftarrow \theta_G - \eta \, \partial V(G, D_1^*)/\partial \theta_G$ ➡ Obtain G_2 — Decrease JS divergence(?)
  - .....



GAN原理

- 在实际中不知道 $P_{data}(x)$ 和 $P_{G_1}(x)$，需要使用采样的方法。

In practice ...

$$V = E_{x\sim P_{data}}[logD(x)] + E_{x\sim P_G}[log(1-D(x))]$$

- Given G, how to compute $\max_D V(G,D)$
  - Sample $\{x^1, x^2, \dots, x^m\}$ from $P_{data}(x)$, sample $\{\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m\}$ from generator $P_G(x)$

Maximize $\tilde{V} = \frac{1}{m}\sum_{i=1}^{m} logD(x^i) + \frac{1}{m}\sum_{i=1}^{m} log\left(1-D(\tilde{x}^i)\right)$

Binary Classifier

Output is D(x) Minimize Cross-entropy

If x is a positive example → Minimize –log D(x)

If x is a negative example → Minimize –log(1-D(x))

GAN原理

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

  for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

  end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

其他改进的GAN算法

- DCGAN

  去除了所有池化层,仅通过卷积层使网络自身学习空间上采样和下采样,并对每一层网络进行批归一化处理。使用Tanh激活函数提高模型学习效率。

- WGAN

  引入Earth-Mover距离来解决采用KL距离和JS距离来刻画真实数据和生成数据分布的相似度时无法产生有效梯度的问题。

- CycleGAN

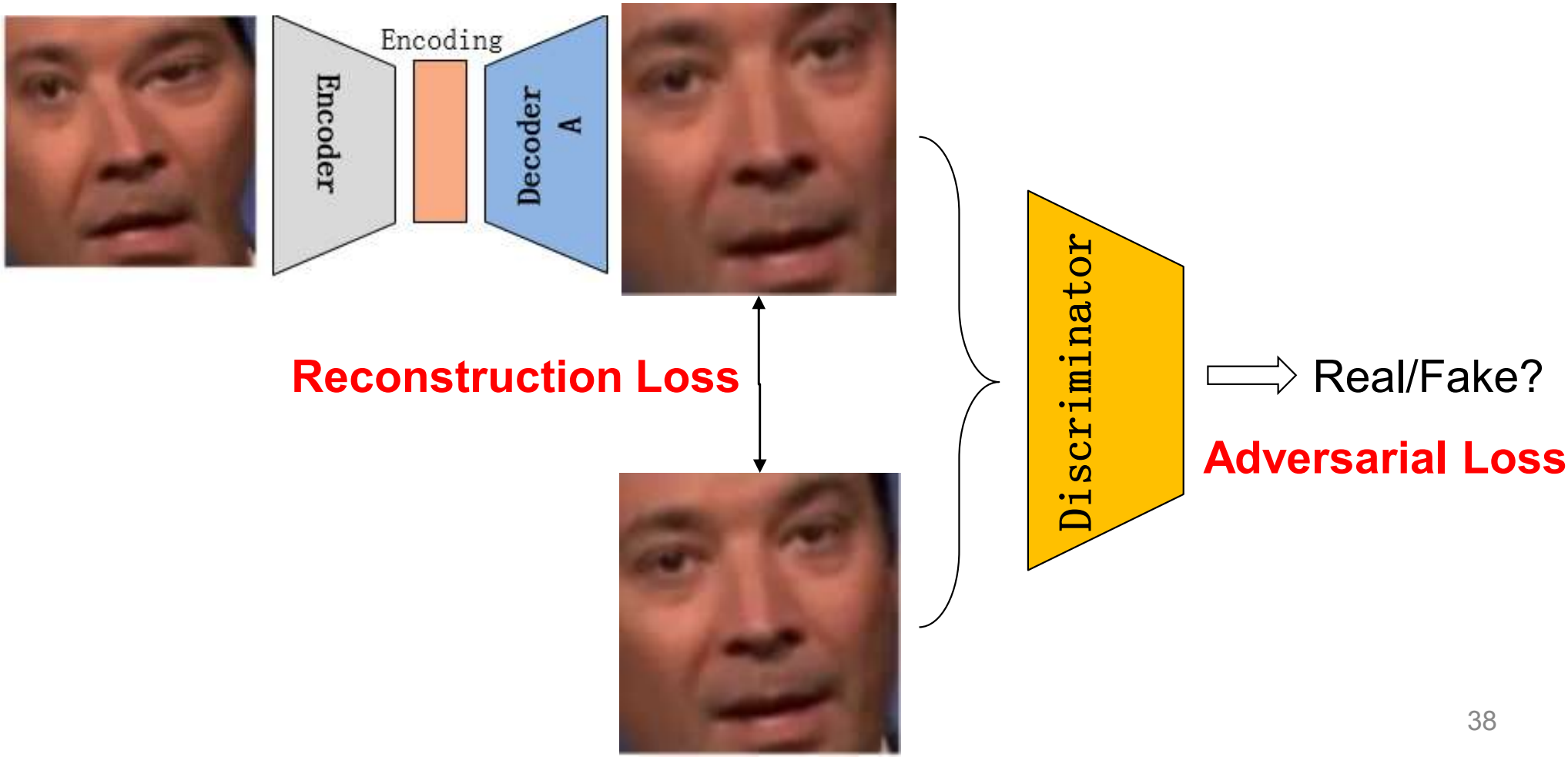
  使用两个不同领域图像的GAN,使各自生成器生成的对方领域的图像尽全力“骗过”对方的判别器。为避免生成器直接从对方领域生成图像的情况, 引入Cycle连续性损失函数。

- StyleGAN

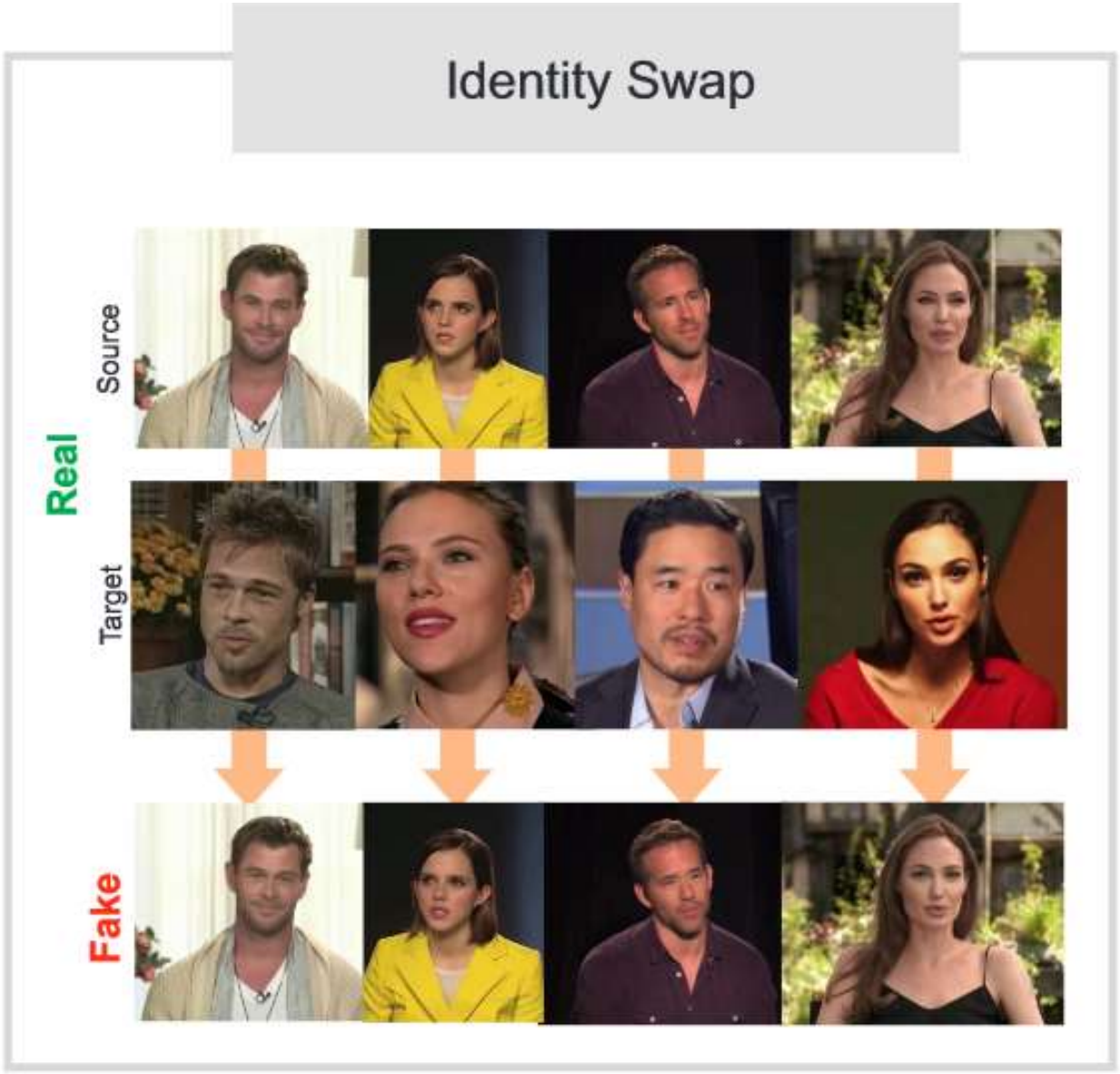
  向生成器的输入加入由 8 个全连接层组成的映射网络,用来生成一个不必遵循训练数据分布的向量,从而降低网络习得的图像特征之间的相关性,减少各类特征之间的干扰。同时,在每次对控制向量卷积后,采用一次自适应实例归一化来添加噪声使生成的图片更逼真。

