Research and Analysis of Deepfake Video Generation and Detection

Chin-Yuan Lin[1], Jen-Chun Lee[2], Shuenn-Jyi Wang[1], Chung-Shi Chiang[2], Chao-Lung Chou[3]*

[1] Department of Computer Science and Information Engineering, National Defense University, Taiwan
[2] Department of Telecommunication Engineering, National Kaohsiung University of Science and Technology, Taiwan
[3] Department of Information Engineering and Computer Science, Feng Chia University, Taiwan
ichilin929@gmail.com, i923002@nkust.edu.tw, sjwang.jason@msa.hinet.net, g950302@gmail.com, clchou@fcu.edu.tw

Abstract

Deepfake technology refers to the use of deep learning to generate fabricated audiovisual content. It can be utilized in virtual reality and movie production. However, if used for malicious purposes, it can impact individuals and society and pose national security issues. Governments worldwide have introduced policies and legislation to address these issues. Detecting deepfake videos is thus critically important. This study analyzes methods of generating and detecting deepfake videos. We used frameworks such as DeepFaceLab and FakeApp to produce deepfake videos. Employing the Xception depthwise separable convolution network model as the foundation, we adopted the FakeVideoForensics method for continuous video analysis, the DeepFakes_FacialRegions method for specific facial feature analysis, and the Improved Xception method, which showed better performance. We have developed an enhanced depthwise separable convolution and facial feature extraction method for deepfake video detection, named Improved Xception Feature Fake Video Detection (IXFFVD). This method was trained using UADFV, FaceForensics++, Celeb-DF, and DFDC datasets, and it was used to detect face-swapped deepfake videos created by frameworks such as DeepFaceLab and FakeApp. Experimental results show that IXFFVD outperforms the compared methods. Although the detection efficacy of this method could be further enhanced, future work will focus on incorporating spatiotemporal detection criteria to improve the detection of deepfake videos created from different datasets.

Keywords: Deepfake, Deep learning, Security, Depthwise separable convolution

1 Introduction

Deepfake technology, based on deep learning, refers to the creation of forged videos using artificial intelligence to replace a person's face in one image with another, creating deceptive and realistic scenes that never occurred. Deepfake utilizes Generative Adversarial Networks (GANs) and autoencoders for face swapping [1]. Deepfake technology offers beneficial applications in virtual reality movie production, such as during the Covid-19 pandemic when film companies faced shooting difficulties and high actor costs, leading Synthesia LLC to release video synthesis technology in 2021 for movie and television production [2]. However, the same technology, when used maliciously, can spread unsettling videos that disrupt social order.

Deepfake primarily employs unsupervised learning with Generative Adversarial Networks [3], and its technology has rapidly evolved from a time-consuming and complex process to one where convincing deepfakes can be created from a single photo, using software like FakeApp, FaceApp, and Zao [4]. Deepfakes are often maliciously used in pornographic videos, causing harm, threats, or humiliation, potentially leading to retaliatory actions by the victims [5].

Political deepfakes represent another common malicious use of this technology, impacting political developments and misleading the public [6]. Misuse of deepfakes in national issues can harm national security, affecting democratic, legal, and sovereign values, and destabilizing peace [7].

During the Russia-Ukraine conflict, deepfakes were used in information warfare, with a deepfake of Ukrainian President Zelensky appearing on Ukrainian TV, announcing the surrender of Ukrainian soldiers [8]. In response, pro-Ukraine groups circulated a deepfake of Russian President Putin calling for Russian soldiers to surrender [9], demonstrating the use of deepfakes by both nations to influence the war (Figure 1).

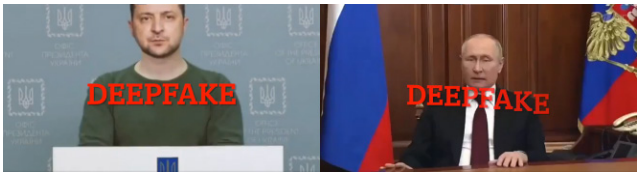

Figure 1. Deepfake videos in the Russia-Ukraine conflict [8-9]

In light of the potential dangers and proliferation of deepfake technology, governments have introduced policies and regulations to mitigate harm and have focused on developing detection technologies. The European Union, in 2021, published a draft of artificial intelligence regulations requiring video creators to proactively label content and mandate transparency in AI applications [10]. In the United States, the 2021 National Defense

*Corresponding Author: Chao-Lung Chou; Email: clchou@fcu.edu.tw
DOI: https://doi.org/10.70003/160792642026032702001

Authorization Act established regulations addressing deepfake technology, mandating the Department of Homeland Security to publish analytical reports on deepfakes and develop detection technologies and corresponding countermeasures [11].

