Recent advance on highly realistic AI-synthesised video approaches makes it hard to distinguish whether a speaker in video is real. Many existing DeepFakes detection approaches can be invalidated by using them as supervision in adversarial training, which in turn improves the performance of DeepFakes. This makes the problem a dilemma. However, human can recognise the identity of familiar persons by observing their motion styles. Inspired by this, the paper proposes a novel method to recognise original speaker’s identity clues in DeepFakes videos by learning individual motion styles. As a biological signature, motion styles can neither be used in the training process of DeepFakes nor be modified to deceive detection methods without reducing realistic performance. Also, this paper proposes a novel pipeline to continuously restore the original frames from DeepFakes videos without knowing the DeepFakes approach in advance. Based on above ideas, our scheme makes it possible to restore DeepFakes videos. Through training the cross-modal transfer module, the appearance embedding can be inferred from the identity code. Extensive experiments are conducted to reveal the effectiveness of the method, it shows for the first time that the original persons can be identified automatically and the original video from DeepFakes videos can be generated.

Introduction: Recent advances in generative adversarial networks have made it significantly easier to create AI-synthesised media (popularly referred to as DeepFakes). The competition between DeepFakes methods and their adversarial methods will be keen for a long period, leaving the social media with periodic floods of fake news and slanders. However, current DeepFakes detection methods focus on finding artificial clues in specific forensic videos, with less attention paid to biological information. In other words, in many occasions, they limit themselves to only effective on existing methods with defects analysis performed, instead of finding a common solution to unseen DeepFakes methods.

Motivated by this, this paper models the biological motion clues in order to provide reliable identity verification results. Thus, our scheme can not only provide a strong and consistent clue for DeepFakes detection problems but also, for the first time, make it possible to restore original media from DeepFakes. The illustration of this task is shown in Figure 1. Given a video modified by DeepFakes, the goal is to identify the original speaker, then infer the face appearance of the original speaker according to the identity code extracted by a specific model, and finally generate the original video. Specifically, as the appearance information in the input DeepFakes video cannot be trusted, all appearance information and other properties (such as poses) can be inferred from previous training results.

Our contributions can be listed as follows: (I) We propose two new tasks — DeepFakes identification and DeepFakes video restoration; (II) we design a spatial convolutional neural network (CNN) to identify speakers with only motion clues provided in DeepFakes identification, which overcomes the issue of previous methods that they are unsuitable for unseen DeepFakes methods; (III) for the first time, we restore the original appearance from DeepFakes, and design a novel transfer net.

Comparison to related works: Some existing studies on DeepFakes detection have been released in the past two years [1–4]. However, most of those approaches are based on manually detecting defects of current DeepFakes methods (or detecting defects by neural networks (NN)). These approaches are very popular but have several issues: (1) Finding defects and designing a special detection method are labour-consuming and have time delay. Researchers need to repeat this process for each DeepFakes method, while new ones continue to come out. (2) They can only perform defects detection on existing methods. Neither human nor NN can find defects on unseen methods including those that have not come out yet and those are not publicly available. (3) Even for existing methods, they may be easily improved once the corresponding defects were reported. (4) These defect-based approaches cannot identify the original speaker because they focus on the defects of each method. On the other hand, Agarwal et al. [5] adopted a two-class Support Vector Machine (SVM) to prove the feasibility that facial expressions and movements can represent an individual’s speaking pattern because it is hard to mimic by actors. But they failed to find a well-designed pipeline and architecture to automatically model and identify multiple persons. Therefore, we propose using NN to model spatial motions and identify speakers in DeepFakes videos. We also make our approach hard to be attacked because modifying the motion styles in the fake video without losing realism is difficult. Furthermore, we utilise the identity clues to perform manipulated video restoration successfully for the first time.

