Deep Fake Generation and Detection: Issues, Challenges, and Solutions

Sonia Salman and Jawwad Ahmed Shamsi, *National University of Computer and Emerging Sciences, Karachi, 422001, Pakistan*

Rizwan Qureshi, *The University of Texas, Austin, TX, 78712, USA*

Detection of fake audio and video is a challenging problem. Deepfake is frequently used for creating fake audios and videos using deep learning techniques. Deepfakes, artificially created audiovisual interpretations can be used in many different ways; such as, damaging the repute of a celebrity, misinformation or hate speech, and it may lead to chaos in the society. Therefore, deepfake detection is of utmost important. This article presents an overview of deepfake detection models and datasets, challenges and opportunities in current methods, and provides some possible solutions. This article is mainly focused on multimodal data modalities to detect audio–visual fakes.

Deepfakes provide an enriched platform for generating synthesized digital content. The great potential of deepfakes induces enormous challenges for developing digital content and discriminating between real and fake contents.

These challenges necessitate a thorough understanding of their generation and detection types, techniques, methodologies, and mechanisms. This article is motivated to address these needs. Through an extensive tutorial, we provide a detailed explanation of the subject. This article is novel in providing a detailed classification of different types of deepfakes, highlighting limitations of existing work[1,2] and proposing an extensive framework for deepfake generation and detection. The work is highly beneficial for the community in understanding the requirements, challenges, and solutions on the topic.

BACKGROUND OF DEEPFAKE VIDEO

Deepfake has been an active area of research and inspired by a wide range of applications. Deepfake research has been aided by a few effective benchmark datasets. These include Celeb DF,[3] FaceForensics++,[4] DFDC,[5] and DF-TIMIT[6] datasets. These are being used to evaluate the performance of audio and video deepfake detection methods.

1520-9202 © 2023 IEEE
Digital Object Identifier 10.1109/MITP.2022.3230353
Date of current version 10 March 2023.

Methods of Deepfake Video Generation

Prominent work in deepfake generation includes DeepFaceLab[7] (DFL) and FaceSwap.[8] These are based on the state-of-the-art deep neural network (DNN) techniques, which learns the probability distribution of a training dataset and sample from the probability distribution to generate fake samples. On the basis of generation techniques, we can classify deepfake generation into four types. These are explained below and also illustrated in Figure 1.

Fake Video/Face-Swapped With Real Audio Deepfake Generation (Type I)

Face-swapped video is the most common type of deepfake, in which target person's face is swapped with source person, whereas the speech of target person remains same. Applications, such as DFL and FaceSwap, swap faces with proper lip-syncing, making it impossible to detect easily.

Generation of swapped-face deepfake uses autoencoders,[9] which consist of two parts: encoder and decoder. The encoder utilizes machine learning (ML) techniques to discover the features of both faces in latent space. These latent features (latent space) are underlying parts of face shared with most people, like eyes, and mouth, around the same place. The encoder identifies similarity between these two faces and compresses target face A into basic latent features that

