Deepfake overview & Reinforcement learning based deepfake fusion and methods

Xinchen Zhang

IPIU, XiDian University

1 Introduction

In recent years, with the rapid development of science and technology and the rapid increase of computing power, the theoretical technology of artificial intelligence has been matured and has been widely used in the fields of finance, medical care, urban services, industrial manufacturing and life services, etc. Artificial intelligence technology is leading a new round of all-round industrial change and driving the human world into the era of intelligence. As a popular research direction in machine learning, deep learning provides powerful technical support for innovations in computer vision, driverless, natural language processing and speech recognition. However, while deep learning technology is leading a new wave of artificial intelligence, it also poses potential threats to personal privacy data, social stability and national security. Since 2017, voice fraud supported by "deep forgery" technology has attracted widespread attention worldwide, and "face-swapping" videos spoofing political figures and public figures have also emerged, causing very negative impacts and even indirectly leading to the military coup in Gabon .

In December 2017, a user named "Deepfakes" created a stir by posting a video of a fake face of Hollywood actress Gal Gadot on Reddit, the fourth most trafficked international Internet community in the world, which started the rise of deep face forgery technology, and the user's username was cited as a synonym for this type of technology " Deepfake". Thus, Deepfake refers to deep forgery of faces, where the face of a target video person is replaced with a specified original video face, or the target face is made to reenact or mimic the actions and expressions of the original face, thus creating a forged video of the target face.

Deepfake can be used to create virtual characters, video rendering, and voice simulation in film production, to "revive" historical figures or deceased friends and relatives, and to enable "face-to-face" communication, creating a new form of communication. Deep forgery techniques can also be used to mislead public opinion, disrupt social order, and may even threaten face recognition systems, interfere with government elections, and subvert state power, which have become the most advanced new forms of cyber attacks. The number of "one-click" content synthesis (image, video, and voice) applications such as FakeApp and Faceswap are common, and even the "one-click" smart stripping software Deepnude appeared in June 2019. Although the software was taken down by the developer after its release, it still caused great fear worldwide. The damage and impact of deep forgery has spread around the world, and the detection and defense of deep forgery has become one of the hot issues for governments, enterprises and individuals around the world.

2 Relate works

2.1 Deepfake

Five variants or combinations of neural networks are commonly used to build deep forgery generative networks: Encoder-Decoder (ED), Convolutional neural network (CNN), Generative adversarial networks (GAN), Image transformation network (Pix2Pix, CycleGAN), and Recurrent neural network (RNN).

2.1.1 Encoder-Decoder

Autoencoder is a type of neural network, and the basic idea is to map the input data directly using one or more layers of the neural network to obtain the output vector as the features extracted from the input data. The basic autoencoder model is a simple three-layer neural network structure: an input layer, a hidden layer and an output layer. The output layer has the same number of dimensions as the input layer. Autoencoder is essentially a data compression algorithm in which the data compression and decompression functions are data-dependent, lossy, and automatically learned from samples. The main uses of autoencoders today are dimensionality reduction, denoising, and image generation.

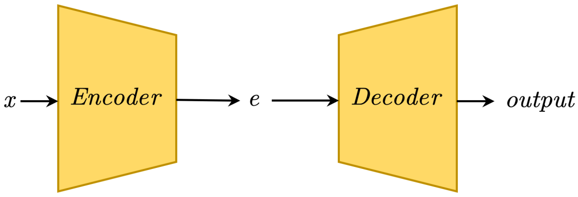

Figure 1: Encoder-Decoder

Deepfake techniques typically use multiple encoders or decoders and manipulate the encoding to affect the output x_g. If the encoders and decoders are symmetric and train the network with the target:

$$De(En(x)) = x \tag{1}$$

the network is called Autoencoders and the output is a reconstruction of x.

The autoencoders extract the latent features of the facial image and reconstruct the facial image using decoders. In order to exchange the face between the source and target images, two encoder/decoder pairs are required.

Two pairs of encoder networks with the same construction, each trained on the image set, and the encoder parameters are shared between the two network pairs.