Problem definition and our pipeline: Given a DeepFakes video, our goal is to identify the original speaker, and infer the original face appearance to simultaneously restore the original video, while keeping emotion, pose, gesture and other attributes unchanged. To tackle this challenging problem, we propose a novel pipeline and factorise it into three novel stages as shown in Figure 2. During the training stage, the pose sequence is firstly extracted from the input ground truth video. We then train the motion encoder to learn individual motion styles from the pose sequence. Next, we use the identity-to-appearance transfer network to bridge the appearance embedding and motion embedding. At last, we alternately feed appearance or TransNet into the generator. During testing, we directly input the video which was manipulated by unknown DeepFakes methods. Specifically, since we only input pose sequences and pose images to synthesise the restored frame, we do not use the appearance encoder during the testing stage.

Stage I: We train an appearance encoder $E_a$ and a motion style based DeepFakes identification detector $E_{id}$, based on the original video sequences $x$, and key-point sequences $p$, respectively.

Stage II: We train a pair of cross-modal transfer nets, respectively, denoted as $T_{p \to a}$ and $T_{a \to p}$, to infer the appearance embedding from the identity code extracted by $E_{id}$. Thus, features from two modalities are projected on the latent space under supervision.

**Fig. 1** A simple illustration of our approach

**Fig. 2** The overall architecture of our novel pipeline

\[\text{Generator} \rightarrow \text{Restored Frame} \]

---

**Haichao Zhang, Zhe-Ming Lu, Hao Luo, and Ya-Pei Feng**

\[1\text{School of Aeronautics and Astronautics, Zhejiang University, Hangzhou, P. R. China} \]

\[2\text{College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, P. R. China} \]

Email: zheminglu@zju.edu.cn
Stage III: Given a DeepFakes video $x_s$, we use $E_{id}$ to get the identity code $\beta$ and transfer it into the predicted appearance embedding $\alpha$ with $T_{\beta \rightarrow \alpha}$. Then, we use $E_{ad}$ to infer a most closed appearance $\hat{\alpha}$ from the real video database. Finally, we disentangle the appearance $\hat{\alpha}$ and $p_{\alpha}$ to generate a restored frame. During training, we alternately put $\alpha$ and $\hat{\alpha}$ into the generator through the AdaIn layer [6]. During testing, we use $E_{ad}$ and $T_{\beta \rightarrow \alpha}$ to get the identity code $\beta$ and the inferred appearance $\hat{\alpha}$.

**Method:** In Stage I, a novel motion style based DeepFakes identity detector is proposed. We utilise a CNN-based network to learn the overall motion styles, including gesture motion style, facial muscle motion style and subconscious head movement. As a body language, people use gesture, emotion, even head movement to express themselves. In many situations, their body language is subconscious, which shows in different ways on different persons based on their individual habits. Even different people have different rest poses according to Ginosar et al. [7]. DeepFakes is a kind of face-swap or puppet-master process. In many kinds of DeepFakes, face landmarks and other key points should be consistent with the original videos to increase the realism of results. So, the motion style is naturally suitable for DeepFakes identification problem. But identifying a person from skeleton sequences without appearance information is difficult, so we need to design a new architecture to learn motion-style patterns.

The architecture of DeepFakes identity detector is shown in Figure 3. Firstly, we use the OpenPose [8] to estimate the key-point sequence $p_{\alpha}$ from the input sequence $x_s$. Then, we transform $p_{\alpha}$ into a key-point image $I_p \in R^{f \times k \times c}$, where $f$ and $k$ are the dimensions of frames and key points of $p_{\alpha}$ respectively, and the dimension $c$ consists of $x$-axis, $y$-axis and the confidence value of each key point. Then, we form a motion representation $I_m$ by temporal differencing on the frame dimension of $I_p$. We put $I_p$ and $I_m$ into a two-stream CNN-based network to extract the motion style embedding $y$ from the whole frame sequence. Here $y$ can be calculated as

$$y = E_{conv} \{ I_m(p_{\alpha}), I_p(p_{\alpha}) \}, I_m \in R^{f-1 \times k \times c}, I_p \in R^{f \times k \times c},$$

(1)

where $E_{conv}$ denotes the convolution operation of the two stream CNN layers.