基于GAN的Deepfakes

- 在基于AE的换脸方法上加入adversarial loss。

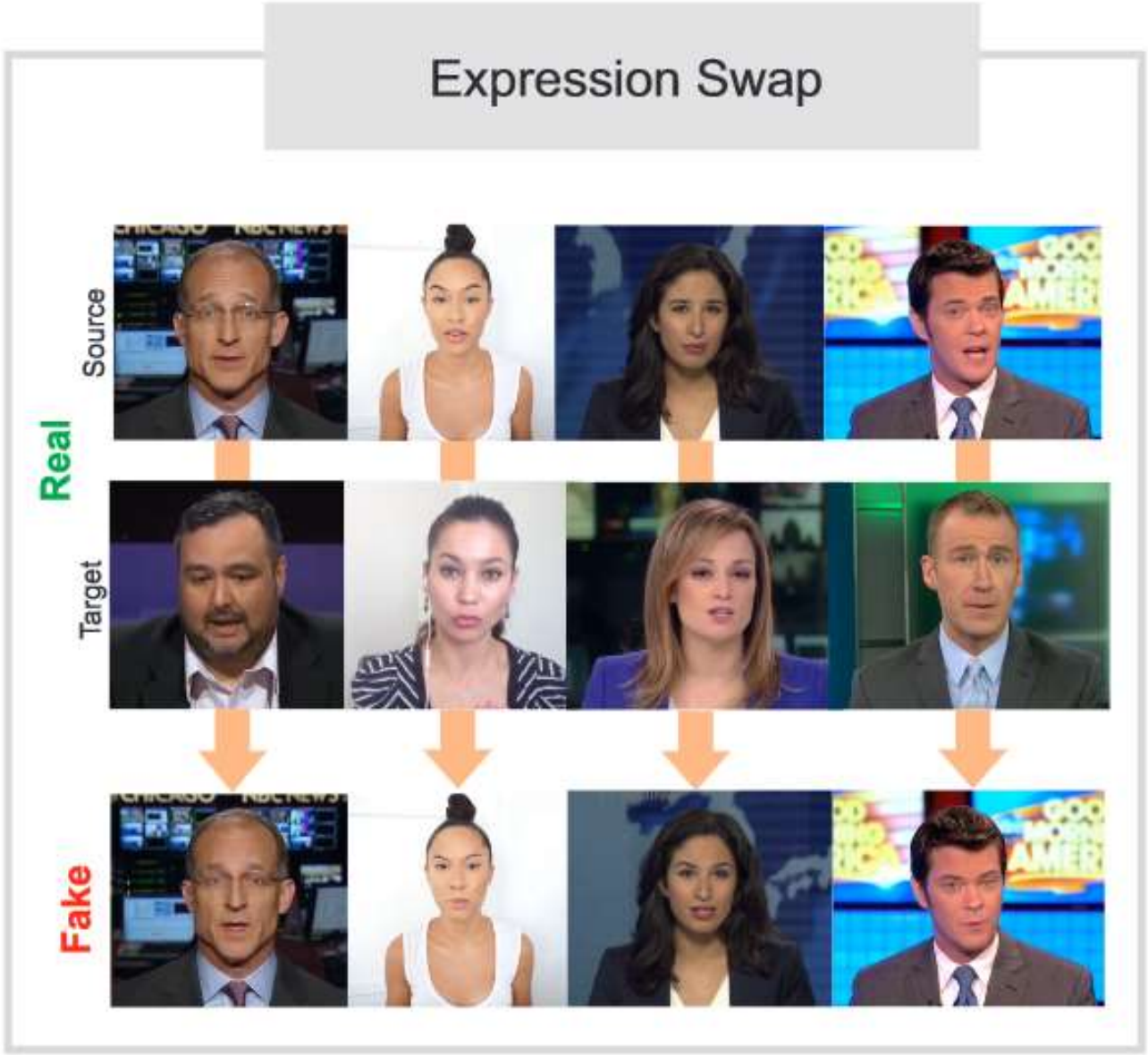


Deepfakes应用效果



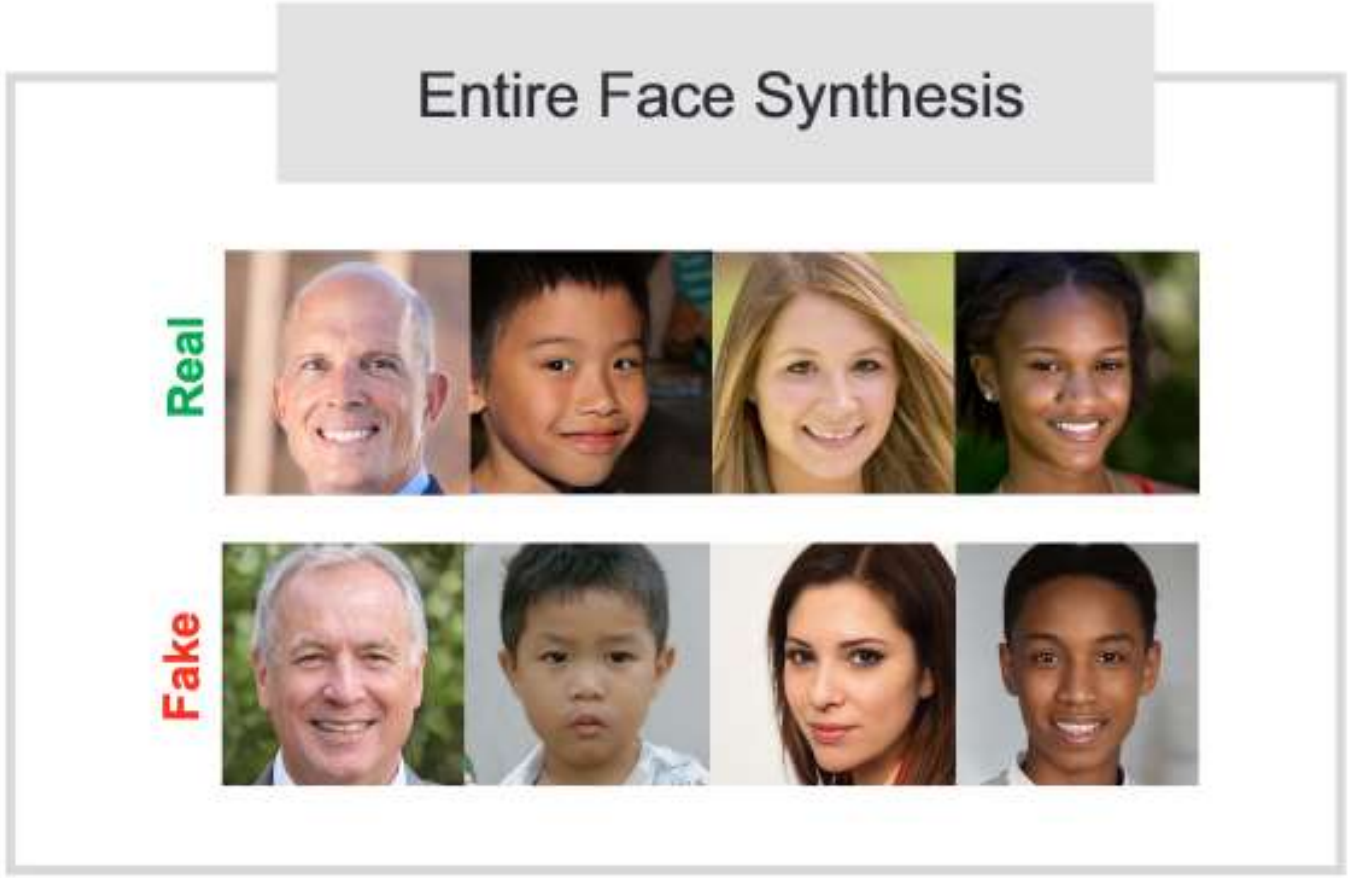
Tolosana, R. et al. “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection.” Inf. Fusion 64 (2020): 131-148.