To reduce the harm caused by deepfake videos, detection technologies have become crucial. Initially, basic differentiation methods like skin tone of the face, eye reflections, and facial afterimages were employed. However, due to the rapid development of deepfake technology, these basic methods have become insufficient. Major technology companies have been launching solutions to address deepfake videos. In 2020, the social media company Facebook released a database of about 100,000 deepfake images, aiming to detect deepfakes using artificial intelligence methods [12]. In the same year, prior to the US presidential election, Microsoft developed deepfake detection software capable of scoring videos to determine whether they are products of deepfake technology [13].

To address the threats posed by deepfake technology, developing techniques for detecting deepfake videos is crucial. This study was conducted with this objective in mind, and its contributions are as follows: (1) Study the types of deepfake videos, their production methods, detection techniques, and datasets for detection, (2) Leveraging the Xception depthwise separable convolution and harnessing the strengths of FakeVideoForensics, DeepFakes_FacialRegions, and Improved Xception detection methods, we have developed an advanced depthwise separable convolution method for deepfake video detection that extracts facial features, termed Improved Xception Feature Fake Video Detection (IXFFVD), (3) Train with datasets such as UADFV, FaceForensics++, Celeb-DF, and DFDC, and detect face-swapped deepfake videos created using frameworks such as DeepFaceLab and FakeApp, with the aim of achieving better performance than existing methods and developing a more suitable detection framework for deepfake videos.

2 Related Work

Deepfake videos are primarily divided into two main categories: image deepfakes and audio deepfakes. Image deepfakes are further classified into four types: synthesis, replacement, editing, and reproduction. Audio deepfakes, on the other hand, are divided into two types: voice replacement and text-to-speech (Figure 2). This study will focus on exploring the technology for detecting face replacement in deepfake videos.

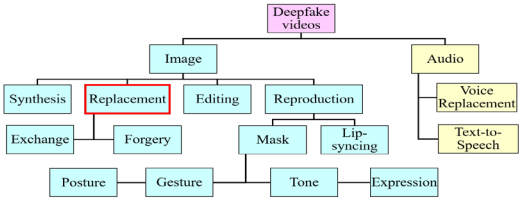

Figure 2. Deepfake videos classification

Depending on the principles of deepfake video production, the methods of creating face replacement deepfake videos can be divided into two types: Face Swapping and Face Morphing. Face swapping deepfake videos are created using autoencoders, which consist of an encoder and a decoder. This process utilizes two autoencoders where the encoder extracts facial features from both the source and target images. These features are then swapped and passed to the decoder for reconstruction to achieve the face swapping effect. Korshunova et al. used convolutional neural networks, training on photo datasets to capture target images and generate face-swapped images (Figure 3) [14]. However, due to the lack of consideration for temporal continuity, it is not possible to produce high-quality deepfake videos. Subsequently, Generative Adversarial Networks (GANs) have become mainstream in deepfake video production. Olszewski et al. leveraged this network architecture to infer the features of each frame of the target video using a single source target feature [15]. FaceSwap [16] improves the performance of autoencoders using adversarial and perceptual losses to generate more realistic faces, while DeepFaceLab is an open-source deepfake video production framework [17].

Figure 3. Face-swapping deepfake videos in [14]

Face morphing refers to the manipulation of facial expressions in videos to forge certain facial movements. Vlasic et al. employed database models of different facial expression parameters for face morphing (Figure 4) [18], but the results lacked temporal coherence. Thies et al. reconstructed facial features using spatial transformation techniques to morph faces [19]; however, their approach only considered changes in expression, resulting in a disconnect with the overall head movement. Kim et al. enhanced temporal and spatial coherence using generative neural networks to create convincingly real facial morphs. This method allows not only for the alteration of facial expressions but also for the movement of the head, resulting in higher overall continuity [20].

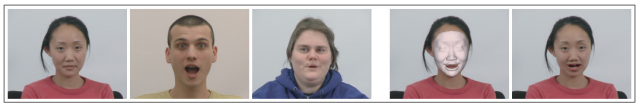

Figure 4. Face morphing examples in [18]

Wang et al. proposed perspectives on deepfake video detection, which can be categorized into detection methods based on deep learning: for instance, Chen et al. conducted detection using datasets [21]; physics-based detection methods: such as Hu et al. who utilized the calculation of the height of corneal reflection between two eyes for deepfake detection [22]; physiological detection methods: as demonstrated by Karras et al., who employed calculations of abnormal facial configurations for detection [23]; and detection methods based on human visual performance: for example, Guo et al. detected deepfakes by segmenting the pupil position from the eyes and analyzing its shape. Yu et al. suggested that techniques for detecting deepfakes could be divided into methods based on time, network, biometric signals, and visual afterimages. In the 2022 study by Xu et al. [24], it was found that the network models most commonly used for deepfake detection and yielding better results include Chollet's XceptionNet from 2017 [25], Afchar et al.'s MesoNet from 2018 [26], Nguyen et al.'s Multi-task from 2019 [27], He et al.'s ResNet from 2016 [28], and Yang et al.'s SVM from 2019 [29], among which XceptionNet is the most widely used network model, appearing in comparisons or as a baseline in over one-third of the studies.