2.1.2 GAN

Generating adversarial networks is a method of unsupervised learning, by having two neural networks play each other. The method was proposed by Ian Goodfellow et al. in 2014. A generative adversarial network consists of a generative network and a discriminative network. The generative network takes random samples from the potential space as input, and its output needs to mimic the real samples in the training set as much as possible. The discriminative network's input is the real samples or the output of the generative network, and its purpose is to distinguish the output of the generative network from the real samples as much as possible. The generative network, on the other hand, tries to deceive the discriminator network as much as possible. The two networks confront each other and continuously adjust their parameters, with the ultimate goal of making the discriminant network unable to determine whether the output of the generative network is real or not.

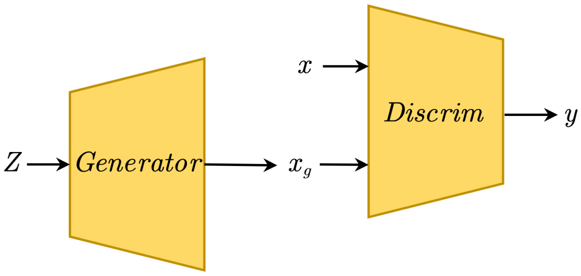

Figure 2: GAN

G aims to trick D to create pseudo-samples, which randomly generate samples of observed data by giving some implicit information; D learns to distinguish between real samples $(x \in X)$ and pseudo-samples$(x_g = G(z), z \sim N)$, and it predicts whether a data sample belongs to the real training sample or not.

Specifically, there is an adversarial loss for training D and G, respectively:

$$L_{adv}(D) = \max\ logD(x) + log(D(G(z))) \tag{2}$$

$$L_{adv}(G) = \min\ log(D(G(z))) \tag{3}$$

Under the adversarial game, both improve their ability by adversarial training, and G learns how to generate samples that are indistinguishable from the original distribution. After training, D is discarded and G is used to generate the content.

Ideally, the generative model would be able to generate data samples that are "false" enough, while the discriminative model would have difficulty in discerning their authenticity, i.e., the probability of being correct is only 50%.

The advantage of GAN is that it does not rely on a priori knowledge, and the parameter updates of the generative model come from the back propagation of the discriminant model rather than directly from the data samples, so the training does not require a complex Markov chain. When applied to images, higher quality and more realistic image samples can usually be generated.

2.1.3 RNN

RNN are neural networks that can handle sequences and variable length data, and in Deepfake productions, RNN are usually used to process audio and video.

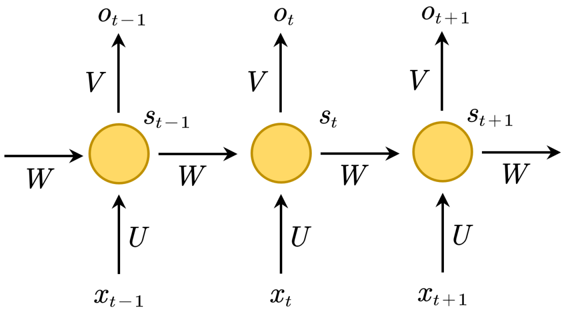

Figure 3: RNN

2.1.4 First Order Motion Model

The task of this method is the task is image animation, the input is a source image and a driving video, the output is a video, where the main character is the source image and the action is the action in the driving video. As shown below, the source image usually contains a subject and the driving video contains a series of actions.