Deepfakes应用效果



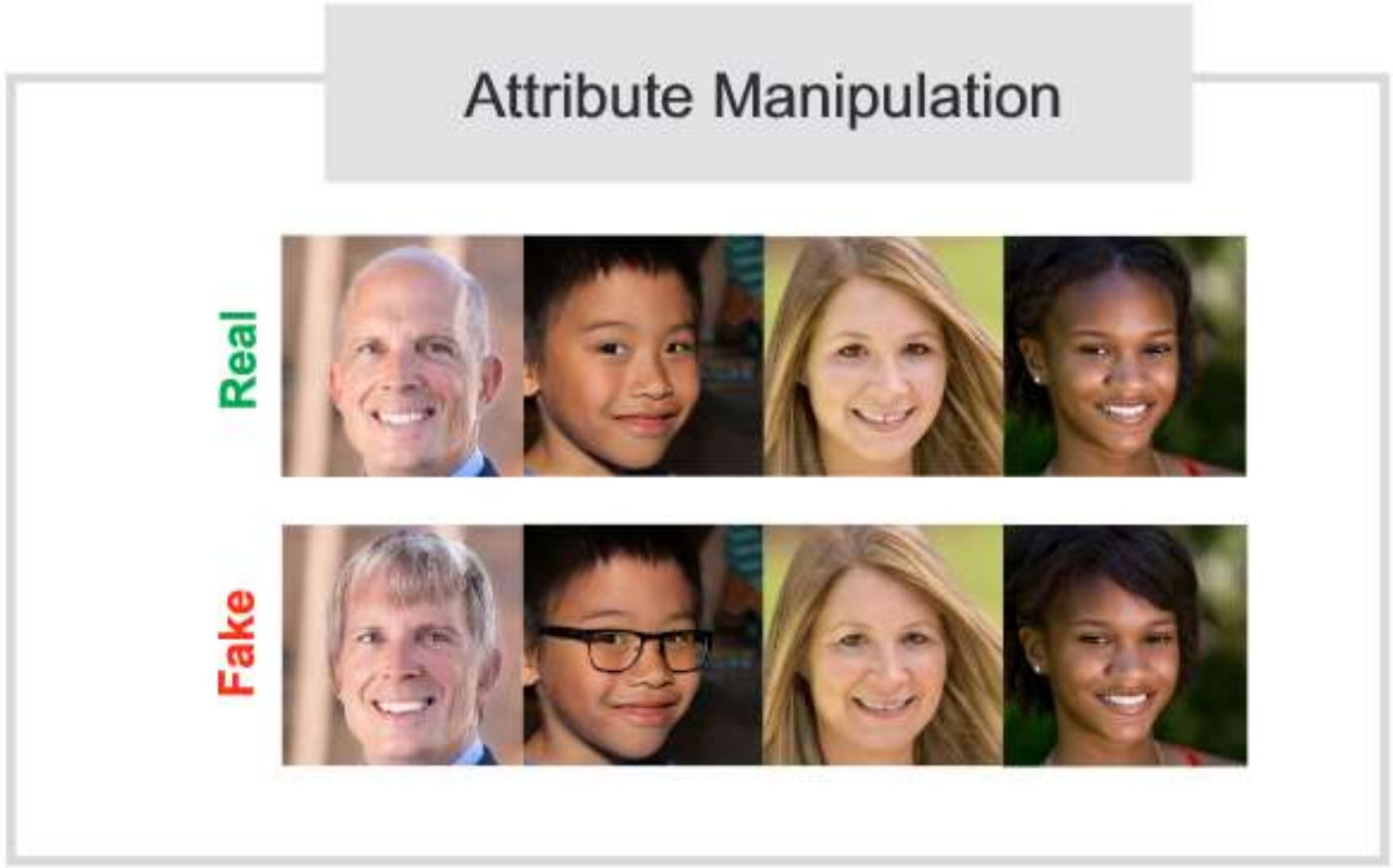
Tolosana, R. et al. “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection.” Inf. Fusion 64 (2020): 131-148.

Deepfakes应用效果



Tolosana, R. et al. “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection.” Inf. Fusion 64 (2020): 131-148.

Deepfakes应用效果



Tolosana, R. et al. “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection.” Inf. Fusion 64 (2020): 131-148.

大 纲

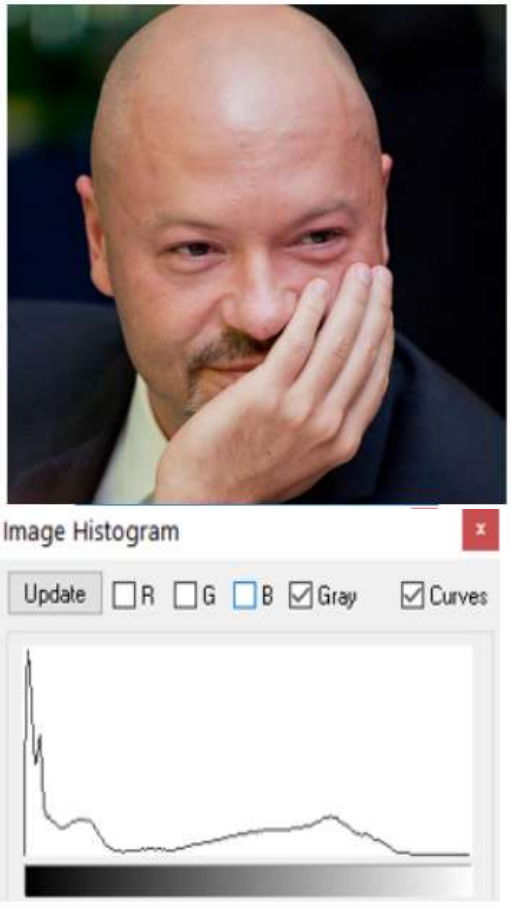
- 深度伪造介绍
- 基于生成模型的深度伪造技术
- **反伪造检测技术**
- 总结与展望