Inception is a deep convolutional network architecture proposed by Google in 2015 [30]. In 2017, Google further improved upon the Inception-v3 version of the network architecture [31], developing Xception, which features depthwise separable convolution (Figure 5). This approach processes channels and spatial dimensions separately, utilizing model parameters more efficiently and enhancing the performance of dataset classification.

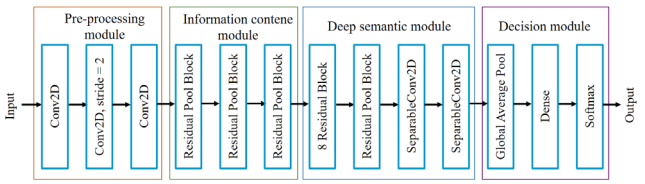

Figure 5. Xception network architecture

In 2019, Rossler et al. collected videos from the internet, specifically from YouTube, using four deepfake creation methods: Face2Face, FaceSwap, DeepFakes, and Neural Textures. They developed the FaceForensics++ deepfake dataset, which includes 1,000 real videos and 4,000 deepfake videos. This dataset was tested using six automatic detection methods: Steg. Features + SVM, Cozzolino, Bayar and Stamm, Rahmouni, MesoNet, and Xception for deepfake identification, among which the method based on Xception showed superior detection performance [32]. Subsequently, in 2019, the Spanish software company BBVA Next Technologies also made public a deepfake detection method named FakeVideoForensics on the GitHub source code hosting platform. This method is based on the Xception approach, operates in two-dimensional space, and was evaluated using the FaceForensics++ dataset. It demonstrated better detection effectiveness for deepfake videos created using deceptive techniques such as FaceSwap, Face2Face, or DeepFakes [33].

In 2022, Tolosana et al. developed a detection method named DeepFakes_FacialRegions, based on four deepfake detection datasets: UADFV [34], FaceForensics++, Celeb-DF [35], and DFDC [36]. They employed three deepfake detection methods: Xception, Capsule Network, and DSP-FWA, and made it public on the GitHub platform [37]. Unlike previous approaches that used the entire face for detection, this method incorporates the facial 68-point landmark analysis tool developed by Baltrusaitis et al. in 2018 [38], focusing on facial regions including the eyes, nose, mouth, and non-facial-feature areas for detection. Moreover, for the newer generation of datasets (Celeb-DF, DFDC), the Xception-based deepfake detection method showed better performance. However, overall, the DeepFakes_FacialRegions detection method can only analyze individual photos, not entire videos.

In 2021, Chen et al. modified the Xception model by removing four residual blocks and replacing the common convolutional layers in the Xception preprocessing module with dilated convolutions from Inception. They utilized a feature pyramid to capture multi-level features, developing an improved Xception model for detecting faces generated by local GANs (Figure 6), which was made public on the GitHub platform [39]. This method uses the FFHQ natural image dataset as a baseline to create a face dataset generated by local GANs (LGGF), showing better detection performance than the original Xception method for deepfake videos. However, this detection method can only analyze individual photos, not entire videos.

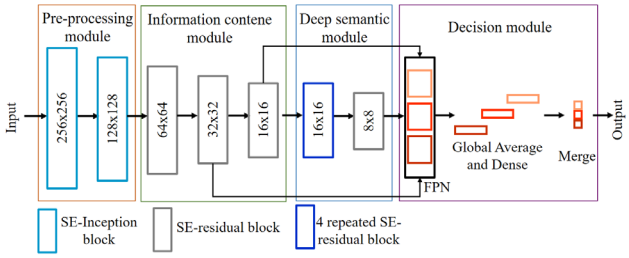

Figure 6. Improved Xception network architecture

In 2024, Ciamarra et al. revealed deepfake videos using the time-surface framework (t-SF), which captures scene surfaces and their temporal variations, identifying inconsistencies caused by deepfake manipulations. These features are used as inputs for binary classifier evaluation in deep neural networks [40]. In 2024, Yan et al. proposed a detection method named LSDA to address the issue of detectors overly relying on specific forgery features. By constructing and simulating variations of forgery features in the latent space, LSDA expands the forgery space and utilizes a binary classifier for forgery detection, thus