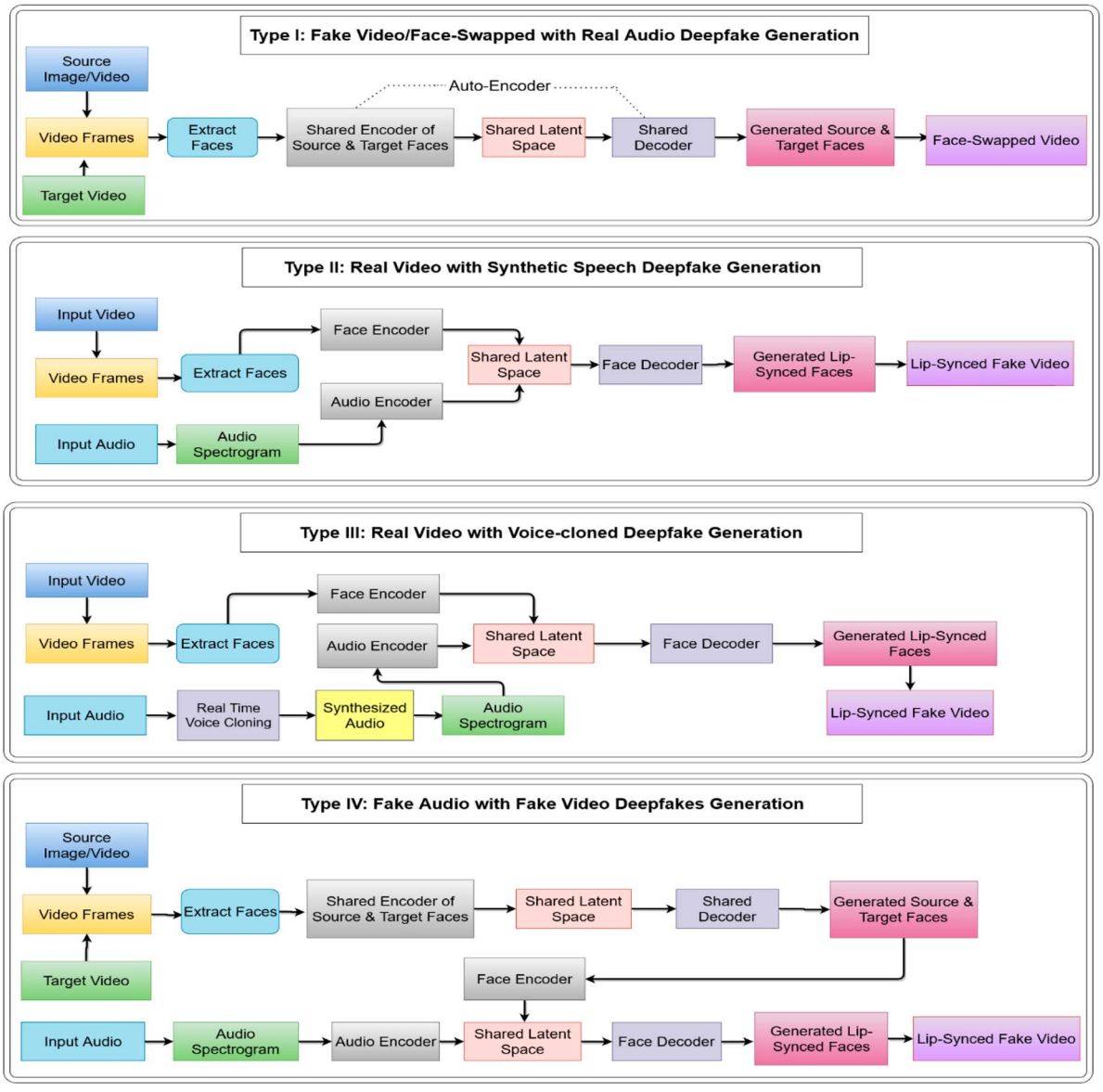

FIGURE 1. In Type I, the extracted faces are passed into a shared encoder, which encodes and stores the main facial features of both the source and target faces in a shared latent space. The decoder then generates faces using those features and swaps them to generate face-swapped video. In Type II, the audio spectrogram receives source audio as an input. The audio is then encoded and input into the shared latent space that contains the target's face's facial features. The decoder then generates lip-synced faces based on the audio input. In this way, the audio/speech can be swapped to create deepfakes. For cloned voices with manipulated words, Type III uses real-time voice cloning. The audio encoder receives the cloned synthetic voice through an audio spectrogram. The audio is then encoded and sent into a shared latent space containing the target's face's facial features. Based on the audio input, the decoder creates lip-synced video. In Type IV, the technique for generating face-swapped videos is the same as in Type I. The modified audio/speech is provided to the generated fake faces to create a fake lip-synced video. In this method, a fake video with synthetic audio is created.

make the target face unique. These latent features are known as feature set. The autoencoder sends that information to decoder, and decoder also utilizes ML techniques to decode features into their original representation. The decoder then merges and overlays source face on top of target face in the video.

Real Video With Synthetic Speech Deepfake Generation (Type II)

The target person's face remains same in voice-swapped deepfake, but voice is swapped with other person's voice. Applications, such as wav2lip and[10] first-order motion model,[11] are utilized for voice-swapped deepfakes. Voice-swapped deepfakes are created by generative adversarial networks (GANs).[12] A target video and source audio are provided as input into the application. First, the frames are extracted from video and fed into the face encoder. The face encoder discovers latent features of the face. The audio signals from video are transformed into a spectrogram. The spectrogram is a visual representation of audio signals in compressed form. This compressed density generates more consistent artificial voices. An audio spectrogram is then fed to an audio encoder for encoding audio features.