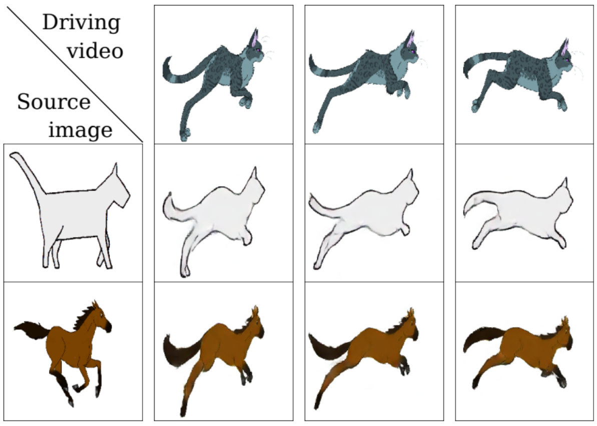

Figure 4: Result of this algorithm

The first column is a given picture, while the first row of images is a given action sequence, which is made to complete a series of actions by face and expression migration, respectively. In other words, the extracted action features are used as the action basis of the images

The authors used a self-monitoring strategy derived from Monkey Net. For training, the authors used a large number of video sequences containing objects of the same object class. The model reconstructs the training videos by combining single frames and learning the underlying representations of the actions. By looking at pairs of frames extracted from the same video, it learns to encode actions as a combination of action-specific keypoint displacements and local affine transformations. During testing, the model was applied to a pair consisting of each frame of the source image and the driving video, and the corresponding image animation was generated based on the source object.

The entire model works as follows:

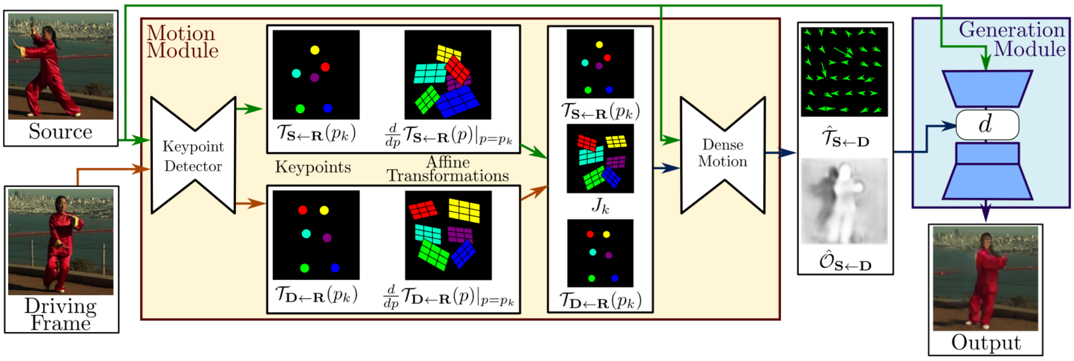
Source
Driving Frame
Motion Module
Keypoint Detector
$\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p_k)$
$\frac{d}{dp}\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p)|_{p=p_k}$
Keypoints
Affine Transformations
$\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p_k)$
$\frac{d}{dp}\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p)|_{p=p_k}$
$\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p_k)$
J_k
$\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p_k)$
Dense Motion
$\hat{\mathcal{T}}_{\mathbf{S}\leftarrow\mathbf{D}}$
$\hat{\mathcal{O}}_{\mathbf{S}\leftarrow\mathbf{D}}$
Generation Module
d
Output

Figure 5: Network of this model

The whole model is divided into two main components, the motion estimation module and the image generation module. In the motion estimation module, the model separates the appearance and motion information of the target object by self-supervised learning and performs feature representation. And in the image generation module, the model models the occlusions that occur during the target motion, and then extracts the appearance information from the given celebrity image and performs video synthesis by combining the previously obtained feature representations.

- Motion estimation module

  there are two outputs:

  - Dense motion field: characterizes the mapping relationship of each key point in the driving image D to the source image S
  - Mapping mask: It shows that in the final generated image, for the driving image D, which part of the pose can be obtained by S distortion and which part can only be obtained by impainting. There is a large deformation from S to D, and the direct mapping will result in a large error. The technique used is to propose a transition frame R, and first establish the mapping from R frame to S frame, R frame to D frame, and then establish the mapping from D frame to S frame.