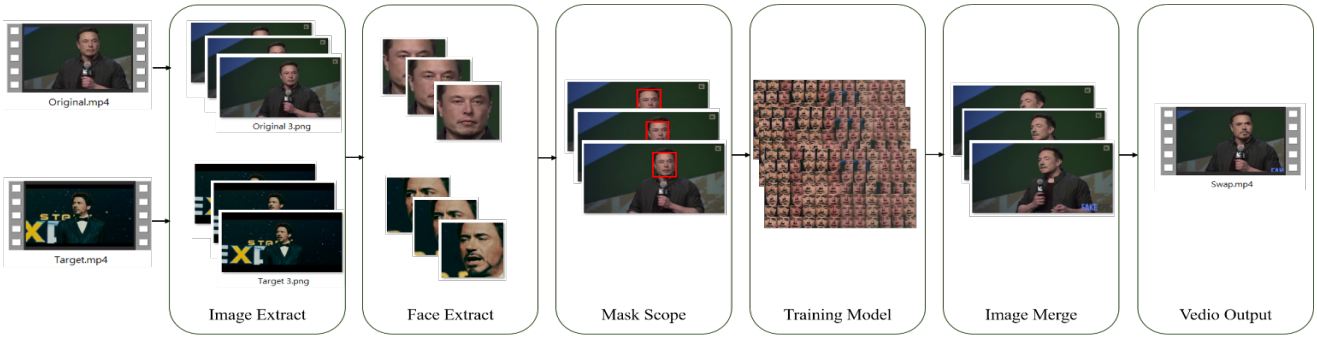

Figure 7. Schematic diagram of the deepfake video production process

learning more generalized decision boundaries and reducing dependence on dataset-specific features [41]. In 2024, Coccomini et al. proposed a deepfake video detection method called MINTIME, which combines spatiotemporal Transformers and convolutional neural networks. This method effectively captures spatiotemporal inconsistencies in videos with multiple individuals and varying face sizes through an identity-aware attention mechanism and position embedding techniques [42].

The quality and quantity of datasets will impact the effectiveness of deepfake detection. In the survey conducted by Xu et al. in 2022, the more popular datasets identified include FaceForensics++ by Rossler et al. from 2019, Celeb-DF by Li et al. from 2020, DFDC by Dolhansky et al. from 2020, and UADFV by Li et al. from 2018. In 2018, Li et al. used the CEW dataset [43], which focuses on detecting eye closure in deepfake videos of faces, to download 50 videos. These videos were produced using the deepfake creation tool FakeApp, with intervals of approximately 30 seconds ensuring at least one blink per interval. The evaluation was conducted using methods based on EAR and CNN. A UADFV dataset was developed, comprising 49 real videos and 49 deepfake videos.

In 2020, Li et al. selected public videos of 59 celebrities from the online video platform YouTube and used the deepfake creation tool DeepFake to enhance the resolution for image synthesis, developing the Celeb-DF deepfake dataset. It was validated using nine deepfake detection methods: Two-stream, MesoNet, HeadPose, FWA, VA, Xception, Multi-task, Capsule, and DSP-FWA. The dataset includes a total of 590 real videos and 5,639 deepfake videos.

In 2020, the well-known social networking company Facebook hosted a deepfake detection dataset competition, paying for the collection of 48,190 private videos. The videos were created using eight deepfake production methods: DF-128, DF-256, MM/NN, NTH, FSGAN, StyleGAN, refinement, and audio swaps, with filters and other means to enhance video recognizability. The selection was implemented using precision metrics, resulting in the development of the DFDC deepfake dataset, which includes 23,654 real videos and 104,500 deepfake videos.

3 Proposed Method

3.1 Face Swap Deepfake Video Production

This study utilizes two deepfake production frameworks, DeepFaceLab and FakeApp [44], to compare the effects of deepfake videos produced by different frameworks. The DeepFaceLab deepfake production framework was made public by Iperov in 2019 on the GitHub source code hosting service platform, using Google's TensorFlow open-source machine learning library. DeepFaceLab performs face swapping using artificial neural networks and is compatible with operating systems such as Windows and Linux, but it produces better results on computers equipped with a GPU; FakeApp deepfake production framework was made public by someone known as Deepfakes on the Reddit social networking site in 2018. FakeApp also uses TensorFlow and artificial neural networks for face swapping, compatible with Windows and Linux operating systems, and has released versions for iOS and Android smartphones. However, the computer version requires a GPU-equipped computer to operate.

The production process for face swap deepfake videos is analyzed as follows (Figure 7):

(1) Image Extraction: Decompose the original and target videos into images.
(2) Face Extraction: Extract facial images from the original and target images.
(3) Mask Scope Detection: Detect the scope of the face in the original video where the target image's face is to be replaced.
(4) Training Model: Conduct deep learning training for face replacement using a network model.
(5) Image Merging: Replace the facial image in the original video with the facial image from the target video.
(6) Video Output: Output the face-swapped deepfake video with the facial replacement completed.