Lip-synced video is created when face and audio elements are combined. The GAN discriminator calculates synced losses between face and audio for synchronization and classifies the video as "real" if the loss is near zero. In this way, voice-swapped lip-synced video is generated.

Real Video With Voice-Cloned Deepfake Generation (Type III)

In voice-cloned deepfakes, the target speaker's words are manipulated with her voice. Audio samples are collected and sent into a real-time voice cloning tool to create a similar voice with different words. The audio encoder receives cloned voice as input, and the rest procedure is similar to voice-swapped deepfakes. The politician's speech was changed in synthesizing Obama[13] in the same way. Voice-cloned deepfakes are more dangerous than voice-swapped deepfakes. Applications, such as Google Tacotron2[14] and DeepVoice,[15] are used to create synthetic speech/voices.

Fake Audio With Fake Video Deepfakes Generation (Type IV)

In this type, both the face and speech of target person are swapped. Swapped-face video is first created, like Type I (see Figure 1) technique, and then fed into face encoder. After conversion into an audio spectrogram, the synthetic audio is passed into an audio encoder. The audio encoder encodes the audio features, and shared latent space contains attributes of both swapped face and audio. The face decoder merges the swapped face and audio characteristics to generate lip-synced video. The discriminator then calculates the synced losses, such as contrastive or binary cross-entropy loss between the face and audio for synchronization. If the loss is near zero, then the video is termed as real by discriminator. In this way, voice-swapped lip-synced video is generated.

Methods of Deepfake Video Detection

Early detection approaches relied on handcrafted features derived from defects in the manipulated video. However, they are less accurate than recent deep learning (DL) techniques. Contemporary methods[16,17] use DL to extract prominent and discriminating features for deepfake detection. Table 1 provides the description and limitations of the state-of-the-art deepfake detection techniques.

LIMITATIONS IN THE EXISTING DEEPFAKE VIDEOS GENERATION AND DETECTION

This section discusses the limitations in existing deepfake generation and detection approaches.

Limitations in the Existing Deepfake Generation

Deepfakes still have some drawbacks, despite continuous improvements in their quality. The following are some of the most prevalent flaws of deepfakes.

Deficiency in Generation of Image Details

Deepfake generation methods may not develop fine facial details due to data loss during the generation's encoding stage. The information is insufficient for high-quality face generation because of compressed representation in an end-to-end framework.

Incapability to Handle Occlusion on Face

It is hard to generate face-swapped video with any object on the face like hands or spectacles, etc. The absence of that particular object/occlusion makes the deepfake video imperfect. Existing methods use a face segmentation network[18] trained on facial data consisting of occlusion-aware face masks, but this approach cannot detect unseen occlusion categories.

Requirement of a Robust DeepFake Generation Model

Existing methods usually use separate models for decoding face and audio. Due to improper lip synchronization and face occlusion, it is hard to produce high-quality video utilizing different models. We need an end-to-end network that can decode multimodal data

TABLE 1. Existing Deepfake Detection Methods with Limitations.

Detection technique	Description	Benefits	Limitations
Detection based on visual artifacts.	Closed observation of deepfakes may reveal minor differences and inconsistencies between background and foreground.	Visible artifacts methods may detect inconsistency in the blending boundary of a modified object, such as image resolution.	These visible artifacts are rapidly diminishing as deepfake algorithms advance, demanding to exploit more intrinsic properties.
Detection based on GAN fingerprints.	Synthesized faces generated by GANs often have GAN generated fingerprints.	By the use of deep features, fingerprints identified in the GAN generated fake images.	Because of different versions of GANS, no universal fingerprint metric can be adopted.
Detection based on biological signals, such as eye blinking.	There can be abnormalities in the eye blinking frequency in deepfakes.	Synthetic biological signals are easier to detect, such as eye blinking and heartbeat.	Advanced versions of deepfakes face generation very precisely model the biological signals, making it harder to detect.
Detection based on adjacent video frames' continuity.	Deepfakes may have flickering, jittering, and different face positions due to the discontinuity among adjacent video frames.	Temporal consistency-based detection methods can recognize discontinuity in adjacent video frames.	Poor performance on low-quality videos as continuity between adjacent frames is affected by video compression.
Detection based on face emotions.	Alignment of facial emotions is improper on swapped faces in deepfakes.	Siamese network-based architecture can detect non-alignment of facial emotions by facial and audio features extraction.	This technique fails if the video has no emotions.
Detection based on out of lip-synced videos.	A deepfake video with synthesized audio may have out-of-sync lips.	The difference between visemes (mouth shapes) and spoken phenomes (utterances) is used for out-of-sync lips detection.	Improved GANs can now generate proper lip-synced deepfakes. Also, any out-of-sync video does not need to be deepfake.
Multimodal detection technique.	Deepfakes created by swapping the face and audio, same as Type IV in Figure 1, are known as multimodal deepfakes.	Multimodal detection techniques first detect swapped faces, then use lip-syncing techniques to identify manipulated speech.	Accuracy suffers as detection techniques extract audio from different datasets instead of the same deepfake video.