伪造检测技术分类

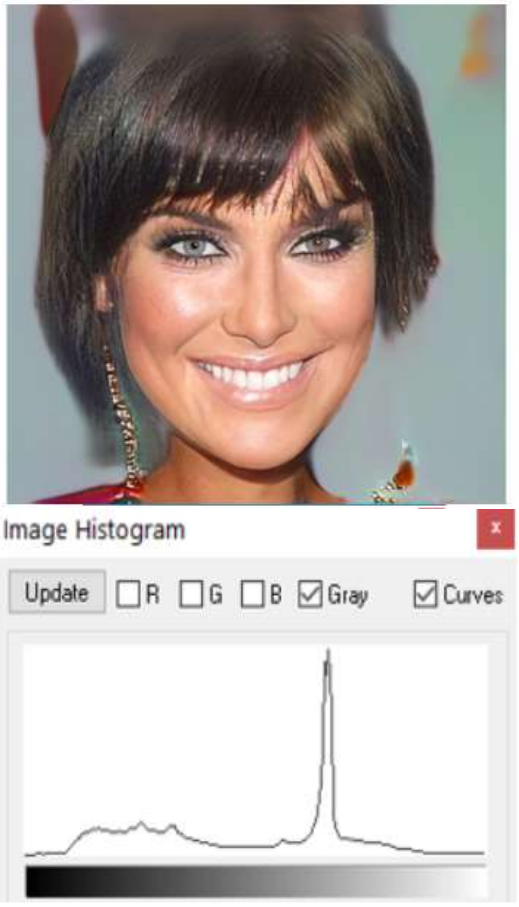
- 基于传统特征的检测
- 基于CNN特征的检测
- 基于真伪对比的检测
- 基于注意力机制的检测
- 基于指纹特征的检测
- 基于跨帧时序特征的视频伪造检测
- 基于仿射变换的视频伪造检测

基于传统特征的检测

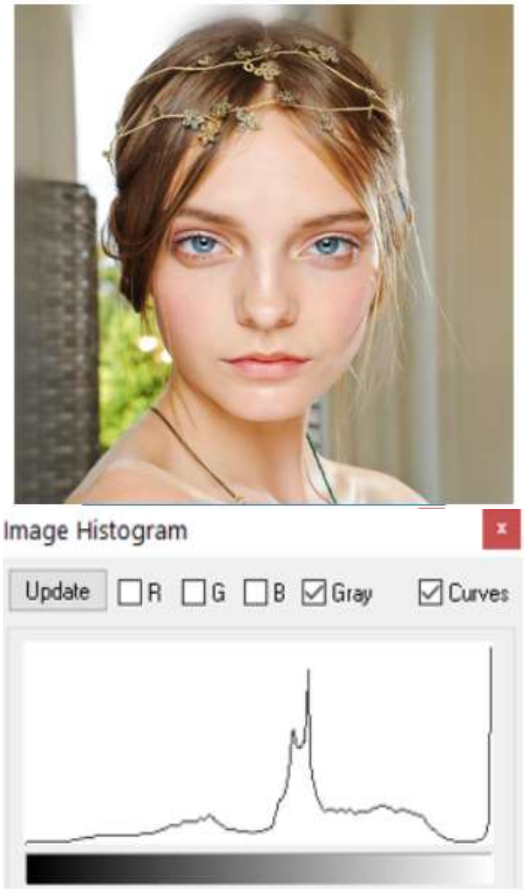
真实图像具有曝光不足或过度的特征，而GAN生成图像即使在背景为白色时也缺乏饱和区域。



真实图像



GAN生成图像

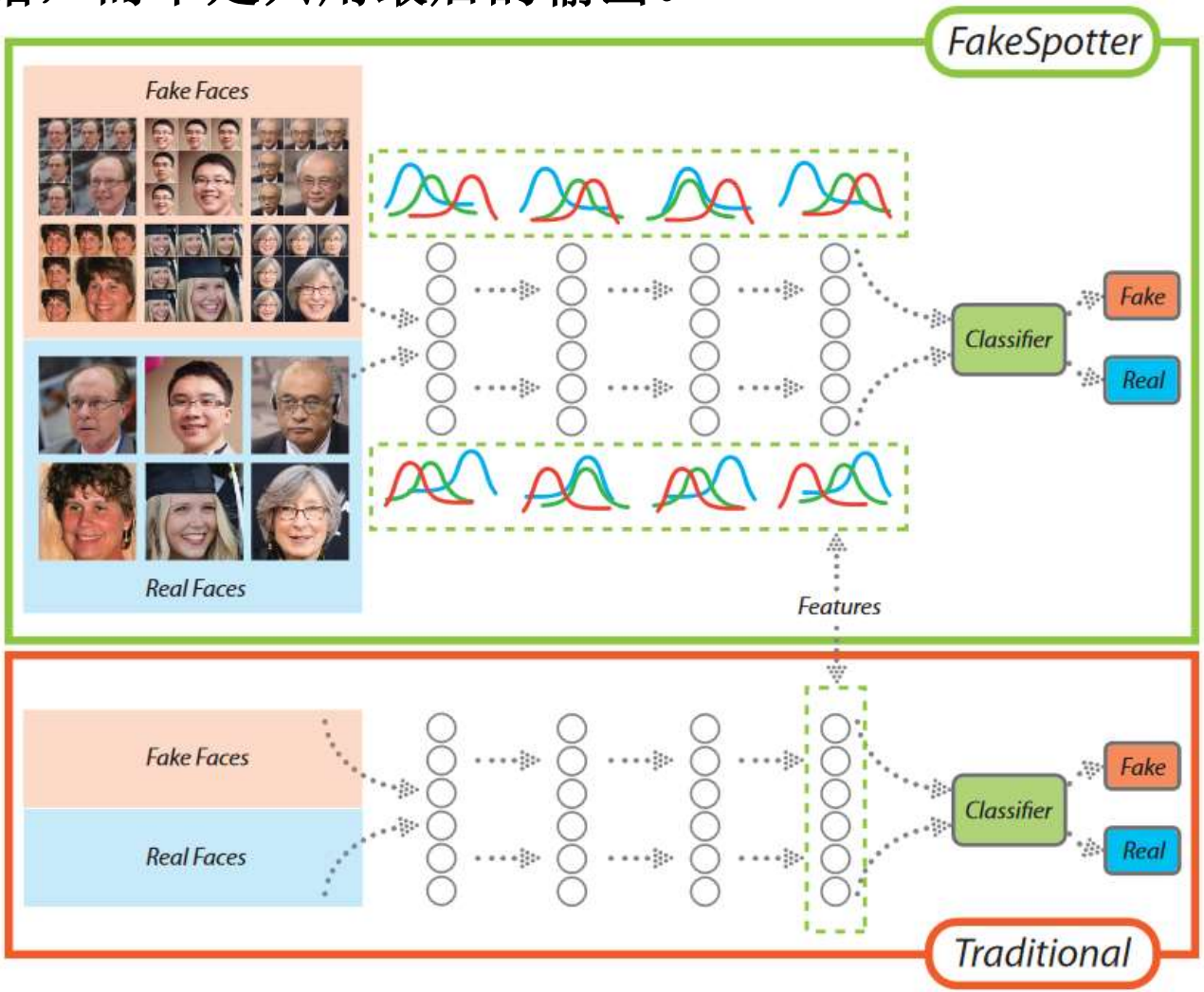


真实图像

McCloskey S, Albright M. Detecting gan-generated imagery using color cues[J]. arXiv preprint arXiv:1812.08247, 2018.