The actual production of deepfake videos revealed that longer training times with deep learning and the use of better computer equipment yield better results. Upon visual comparison, face swap deepfake videos produced by DeepFaceLab (Figure 8) featured more distinct facial contours compared to those created with FakeApp (Figure 9).

Figure 8. Face swap deepfake video produced by DeepFaceLab

Figure 9. Face swap deepfake video produced by FakeApp

3.2 Deepfake Video Detection Methods

This research utilizes the Xception depthwise separable convolutional network model, widely recognized for its efficacy in detecting deepfake videos. It integrates the FakeVideoForensics method to identify characteristics across sequential video frames, meeting practical detection requirements, and employs the DeepFakes_FacialRegions technique's 68-point landmark strategy for the extraction of distinct facial features from images. The network model undergoes enhancements through the augmentation of Dropout layers, thereby reducing overfitting [45] and improving the model's generalization capability. These improvements facilitate the development of an advanced depthwise separable convolution and facial feature extraction technique for deepfake video detection, termed Improved Xception Feature Fake Video Detection (IXFFVD) (Figure 10). Additionally, this methodology is tested across four widely recognized datasets: UADFV, FaceForensics++, Celeb-DF, and DFDC, to detect face swap deepfake videos generated by platforms such as DeepFaceLab and FakeApp. This method is designed to be compatible with operating systems including Windows and Linux, necessitating the use of a GPU-equipped computer for execution.

The loss function in the detection method's network model can be expressed by the following mathematical formula:

$$L = -\sum_i \left[yi \log(pi) + (1 - yi) \log(1 - pi) \right] \quad (1)$$

where L is the loss function used to measure the difference between the predicted value and the actual value; yi is the true label of the i-th sample; pi is the predicted label of the i-th sample; $\log(pi)$ represents the logarithmic probability of predicting a positive class; $\log(1-pi)$ represents the logarithmic probability of predicting a negative class. This method utilizes a loss function for supervised learning, effectively addressing binary classification problems by reducing the loss value to enhance detection capabilities, and employs the 68-point landmark method to improve feature values and enhance model performance.

This study utilizes the detection process of the Improved Method for Deepfake Detection with Depthwise Separable Convolution and Facial Feature Extraction (IXFFVD) (Figure 11), detailed as follows:

(1) Select Dataset: Choose from different datasets (UADFV, FaceForensics++, Celeb-DF, DFDC) for comparison by loading weights.
(2) Load Weights: Load pre-trained model weight files based on specified weights.
(3) Establish Network: Create an enhanced depthwise separable convolution network through selected weights.
(4) Read Video: Read relevant information for each frame of the test video.
(5) Image Processing: Perform preprocessing such as resizing and color adjustment for each frame.
(6) Make Predictions: Use the loaded model to make predictions for each frame and calculate the authenticity of the test video.
(7) Produce Results: After comparing each frame, obtain a true or false score and produce the final detection outcome and output the result video.