in a high-dimensional data space for realistic deepfakes generation.

Limitations in the Existing Deepfake Detection

Deepfake detection methods also suffer from following issues regardless of their approach.

Deepfake Detection Suffers From Postprocessing Operations

Postprocessing operations can be applied in deepfake video for enhancement by removing minor artifacts, such as blurriness in background and fine-tuning the video frames by adding filters in 3-D setting. Deepfakes look highly realistic and challenging to detect after postprocessing operations.

Limitations in the Existing Datasets

There are several limitations in existing audio–visual deepfake datasets. For example, video and synthesized audio are not lip-synced, and participants are not facing the camera in most recorded deepfake videos, etc. As a result, a benchmark dataset of deepfake videos, such as ImageNet, for deepfake detection is required to address these issues.

COMPREHENSIVE DEEPFAKE GENERATION AND DETECTION MODEL

Our proposed deepfake generation model (see Figure 2) overcomes the limitations of existing generation methods discussed previously. It is an end-to-end network that decodes audio–visual data in a high-dimensional

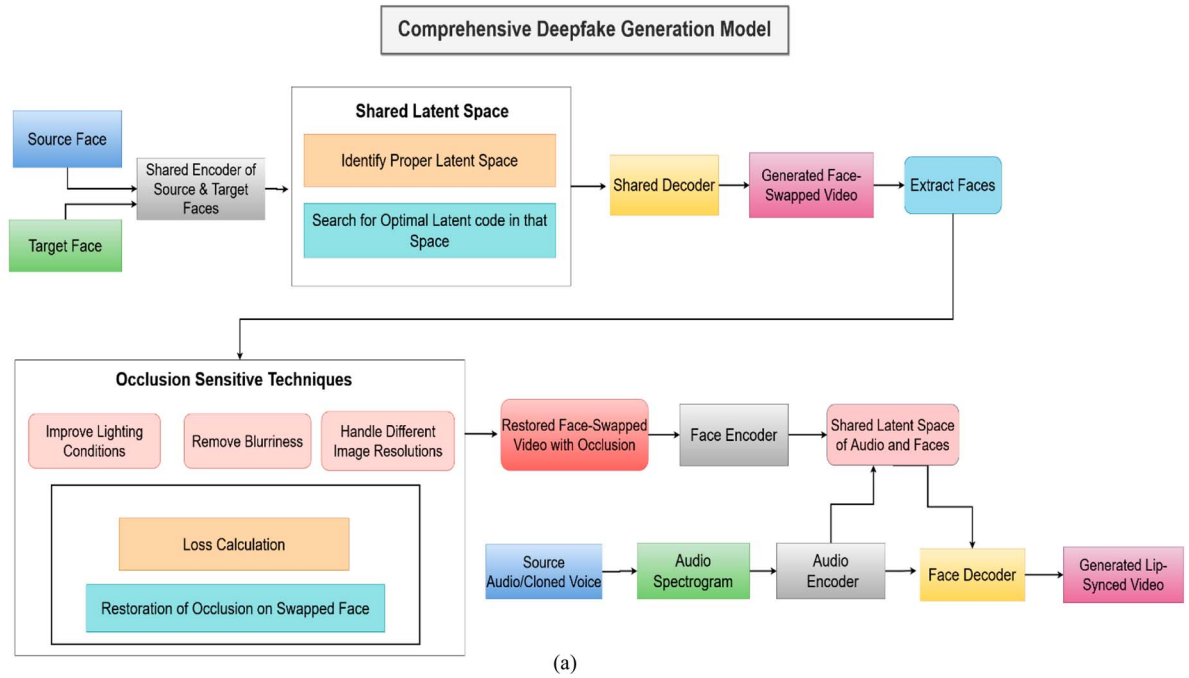

FIGURE 2. (a) Proposed deepfake generation model generates face-swapped video with fake audio. It searches the optimal latent code in the latent space for targeting facial attributes for a perfect face swap and also deals with the occlusion, orientation, and background. (b) Block diagram of the proposed audio–visual deepfake detection model. It consists of two modules: The computer vision model for swapped-face detection and audio processing model for fake audio detection.

latent space and generates swapped faces with fake audio deepfakes.