- Image generation module

  An image generation model that generates a new image based on the input image and the information obtained in the first part

2.2 Anti-Deepfake

Anti-deepfake refers to the Deepfake face video defense technology. According to the different defense strategies, the existing defense technologies can be broadly divided into two categories: passive detection and active defense. Another type of active defense technology focuses on ex ante defense, i.e., adding hidden information, such as watermark and anti-noise, to actively trace the source or make it impossible for malicious users to forge face videos with added noise, so as to achieve the purpose of protecting faces and realizing active defense.

Passive detection techniques refer to techniques that identify forged face videos by obtaining information or extracting features only from the face video itself, a task that is essentially a binary classification task.

The core idea of active defense is to add a certain degree of signal interference to the face video before it is released, so that malicious users cannot use the interfered face material for forgery, or even forgery can be successfully traced back to the forger, in order to achieve the goal of "pre-defense". Active defense techniques can also be categorized as active interference and active forensics. Active interference refers to adding noise interference to the published face material, such as counterattack or data poisoning, so that malicious users cannot forge the face. Active forensics refers to embedding a certain identifier in the training data so that such identifier can be detected in the generated forgery results.

2.2.1 ID-Reveal

The method introduces ID-Display, a new approach that learns temporary facial features through metric learning and adversarial training strategies to specifically describe how a person moves while speaking. The advantage is that we do not need any fake training data, but only training on real videos. Moreover, the algorithm utilizes high-level semantic features, which makes it robust to widespread and destructive forms of post-processing.

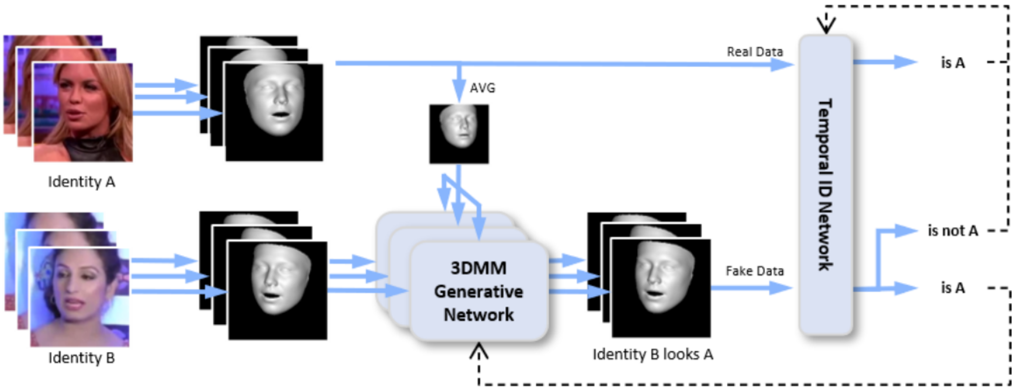

Figure 6: Network of this model

The network consists of three main structures:

- Feature extraction

  The input is a video, and facial features are extracted for each frame in the video. Then a 3D morphological model is used to map each face into a low-dimensional representation (i.e., the morphological map of the face shown in the figure). This representation contains information about the shape, expression, and appearance of the face. The next step is to retrieve these information parameters of the face from this low-dimensional representation and map this information again into a vector of 62 parameters.

- Temporal ID network

This network is used to compare the similarity between the input features and to act as a discriminator for adversarial learning with the 3DMM generative network presented next. The process is: the two feature vectors are mapped and the similarity between them is compared, the similarity is compared with the labels, and if the decision is wrong, the parameters of the 3DMM generative network are updated to make it generate features that better distinguish the key information between true and false.

- 3DMM generation networks

  The role of this network is to generate videos similar to those tampered with by deepfake, as shown on the figure: the facial features of identity A, are placed on top of the facial background of identity B, i.e., information consistent with the individual's visual identity but not the biometric features. It is generally used twice to change individual i into identity c and will be changed into 3DMM features, after which the generated 3DMM features are then re-transformed into i. The generator aims to increase the similarity, while the temporal ID network training hinders the generator, and the ultimate goal of adversarial training is to improve the ability of the temporal ID network to distinguish real identity from false identity

2.2.2 Multi-attentional Deepfake Detection

In this algorithm, depth pseudo-detection is described as a fine-grained classification problem, and a new multi-attention depth pseudo-detection network is proposed. Specifically, it consists of three key components:

- Multiple spatial attention heads that enable the network to focus on different localities,
- Texture feature enhancement blocks that amplify subtle artifacts in shallow features
- Aggregated attention graph-guided low-level texture features and high-level semantic features. In addition, a new region independence loss and attention-guided data enhancement strategy is further introduced to address the learning difficulties of this network.