We then use a full-connect net with a softmax function, which is after $E_{conv}$ to classify the identification embedding. Specifically, we denote part of full-connect layers and convolutional layers as the identity encoder $E_{id}$. The calculation of the identity embedding $\beta$ can be simplified to:

$$\beta = E_{id}(p_{\alpha}).$$

Here $E_{id}$ refers to the identity encoder.

According to the appearance encoder, our generator starts training with the appearance encoder $E_{ad}$ to get initial parameters. With the input $K$ frames $f_s (1 \leq n \leq K)$ of $x_s$, the appearance $\alpha$ can be formulated as

$$\alpha = E_{ad}(x_s).$$

Here $E_{ad}$ refers to the appearance encoder.

We need to design a network named cross-modal transfer net (TransNet) to transfer the identity embedding $\beta$ into the appearance $\alpha$ because the appearance information from DeepFakes videos is not trustworthy. Thus, we propose a novel cross-modal transfer net to transfer the motion-style identity information to the appearance code and vice versa. We point out that both appearance and motion-style code share the same identity information. We argue that there is a mapping between $\alpha$ and $\beta$, so we transfer them to each other through a latent identity variable.

As shown in Figure 2, we use a dual encoder and a decoder to simulate this mapping. The mapping of our transfer network can be defined as

$$\hat{\beta} = T_{\alpha \rightarrow \beta}(E_{ad}(x_s)), \hat{\alpha} = T_{\beta \rightarrow \alpha}(E_{id}(p_{\alpha})), \text{where the subscript of } T \text{ is the transfer direction, and } \hat{\alpha} \text{ refers to the referred appearance embedding.}$$

The objective function of our TransNet can be formulated as

$$L_{\text{Trans}} = \lambda_1 L_{\beta \rightarrow \alpha} + \lambda_2 L_{\alpha \rightarrow \beta} + \lambda_3 L_{\hat{\beta} \rightarrow \alpha},$$

(2)

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are hyper-parameters. The feature projection losses $L_{\beta \rightarrow \alpha}$, $L_{\alpha \rightarrow \beta}$ and the appearance reconstruction loss $L_{\hat{\beta} \rightarrow \alpha}$ are calculated as

$$L_{\beta \rightarrow \alpha} = L_{\beta} (E_{ad}(x_s), T_{\beta \rightarrow \alpha}(E_{id}(p_{\alpha}))) \quad (3)$$

$$L_{\alpha \rightarrow \beta} = L_{\alpha} (E_{id}(p_{\alpha}), T_{\beta \rightarrow \alpha}(E_{ad}(x_s))) \quad (4)$$

Here $L_{\beta}$ is the $L_2$ loss function.

As shown in Figure 2, the generator continuously synthesises the original frame with the pose, by taking an image which is drawn according to the key points estimated by the OpenPose. The AdaIN [6] layer aligns the content information of the appearance code $\alpha$ or $\hat{\alpha}$, and transfers it into the auto-encoder as adaptive parameters. Meanwhile, to train a generator capable of generating the frame from motion styles directly, the appearance code is alternatively extracted by the appearance encoder or the motion style based identity encoder with TransNet. We use two discrim- inators to provide the adversarial loss and supervise the consistence of poses between the synthesised frame and the image with key points.