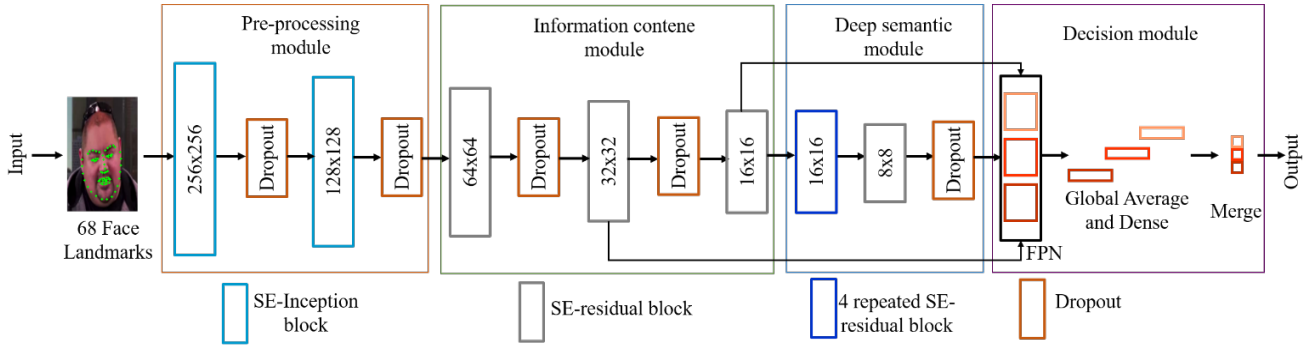

Figure 10. Network architecture diagram of the IXFFVD detection method

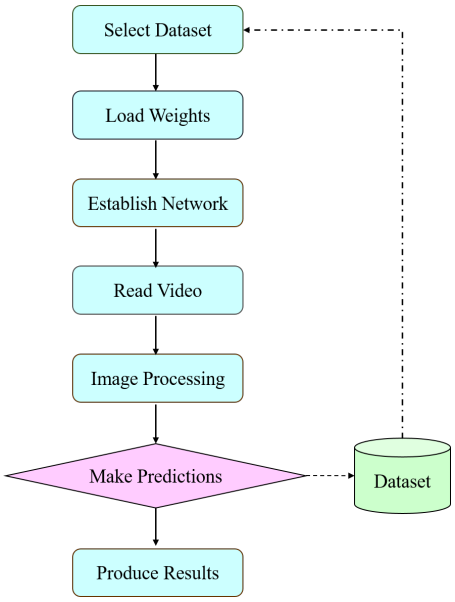

Figure 11. Schematic diagram of the detection process

In this study, the IXFFVD detection method outperforms other methods by detecting continuous videos, utilizing the 68-point landmark method to extract specific facial features, and evolving the Improved Xception network model, achieving the best experimental results (Table 1).

Table 1. Comparison of detection methods

Method	Network model	Comparison
FakeVideo Forensics	Xception	Capable of detecting continuous videos
DeepFakes_ FacialRegion	Xception, Capsule, DSP-FWA	Capable of detecting specific facial features
Improved Xception	Improved Xception	Demonstrates improved detection performance
IXFFVD	Improved Xception	Detects continuous videos, extracts specific facial features, and achieves the best overall performance in this study

4 Experimental Results and Analysis

4.1 Experimental Steps

Using equipment with an Intel Core i7-10875H CPU, NVIDIA GeForce RTX 2070 GPU, and 64 GB of RAM, this study produced face-swap deepfake videos using two frameworks: DeepFaceLab and FakeApp. It employed the Improved Method for Deepfake Detection with Depthwise Separable Convolution and Facial Feature Extraction (IXFFVD) detection framework and conducted tests using four datasets: UADFV, FaceForensics++, Celeb-DF, and DFDC.

The IXFFVD model was trained using the Adam optimizer (categorical cross-entropy loss, initial learning rate 0.001) with a step-decay schedule halving the rate every 10 epochs. Training ran for up to 20 epochs with a batch size of 8. Input frames were resized to 256×256 pixels and normalized to [0, 1]. Balanced class weights addressed class imbalance, and a dropout rate of 0.5 was applied throughout the network. Early stopping (patience = 10, monitored on validation loss) retained the best model weights. For each video, 10 frames were uniformly sampled and preprocessed with the dlib 68-point facial landmark detector.

To evaluate the performance of the proposed IXFFVD deepfake detection model, the datasets UADFV, FaceForensics++, Celeb-DF, and DFDC were divided into training, validation, and testing sets in a 5:1:4 ratio, with three indicators adopted for assessment: accuracy, precision, and recall. In this study, the positive class denotes real samples and the negative class denotes deepfake samples. Let N be the total number of samples in the dataset, TP the number of true positives (correctly detected real samples), TN the number of true negatives (correctly detected deepfake samples), FP the number of false positives (incorrectly detected real samples), and FN the number of false negatives (incorrectly detected deepfake samples). The formulas for the three indicators are as follows:

$$Accuracy = \frac{T_P + T_N}{N} \tag{2}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{3}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{4}$$

4.2 Deepfake Video Detection and Analysis

In this study, the method was trained on videos from four datasets: UADFV, FaceForensics++, Celeb-DF, and DFDC. Feature extraction was performed on each frame, and prior to training, the 68-point landmark method was used to enhance feature values (Figure 12) to improve model performance. Finally, through the IXFFVD deepfake video detection framework, deepfake videos with face swaps created by DeepFaceLab and FakeApp were detected. The deepfake videos produced by DeepFaceLab and FakeApp were successfully detected by the deepfake video detection model trained on the four datasets: UADFV, FaceForensics++, Celeb-DF, and DFDC (Figure 13).

Considering the detection methods FakeVideoForensics, DeepFakes_FacialRegions, Improved Xception, and this study's IXFFVD, an analysis was conducted based on three deep learning metrics: accuracy, precision, and recall, to assess the overall detection outcomes. The IXFFVD method developed in this research demonstrated superior performance across the metrics of accuracy, precision, and recall, compared to the FakeVideoForensics, DeepFakes_ FacialRegions, and Improved Xception detection methods (Table 2).