基于CNN特征的检测

针对GAN产生的Fake face，FakeSpotter使用神经网络中间层的激活作为判断依据，而不是只用最后的输出。

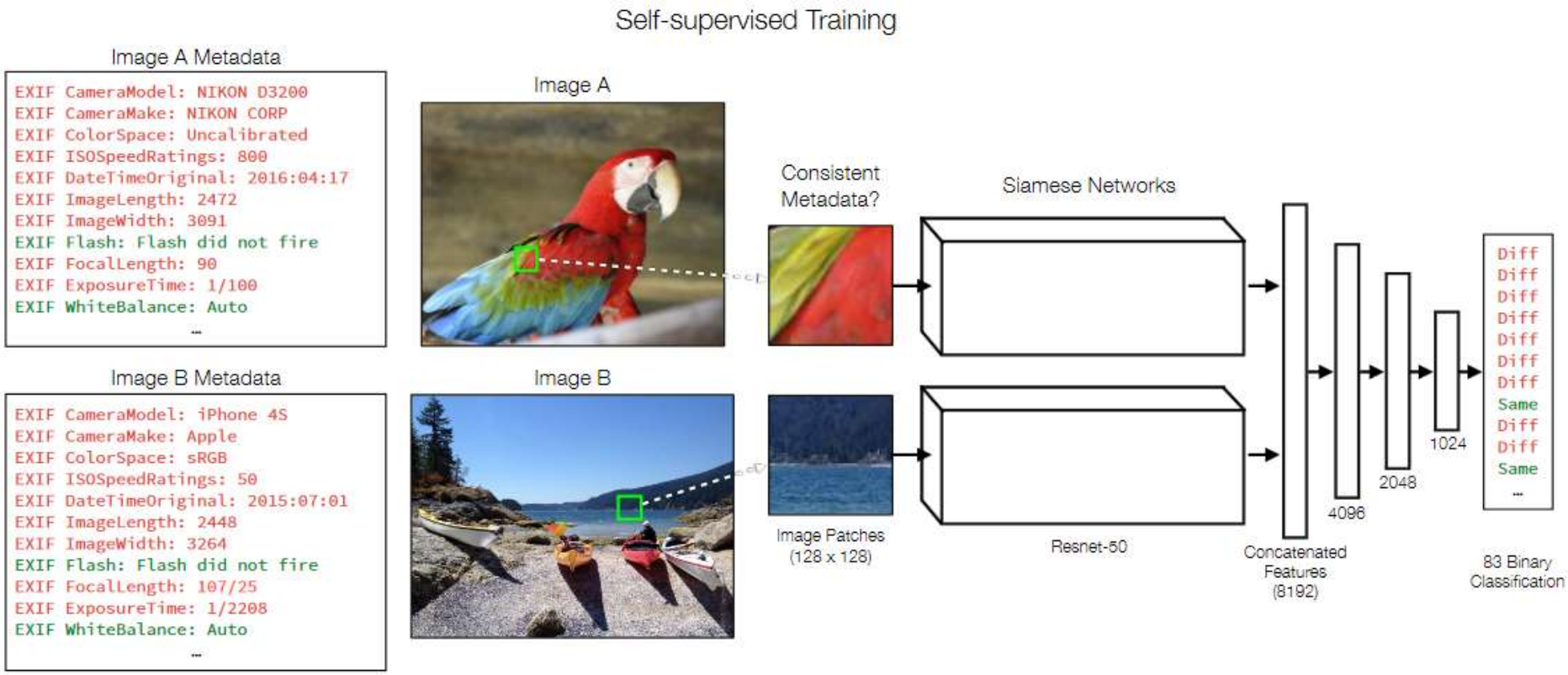


Wang, R., et al. "Fakespotter: A simple yet robust baseline for spotting AI-synthesized fake faces. IJCAI 2020

基于真伪对比的检测

利用图片本身的EXIF信息，训练一个模型判断图片是否是自一致性的（self consistent），即判断图片的每一个部分是否都是由同一个图像处理pipline产生的。

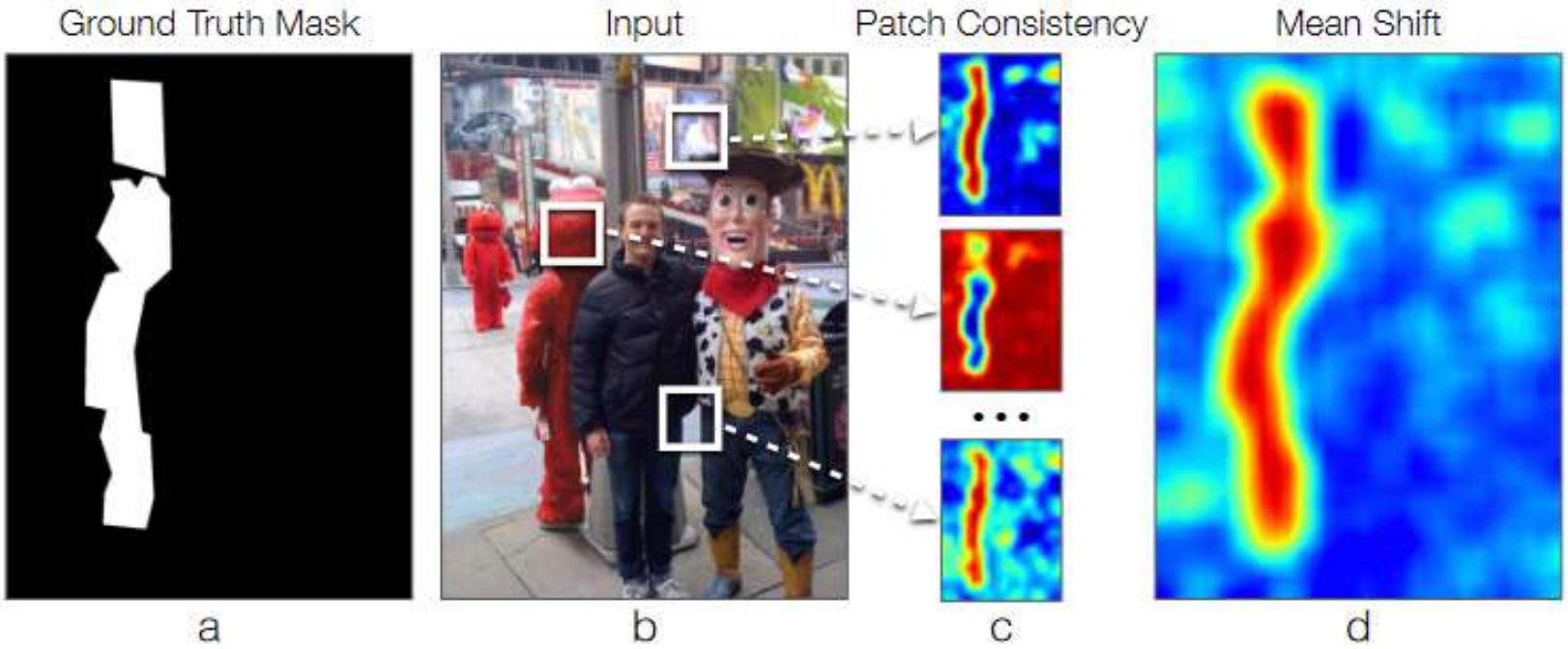
EXIF信息是照相机独有的，在图片成像的时候被嵌入到图片中，包含成像设备、焦距、JPEG质量等。



Huh, Minyoung, et al. "Fighting fake news: Image splice detection via learned self-consistency." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

基于真伪对比的检测

利用图片本身的EXIF信息，训练一个模型判断图片是否是自一致性的（self consistent），即判断图片的每一个部分是否都是由同一个图像处理pipline产生的。