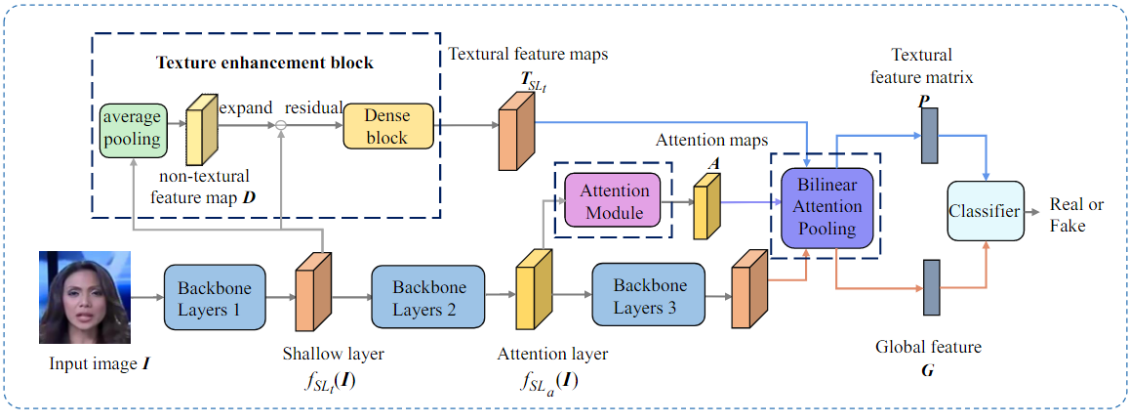

Figure 7: Network of this model

Augmenting shallow features with textures, using deeper features to generate attention maps, and bilinearly interpolating the attention maps and multiplying them point by point with texture features.

To avoid different feature maps focusing on the same region, Regional Independence Loss (RIL) is used to make the distance between feature maps as far as possible, and the attention maps generated by the same channel for different images are as close as possible.

In addition, Attention Guided Data Augmentations is added, i.e., an attention map is randomly selected first, and then the corresponding region of the original map is Gaussian blurred to further decouple the different attention maps.

3 My idea

Since the current deepfake detection algorithms are increasing dramatically year by year with excellent results, while the growth rate of deepfake generation algorithms is growing slowly, my idea focuses on how to generate better deepfakes.

The main area of my thinking is how to better fuse two images (the image to be replaced and the replaced face image) from a given set of images. I think the main problem of fusion is how to extract the facial features from the replaced face, how to locate the position of the face to be replaced, and how to unify the skin color and other stylistic features of the face before and after the replacement.

Based on the above three objectives, I designed a reinforcement learning based deepfake fusion network.

3.1 Feature extraction module

First, we need to get the specific face contour from the two input images. Here, we use a network similar to FPN, firstly, we use a pyramidal FPN network for coarse classification, and each layer of the downward passing branch of the FPN network uses convolution and up-sampling operations for fine classification. Finally, we cascade the obtained feature maps and obtain the final two output face contour regions after one layer of convolution. This region is the final approximate region that needs to be replaced.