Deepfake Generation Approaches

Following are a few proposed techniques for an extensive deepfake generation model.

Face Swapping: Generation of Fine Facial Details via Latent Space Embedding

Shared latent space contains the common attributes of both source and target faces. Decoder generates the swapped face from this latent space that may not include precise facial details required for proper face swap. The solution is to locate the best suitable latent code from an expanded latent space by identifying entire face representation.

Face Swapped Video: Deal With Challenges, Such as Occlusion and Image orientation

An occlusion is anything that appears on the face, such as spectacles and a hand. The existing swapped-face generation method often fails to handle occlusions properly. Our robust model uses occlusion restoration techniques to restore any disappeared occlusion on the swapped face. Similarly, the model improves lightning conditions, removes blurriness, and handles different orientations and image resolutions.

Face Swapped With Swapped Audio/ Cloned Voice Generation.

Finally, fake or cloned voice is input into face-swapped video. The audio and facial features combine in a shared latent space, which then passed to the face decoder. The decoder then generates the perfect lip-synced deepfake video.

In the block diagram (see Figure 2), the source and target face from a video/image are passed to an encoder that contains both faces' shared attributes. Shared latent space includes common properties of both faces in compressed form. The goal is to determine the best latent space for generating swapped-face correctly. The swapped face looks similar to target face if we keep excessive features of the target face. Hence, we need to find the optimal latent code inside that latent space. The decoder receives target face attributes selected from the optimal latent space as an input and then merges the generated face onto target face in the video. The occlusion restoration technique is added to the model that determines the presence of occlusion in generated deepfake video.

First, it compares the target's real face with the swapped-face to find the disappearance of any occlusion in the deepfake video. Using this, loss is computed and compared with a threshold to determine an anomaly. Facial features will be evaluated again to restore that face occlusion.

Deepfake Detection Approaches

We have also proposed a comprehensive deepfake detection model [see Figure 2(b)], which detects fake audios and swapped faces. Following are the proposed approaches.

Training on Diversified Dataset

Deepfakes created with advanced tools, such as DFL are of high quality, whereas, Deepfakes created with smartphone applications are of low quality and compressed. Hence, a DNN must be trained on diversified deepfake video dataset that contains videos of any resolution quality and compression rate to assure our model's accuracy.

Use of Frequency Domain

Artifacts of swapped face are destroyed by modifications, such as data compression, rendering them undetectable in the RGB domain but detectable in the frequency domain. Hence, frequency information should be used as a complementary modality to reveal artefacts that are no longer visible in RGB domain.

Integration of Computer Vision and Audio Processing Modules

As deepfakes can have swapped faces and audio, our proposed deepfake detection model is robust because it integrates audio and visual modalities. Vision transformer is the current CV model, whereas signal processing techniques are used for audio processing. Vision transformers' self-attention mechanism enables the network to acquire higher level information faster, resulting in improved performance in less processing time. Our model will detect fake audio by employing signal processing techniques, such as melfrequency cepstral coefficient in audio processing module. Spectrogram and wavelength of fake audio signal are processed using signal processing methods.

The input video is passed to the CV block, which processes the images, as shown in Figure 2(b). Facial landmarks are identified after face extraction. Frequency-domain characteristics are calculated using face landmarks, and those features, along with visual attributes, are supplied into the self-attention DNN classifier model.

The audio spectrogram is computed using the audio features of extracted video frames and then passed as input to a sequential DNN classifier model that predicts the audio as real or fake.

RESULTS AND DISCUSSION

We have adopted SyncNet[19] as the base model because it detects fake contents using two categories, i.e., Type 2 and Type 3. For comparison, we have implemented a solution consisting of two-stream convolutional neural network architecture, which determines the audio–video synchronization between mouth movements and speech in the input video.

For both the models, audio and video parts are separated and the Euclidean distance between them is calculated. If the distance is greater than the threshold value, then the video is classified as fake. In our model, we also compute the frequency domain of facial regions using a high-pass filter, which enables us to detect face-swapped videos (Type 1). This enhancement in our model enables us to yield high true positive values as compared to SyncNet.