During the alternative training process, our generator takes the drawn image with key points of a single frame to get the pose information, $\alpha$ and $\hat{\alpha}$ through AdaIn layers. Our generator can be formulated as

$$\hat{f} = G(p_n, T_{\beta \rightarrow \alpha}(E_{ad}(x_s))).$$

(5)

Here $\hat{f}$ is the synthesised image and $p_n$ is the pose of the single frame $f$. During the training process of the generator, we disentangle the appearance embedding $\alpha$ and the image with the key points $p_n$ to synthesise the video frames with appearance of the original identity, in which the appearance is alternatively taken from $E_{ad}$ and $E_{id}$ through $T$. So, our alternative adversarial training loss $L_a$ is calculated as

$$L_a = L(\alpha) + L(\hat{\alpha}),$$

(6)

where $L$ is formulated as

$$L = \lambda_{adv} L_{adv} + \lambda_{face} L_{face} + L_{Trans},$$

(7)

where $\lambda_{adv}$ and $\lambda_{face}$ are hyper-parameters, and $L_{adv}$ and $L_{face}$ refer to the adversarial loss and the facial perceptual loss, respectively.

When our generator $G$ takes the appearance embedding from $\alpha$, the adversarial loss $L_{adv}$ for $G$ is calculated as

$$L_{adv}(p_n, x_s) = -E_{\text{face}}(x_s) \log (D(G(p_n, x_s))).$$

(8)

Here $D$ is our discriminator, and when the appearance embedding is taken from $\hat{\alpha}$, $L_{adv}(p_n, x_s)$ is the distribution of $x_s$.

Following [9], we add a facial perceptual loss with a pre-trained VGG-19 face recognition network to calculate the facial appearance distance between the synthesised frame and the ground-truth, while keeping the identity consistent with the appearance code $\alpha$.

**Experimental results and discussions:** In our simulation, we use the speaker-specific gesture dataset [7] which is the only large person-specific ‘in-the-wild’ speech video dataset with full gestures, faces and upper body motions. 50% of the dataset is adopted as the training set, 20% for validation and 30% for testing.

For the DeepFakes detection task, we use a pre-trained face recognition model to calculate the facial identity distance between the person in input DeepFakes videos and each identity in the database. If the ground truth identity is different from our results, we think the input video was modified by DeepFakes methods. Because our model identifies the speaker in DeepFakes videos for the first time, we propose using the Top-1 accuracy to measure the identification ability.

Table 1 shows DeepFakes detection and identification performance. The ‘Not-working’ in the table means this baseline cannot perform on
this task. Our motion-based style identification model gets higher performance, and it is the only model capable of DeepFakes identification in above baselines. FF++ [4] and FWA [10] are current existing methods based on detecting the trace left by DeepFakes [11]. They cannot identify the original speakers in DeepFakes videos. Our method outperforms the baselines in the DeepFakes detection task, and it is the only method capable of DeepFakes identification.

Figure 4 shows the qualitative generation results. This experiment is implemented in an unequal condition that our model only takes the input DeepFakes video, while other baselines have extra labels or manual help. The face part in Figure 4(c) is synthesised by Faceswap with manually selected pre-trained weights according to identity labels each time. Figure 4(d) is generated by StarGAN with extra identity labels. Figure 4(e) shows our results whose all visual information is inferred from skeletons.

The inferred information from previous training makes our results have higher degree of restoration. Our model generates the original frames with high degree of restoration in an unequal condition. As there is no existing work exactly comparable with our model in the DeepFakes restoration task, we implement three baselines with extra labels or manual operations to make them comparable with our model. During testing, our model only takes the pose of input DeepFakes videos, which means there is no DeepFakes video in our training dataset. This proves that our model is capable of accurately and automatically restoring the original video frames from DeepFakes videos. It also indicates that our model has the ability to identify the speaker in DeepFakes videos processed by unseen DeepFakes methods and infer appearance from previous training to restore the original frame.

Table 2 shows the quantitative analysis of our generation task. Our method gets higher performance compared to the StarGAN [12] baseline in an unequal condition. Our model gets a better score in structural similarity and Fréchet inception distance (↑), even in a condition where all appearance and RGB information are inferred by our motion style based encoder. We do not use the DeepFake baseline because it only generates little part of the image. It indicates that our TransNet and generator proposed in this work enable a robust learning on generation of restored DeepFakes frames.