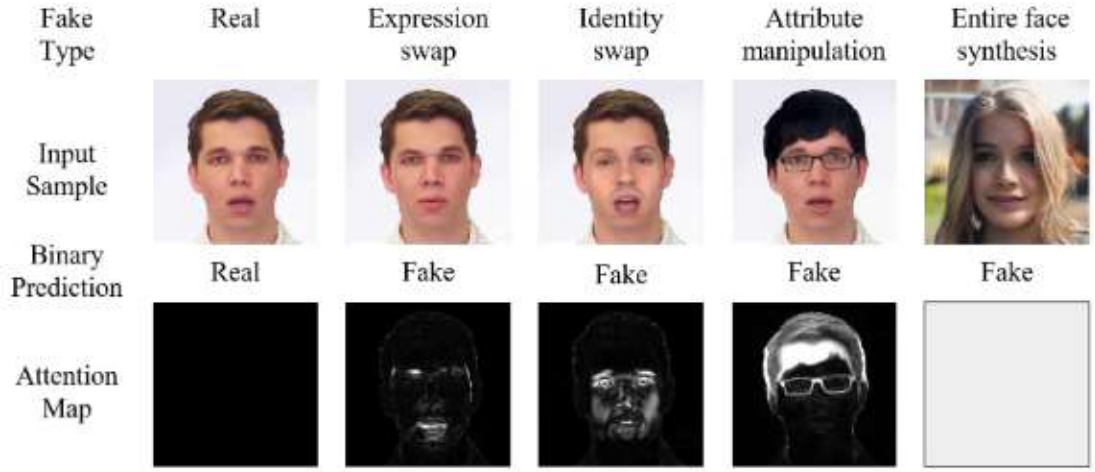


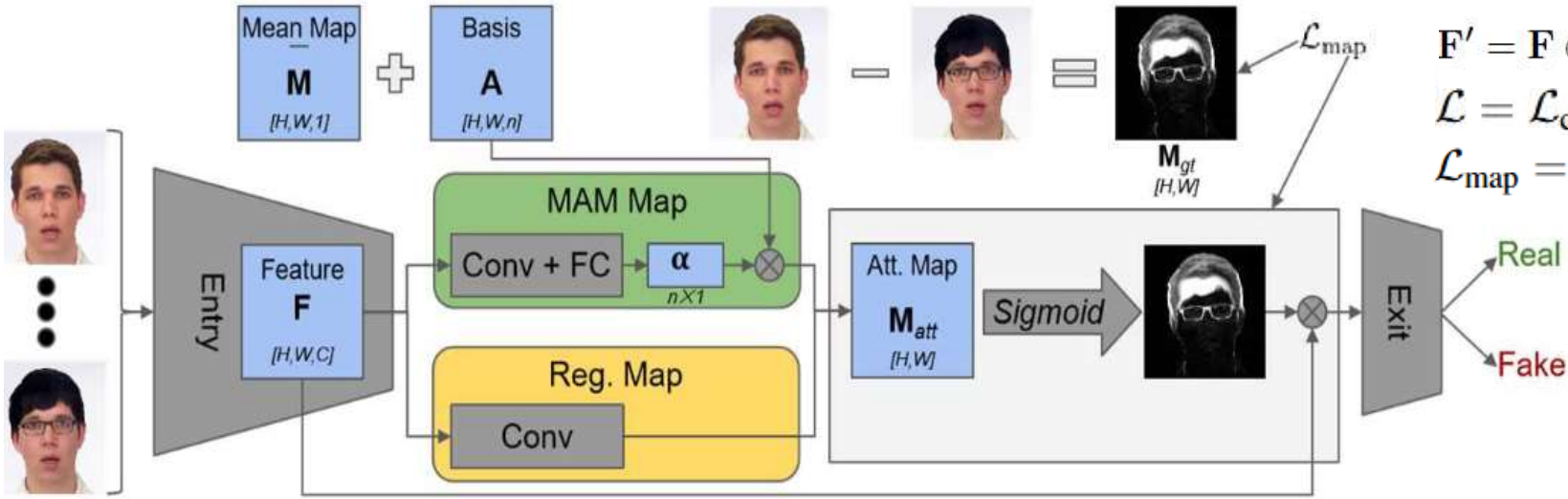
测试时，从输入图像中在网格中采样patch，并估计目标patch和采用patch的一致性，使用Mean Shift将一致性图聚合成最终预测。

Huh, Minyoung, et al. "Fighting fake news: Image splice detection via learned self-consistency." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

基于注意力机制的检测

利用attention机制对图像的伪造区域进行定位，并根据attention map进行伪造分类。

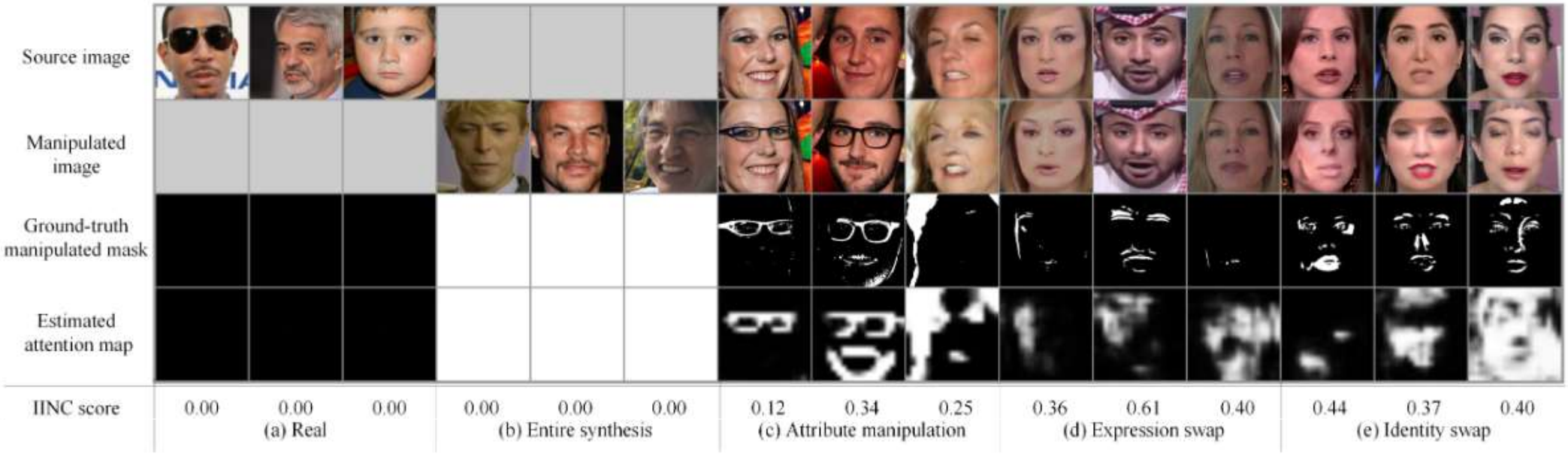




$$\mathbf{M}_{att} = \bar{\mathbf{M}} + \mathbf{A} \cdot \alpha,$$
$$\mathbf{F}' = \mathbf{F} \odot \text{Sigmoid}(\mathbf{M}_{att}),$$
$$\mathcal{L} = \mathcal{L}_{\text{classifier}} + \lambda * \mathcal{L}_{\text{map}},$$
$$\mathcal{L}_{\text{map}} = ||\mathbf{M}_{att} - \mathbf{M}_{gt}||_1,$$

Dang, Hao, et al. "On the detection of digital face manipulation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*. 2020.

基于注意力机制的检测



$$\mathrm{IINC} = \frac{1}{3-|\mathrm{U}|} * \begin{cases} 0 & \text{if } \overline{\mathrm{M_{gt}}} = 0 \text{ and } \overline{\mathrm{M_{att}}} = 0 \\ 1 & \text{if } \overline{\mathrm{M_{gt}}} = 0 \text{ xor } \overline{\mathrm{M_{att}}} = 0 \\ (2 - \frac{|\mathrm{I}|}{|\mathrm{M_{att}}|} - \frac{|\mathrm{I}|}{|\mathrm{M_{gt}}|}) & \text{otherwise,} \end{cases}$$

where I and U are the intersection and union between the ground truth map.

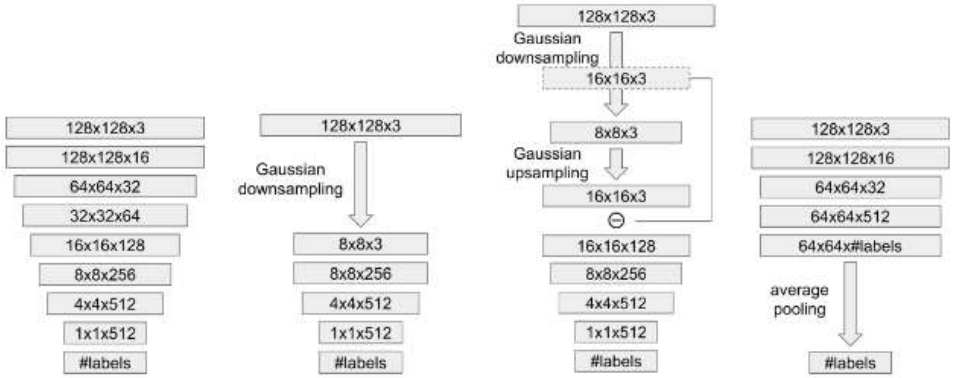
Dang, Hao, et al. "On the detection of digital face manipulation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*. 2020.