We compared the performance of SyncNet with our proposed method using accuracy, precision, recall, and F1-score by following the formulas given in (1)–(4) in the following. Accuracy reveals how frequently the model was overall correct. Precision measures how well the model predicts a specific category. Recall shows how frequently the model is able to detect a particular class, and the F1-score is used to determine a model's performance by considering precision and recall

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} * 100$$

$$\text{Precision} = \left(\frac{\text{TP}}{\text{TP} + \text{FP}}\right) * 100$$

$$\text{Recall} = \left(\frac{\text{TP}}{\text{TP} + \text{FN}}\right) * 100$$

$$\text{F1} - \text{Score} == \left(\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}\right) * 100.$$

TP is true positive when the video is real and model prediction is also real. TN is true negative when the video is fake and prediction is also fake. FN is false negative where the video is real but prediction is fake and FP is false positive when the video is fake but predicted as real.

For experiments, we have created a dataset of 200 videos consisting of real as well as fake videos with all the four types. Experiments indicate that our proposed model yields an accuracy of 66.5%. In comparison, SyncNet achieves the accuracy of 32.5% only. Precision and recall of our proposed model are 64% and 75%, whereas SyncNet achieves 24.6% and 17%, respectively. F1-Score of our proposed model is 68.9% and that of SyncNet is 20%.

The experimental evaluation confirms that our proposed model is extensive and enriched as compared to SyncNet. We plan to continue along these directions for refinements in deepfakes detection and generation.

REFERENCES

1. Y. Peipeng, J. Fei, Z. Xia, Z. Zhou, and J. Weng, "Improving generalization by commonality learning in face forgery detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 0, pp. 547–558, 2022.
2. K. Janavi, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian J. Sci. Eng.*, vol. 47, no. 3, pp. 3447–3458, 2022.
3. L. Yuezun, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3204–3213.
4. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1–11.
5. B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.
6. P. Korshunov and S. Marcel, "Deepfakes: A new threat to face recognition? assessment and detection," 2018, *arXiv:1812.08685*.
7. I. Perov et al., "DeepFaceLab: A simple, flexible and extensible face swapping framework," 2020.
8. Faceswap. [Online]. Available: https://github.com/deepfakes/faceswap
9. A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022.
10. K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 484–492.
11. S. Aliaksandr, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Adv. Neural Inf. Process. Syst.*, 2019, *arXiv:2003.00196*.

12. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Commun. ACM*, 2014, pp. 2672–2680.
13. S. Supasorn, S. Seitz, and I. Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.
14. W. Yuxuan et al., "Tacotron: Towards end-to-end speech synthesis," 2017, *arXiv:1703.10135*.
15. W. Ping et al., "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 214–217.
16. H. Alexandros, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5037–5047.
17. A. Shruti, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 2814–2822.
18. N. Yuval, T. Hassner, and Y. Keller, "FSGANv2: Better subject agnostic face swapping and reenactment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 560–575, Jan. 2023.
19. J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 251–263.

SONIA SALMAN is a Ph.D. candidate in computer science at FAST—National University of Computer and Emerging Sciences, Karachi, 422001, Pakistan. She is a lecturer at VU. Her research interest includes deep learning and deepfakes. Salman received her M.S. degree in computer networks from Virtual University, Lahore, Pakistan. Contact her at sonia.salman@nu.edu.pk.

JAWWAD AHMED SHAMSI is a professor of computer science and dean computing at FAST—National University of Computer and Emerging Sciences, Karachi, 422001, Pakistan. His research focuses on developing scalable and intelligent systems through integration of computing technologies. Shamsi received his Ph.D. degree from Wayne State University, Detroit, MI, USA. Contact him at jawwad.shamsi@nu.edu.pk.

RIZWAN QURESHI is currently working at MD Anderson Cancer Center, The University of Texas, Austin, TX, 78712, USA, and is a postdoctoral researcher at Hamad Bin Khalifa University, Qatar Foundation Doha, 34110, Qatar. His research interests include the intersection of AI and life sciences, deep learning, signal and image processing, and cancer data science. Qureshi received his Ph.D. degree from the City University of Hong Kong, Kowloon Tong, Hong Kong. Contact him at riahmed@hbku.edu.qa.