**Table 1.** Performance of deep fake identification and detection with Top-1 metric

| Method   | Detection rate | Identification |
|----------|----------------|----------------|
| FF++ [4] | 96.4           | Not working    |
| FWA [10] | 95.6           | Not working    |
| MesoNet [3] | 86.3        | Not working    |
| Ours     | 98.9           | 98.9           |

**Table 2.** Quantitative results of generation baselines

| Method         | FID↑ | SSIM↑ |
|----------------|------|-------|
| StarGAN [12]   | 59.68| 0.56  |
| Ours           | 52.36| 0.62  |

Abbreviations: FID, Fréchet inception distance; SSIM, structural similarity.

**Conclusion:** Methods capable of identifying the person from DeepFakes videos are urgently needed. Thus, restoring the appearance of original speakers from DeepFakes videos is an interesting and difficult topic which can be a supplement to current forensic technology. Our approach overcomes the shortcoming of previous DeepFakes detection works that they need to know which kind of DeepFakes scheme is used to produce DeepFakes videos. Also, we utilise the link between motion styles and identities to restore the original appearance from pose sequences in case of fake information polluting the restored results.

We present a pipeline to model the movement styles of speakers and identify the identities in DeepFakes videos based on their movement styles. Also, we infer the appearance embedding of the original speaker from the identity embedding to restore the speaker in the original video. We make experiments on the most popular appearance changed DeepFakes videos and it shows that our method can automatically synthesise the original video based on inferred appearance and accurate identification.

**Acknowledgements:** This work was partially supported by the National Nature Science Foundation of Zhejiang Province, P.R. China (grant no. LQ20F030010) and the Fundamental Research Funds for the Provincial Universities of Zhejiang Province (grant no. GK199900299012-013).

**Author contributions:** Zhe-Ming Lu: project administration and supervision; Hao Luo: formal analysis and validation; Ya-Pei Feng: funding acquisition, methodology and writing, review and editing

© 2021 The Authors. *Electronics Letters* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 23 October 2020 Accepted: 15 December 2020 doi: 10.1049/ell2.12015

**References**

1. Zhou, P., et al.: Two-stream neural networks for tampered face detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2017, pp. 1831–1839 (2017). https://doi.org/10.1109/CVPRW.2017.229
2. Li, Y., Chang, M. C., Lyu, S.: Ictu Oculi: exposing AI created fake videos by detecting eye blinking. Proceedings of the 10th IEEE International Workshop on Information Forensics and Security, Hong Kong, pp. 1–7 (2019). https://doi.org/10.1109/WIFS.2018.8630787
3. Afchar, D., et al.: Mesonet: a compact facial video forgery detection network. Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security. IEEE, Piscataway, NJ, pp. 1–7 (2018). https://doi.org/10.1109/WIFS.2018.8630761
4. Rossler, A., et al.: Faceforensics++: Learning to detect manipulated facial images. Proceedings of IEEE International Conference on Computer Vision, Seoul (2019). https://doi.org/10.1109/ICCV.2019.00009
5. Agarwal, S., et al.: Protecting world leaders against deep fakes. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 38–45 (2019).
6. Huang X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. Proceedings of IEEE International Conference on Computer Vision, pp. 1501–1510 (2017). https://doi.org/10.1109/ICCV.2017.167
7. Ginosar, S., et al.: Learning individual styles of conversational gesture. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA (June 2019), https://doi.org/10.1109/CVPR.2019.00361
8 Cao, Z., et al.: Realtime multi-person 2d pose estimation using part affinity fields. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, pp. 7291–7299 (2017), https://doi.org/10.1109/TPAMI.2019.2929257

9 Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017, pp. 105–114 (2017).

10 Li, Y., Lyu, S.: Exposing DeepFake videos by detecting face warping artifacts. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019).

11 Deepfakes github. https://github.com/deepfakes/faceswap.

12 Choi, Y., et al.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018), https://doi.org/10.1109/CVPR.2018.00916