基于指纹特征的检测

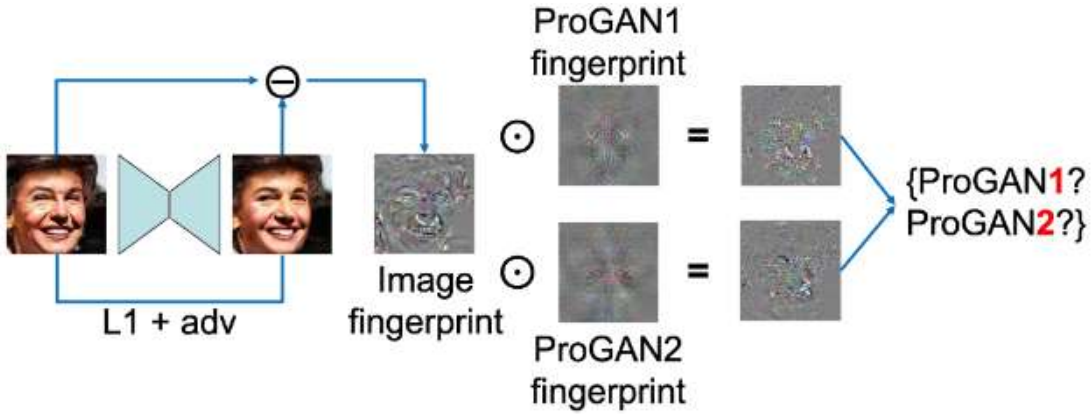
每个GAN模型受训练设置（数据、网络结构、损失函数、参数等）的影响，都会有其独特的指纹，这种指纹会影响其生成的图像，造成独特的图像指纹。因此可以将GAN模型指纹和图像指纹进行匹配。

提取指纹的方法：

- **隐式：**用CNN提取图像特征向量，作为要验证的图像指纹，然后输入最后的分类器，将GAN模型指纹作为该分类器的参数。



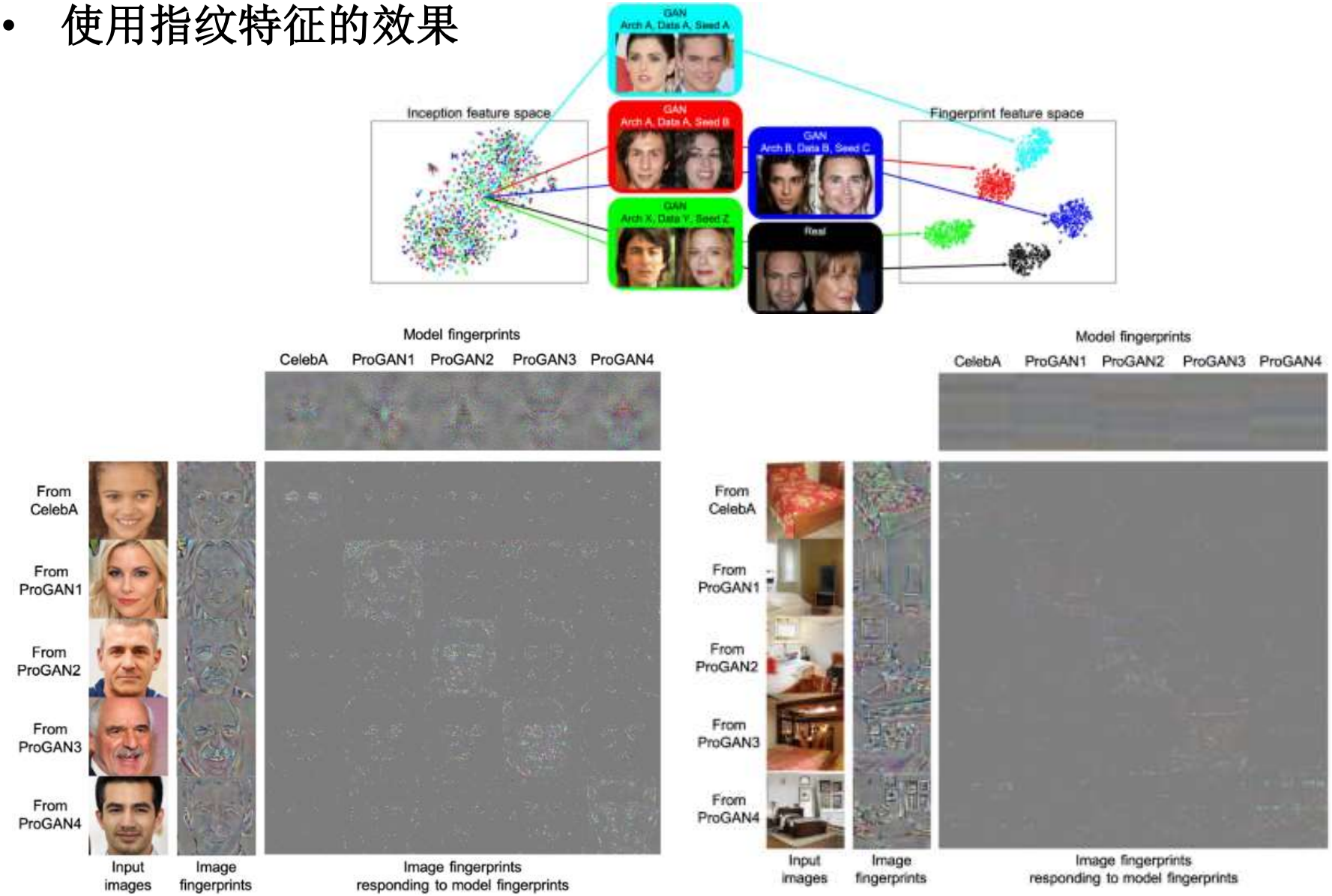
- **显式：**使用Autoencoder对原图像进行重建，然后与原图提取残差，将此残差作为图像指纹。然后将图像指纹与模型指纹逐像素相乘。



Yu, Ning, Larry Davis, and Mario Fritz. "Attributing fake images to gans: Analyzing fingerprints in generated images." ICCV 2019.

基于指纹特征的检测

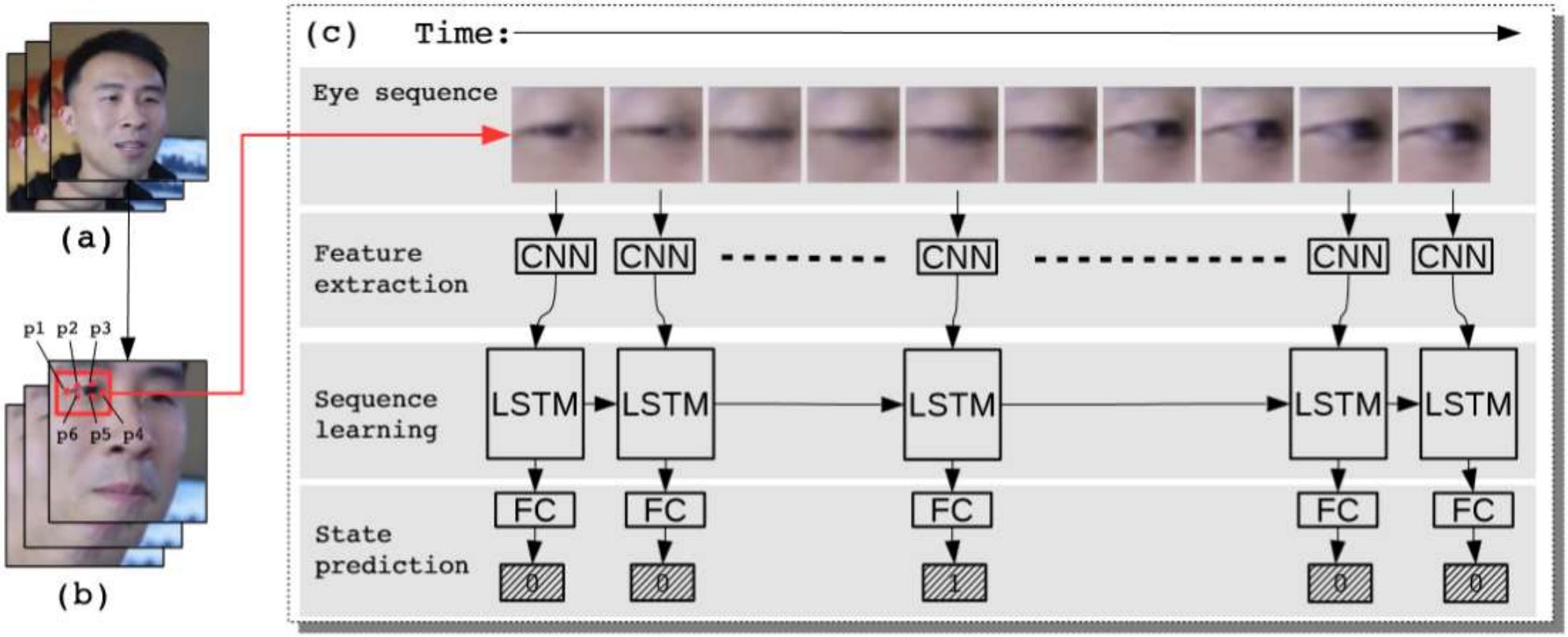
- 使用指纹特征的效果



Yu, Ning, Larry Davis, and Mario Fritz. "Attributing fake images to gans: Analyzing fingerprints in generated images." ICCV 2019.