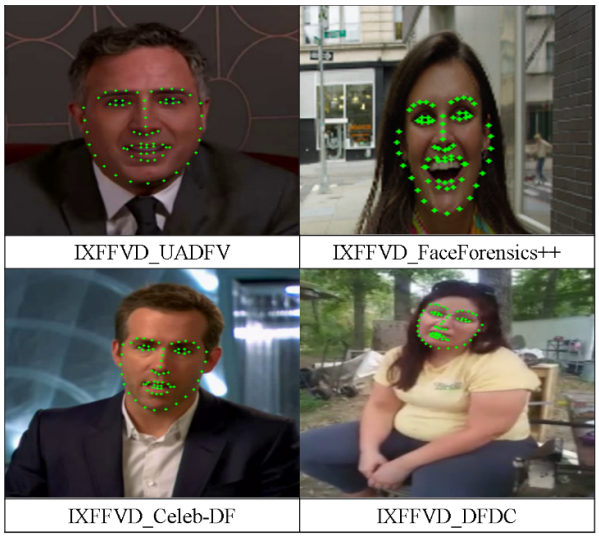

Figure 12. Illustration of dataset image training results

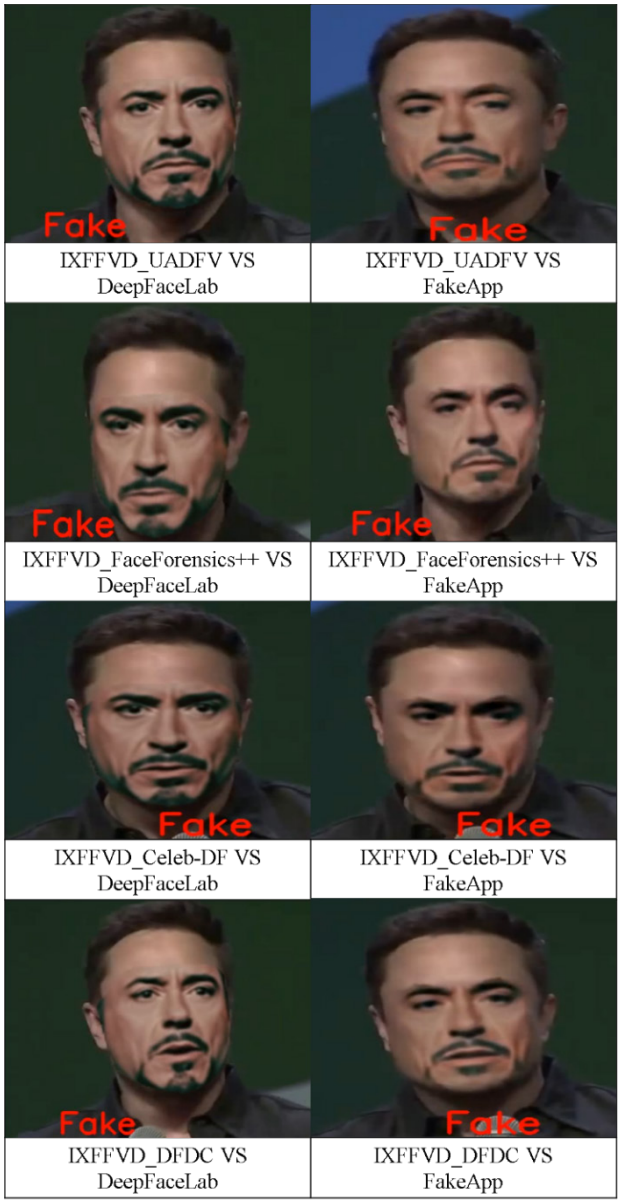

Figure 13. Deepfake video detection results

Table 2. Comparison of detection performance

Category	Accuracy	Precision	Recall
FakeVideoForensics-FaceForensics++	0.4444	0.6667	0.3333
DeepFakes_FacialRegions-DFDC	0.5250	0.5136	0.9500
Improved Xception-Celeb-DF	0.8187	0.8160	0.9386
IXFFVD-UADFV	0.9777	0.9819	0.9684
IXFFVD-FaceForensics++	0.9870	0.9870	1.0000
IXFFVD-Celeb-DF	0.8498	0.8653	0.9353
IXFFVD-DFDC	0.9740	0.6716	0.7287

5 Conclusion

This study builds on the foundation of Xception's depthwise separable convolution and incorporates the 68-point landmark method for facial feature extraction, developing a method suited for deepfake video detection—IXFFVD. It was trained using four datasets: UADFV, FaceForensics++, Celeb-DF, and DFDC, capable of detecting face-swap deepfake videos produced by frameworks such as DeepFaceLab and FakeApp. Compared to other methods, the experimental results demonstrate superior performance. However, analysis indicates that different detection datasets may affect the outcomes. Future work will focus on enhancing the IXFFVD method by incorporating spatiotemporal detection criteria to improve detection efficacy across various datasets, thereby boosting the performance of deepfake video detection.