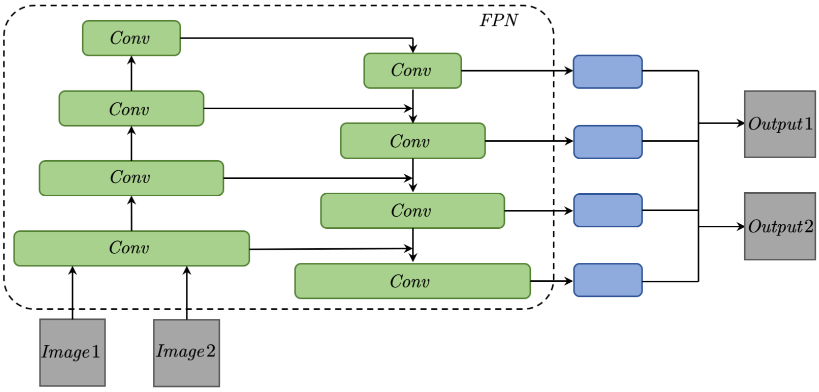

Figure 8: Network of feature extraction module

3.2 Reinforcement learning-based decision network

The second part is a decision network based on reinforcement learning. In this module, the two face region contours obtained above need to be merged, but in the merging, the position information of the face contour needs to be considered (i.e., whether it needs to be moved in four directions: up, down, left, and right, and whether it needs to be stretched and extended), and considering that this problem is a content that requires constant fine-tuning, I try to use reinforcement learning to continuously adjust the features to achieve the best adjustment results.

The net works as follows:

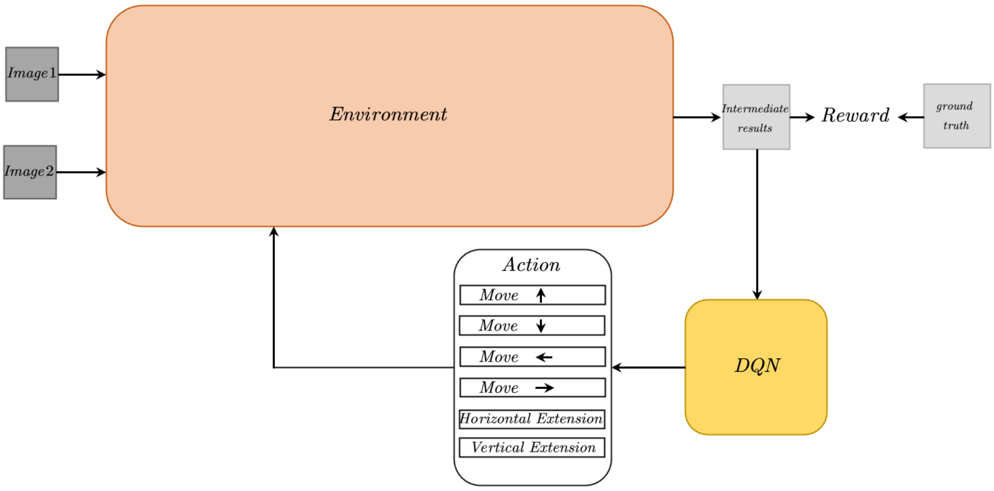

Figure 9: Reinforcement learning-based decision network

In this network, the original input is two face images to be fused, and the proposed segmentation method is used to first segment the face region contour of the two images, and this contour is first simply superimposed to obtain an intermediate result to use this intermediate result as the State for reinforcement learning, where the Deep Q Network (DQN) in reinforcement learning is chosen as the decision network, and the state is passed into the DQN The reward is set as a function of the difference in position between the intermediate result and the ground truth, and the smaller the difference in position between the two, the larger the reward.

3.3 Style transfer module

Considering that there are characteristic differences in the distribution of skin tones, wrinkles, etc. of faces in the two images, existing style migration algorithms are considered to keep the color styles of the overlay faces and the background faces consistent. Within the author’s knowledge, it is possible to use only some style migration algorithms such as AdaIN, CycleGAN, etc., as well as to dynamically adjust the final fusion effect by setting perceptual loss, style loss.

4 Conclusion

In this paper, we briefly state the algorithm and development of deepfake and anti-deepfake, and propose our own method for face feature contour extraction by using FPN network, propose a dynamic strategy adjustment method based on reinforcement learning, and use a style migration method to keep the style of two images such as color consistent. Technology is a double-edged sword, and only by using the deepfake algorithm in the right place can we really play the role of technology for social development.