基于跨帧时序特征的视频伪造检测

深度伪造模型通常使用静态的面部图像集进行训练，无法实现对眨眼、呼吸和心跳等生理信息的准确伪造，故可以基于生理信息的合理性来构建检测方法。

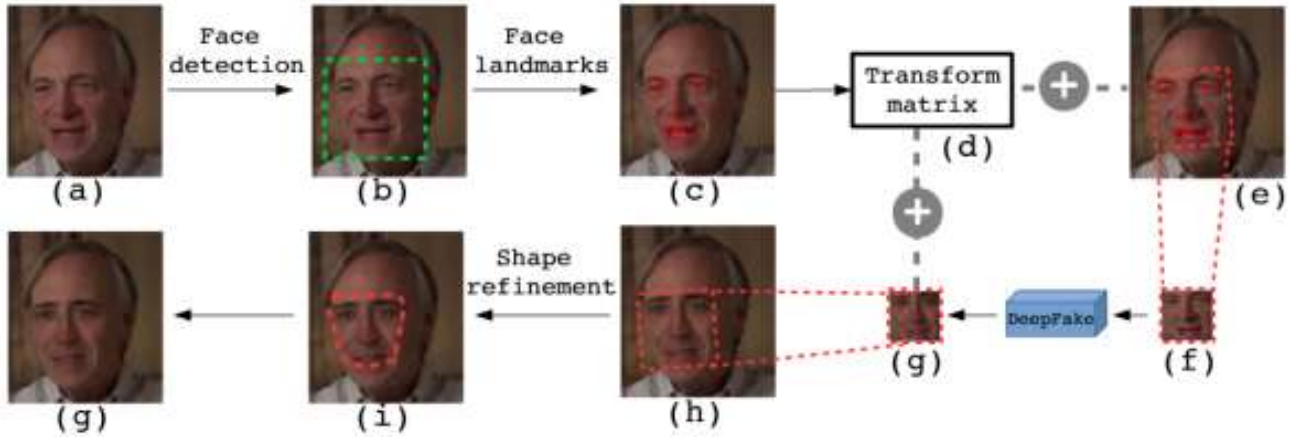


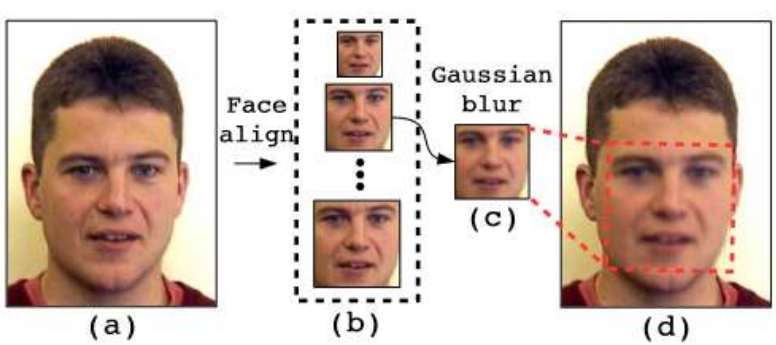
LIY, CHANG MC, and L. Y. U. S. InIctuOculi. "ExposingAICreated FakeVideosbyDetectingEyeBlinking." *2018IEEEInterG national Workshop on Information Forensics and Security (WIFS). IEEE*. 2018.

基于仿射变换的视频伪造检测

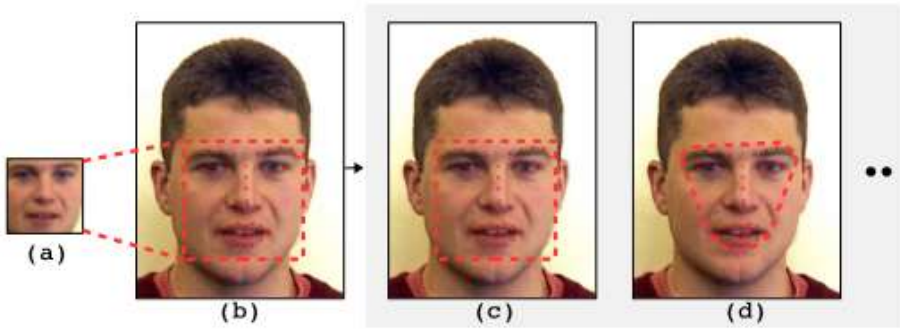
Deepfakes只能合成有限分辨率的人脸图像，并且必须对其进行仿射变换以匹配源人脸配置。由于扭曲面区域和周围环境的分辨率不一致，导致在生成的视频中留下了独特的伪影（artifacts）。通过训练一个神经网络来捕获伪影即可进行检测。

Deepfakes
算法流程





构建负样本



增强的负样本

Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts[J]. arXiv preprint arXiv:1811.00656, 2018.

大 纲

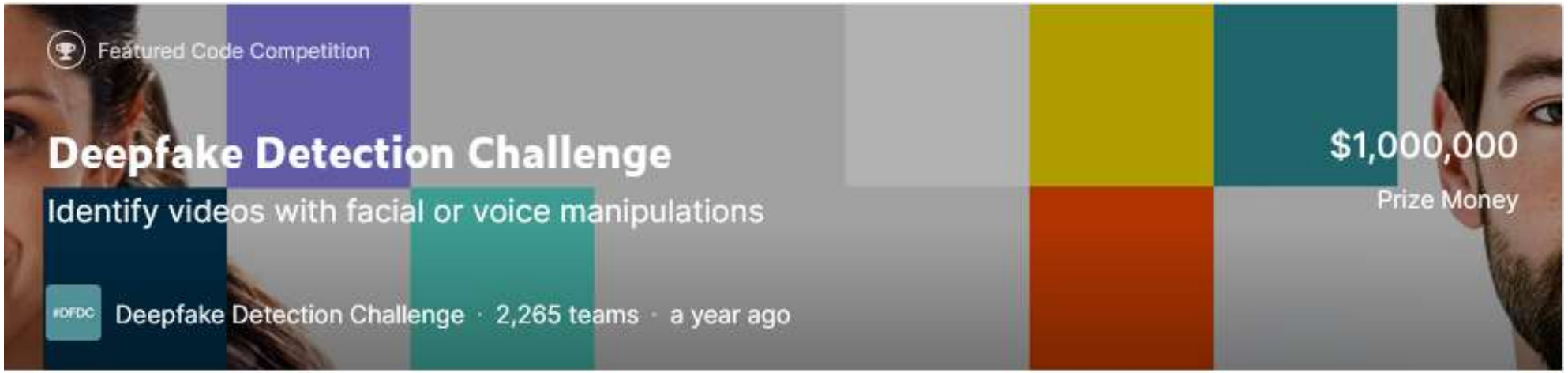
- 深度伪造介绍
- 基于生成模型的深度伪造技术
- 反伪造检测技术
- **总结与展望**

总结与展望

- Deepfakes技术的逼真效果主要得益于深度生成模型尤其是GAN的发展。
- Deepfakes的恶意应用催生了深度伪造检测技术的发展：
  - 基于传统特征的检测
  - 基于CNN特征的检测
  - 基于真伪对比的检测
  - 基于注意力机制的检测
  - 基于指纹特征的检测
  - ……
- 深度伪造检测领域未来研究方向：
  - 研究泛化性好的检测算法
  - 研究鲁棒性强的检测算法
  - 研究主动防御算法
  - 研究深度伪造图像和伪造语音的融合检测技术

总结与展望

深度伪造图像/视频的检测研究持续受到包括学界、工业界、政府、用户等社会各界的关注，并且依旧具有挑战性。



https://www.kaggle.com/c/deepfake-detection-challenge/overview

Thanks!

Q&A