Acknowledgements

This work was supported by the National Science and Technology Council of Taiwan, grant no. NSTC 112-2221-E-035-084-MY2.

References

[1] P. Yu, Z. Xia, J. Fei, Y. Lu, A Survey on Deepfake Video Detection, *IET Biometrics*, Vol. 10, No. 6, pp. 607-624, November, 2021. https://doi.org/10.1049/bme2.12031

[2] Synthesia LLC, Synthesia software, https://www.synthesia.io, 2021.

[3] R. Chesney, D. K. Citron, 21st Century-Style Truth Decay: Deep Fakes and the Challenge for Privacy, Free Expression, and National Security, *Maryland Law Review*, Vol. 78, No. 4, pp. 882-891, 2019. https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=3834&context=mlr

[4] B. U. Mahmud, A. Sharmin, Deep Insights of Deepfake Technology: A Review, *Dhaka University Journal of Science*, Vol. 5, No. 1-2, pp. 13-23, 2020.

[5] E. Meskys, A. Liaudanskas, J. Kalpokiene, P. Jurcys, Regulating Deep Fakes: Legal and Ethical Considerations, *Journal of Intellectual Property Law & Practice*, Vol. 15, No. 1, pp. 24-31, January, 2020. https://doi.org/10.1093/jiplp/jpz167

[6] D. M. J. Lazer, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, J. L. Zittrain, The Science of Fake News, *Science*, Vol. 359, No. 6380, pp. 1094-1096, March, 2018. https://doi.org/10.1126/science.aao2998

[7] M. E. Bonfanti, The Weaponisation of Synthetic Media: What Threat Does This Pose to National Security? *Elcano Royal Institute*, 2020. https://www.realinstitutoelcano.org/en/analyses/the-weaponisation-of-synthetic-media-what-threat-does-this-pose-to-national-security/

[8] M. Thalen, Hackers Drop Deepfake of Zelenskyy Ordering Troops to Surrender on Ukrainian News Site, *Daily Dot*, March 16, 2022. https://dailydot.com/hackers-zelensky-deepfake-surrender-ukraine-war

[9] M. Holroyd, Deefake Zelenskyy Surrender Video is the First Intentionally used in Ukraine War, *Euronews*, March 16, 2022. https://www.euronews.com/my-europe/2022/03/16/deepfake-zelenskyy-surrender-video-is-the-first-intentionally-used-in-ukraine-war

[10] M. Kop, EU Artificial Intelligence Act: The European Approach to AI, *Vienna Transatlantic Technology Law Forum*, Transatlantic Antitrust and IPR Developments, Stanford University, No. 2, 2021.

[11] S. Briscoe, U.S. Laws Address Deepfakes, *Security Management*, 12 January 2021. https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/

[12] W. D. Heaven, Facebook Just Released a Database of 100,000 Deepfakes to Teach AI How to Spot Them, *MIT Technology Review*, June 12, 2020. https://www.technologyreview.com/2020/06/12/1003475/facebooks-deepfake-detection-challenge-neural-network-ai/

[13] A. Sebenius, Microsoft Releases Deepfake Detection Tool Ahead of Election, *Bloomberg*, September 2, 2020.

[14] I. Korshunova, W. Shi, J. Dambre, L. Theis, Fast Face-Swap Using Convolutional Neural Networks, *IEEE International Conference on Computer Vision (ICCV)*, 2017, Venice, Italy, pp. 3677-3685. https://doi.org/10.1109/ICCV.2017.397

[15] K. Olszewski, Z. Li, C. Yang, Y. Zhou, R. Yu, Z. Huang, S. Xiang, S. Saito, P. Kohli, H. Li, Realistic Dynamic Facial Textures from a Single Image Using Gans, *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 5439-5448. https://doi.org/10.1109/ICCV.2017.580

[16] M. Kowalski, *Faceswap*, https://github.com/MarekKowalski/FaceSwap, 2021.

[17] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, W. Zhang, DeepFaceLab: Integrated, Flexible and Extensible Face-swapping Framework, *Pattern Recognition*, Vol. 141, Article No. 109628, September, 2023. https://doi.org/10.1016/j.patcog.2023.109628

[18] D. Vlasic, M. Brand, H. Pfister, J. Popovic, Face Transfer with Multilinear Models, *ACM SIGGRAPH Courses*, Boston, Massachusetts, 2006, pp. 24-33. https://doi.org/10.1145/1185657.1185864

[19] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2Face: Real-time Face Capture and Reenactment of RGB Videos, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2387-2395. https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.262

[20] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, C. Theobalt, Deep Video Portraits, *ACM Transactions on Graphics*, Vol. 37, No. 4, pp. 1-14, August, 2018. https://doi.org/10.1145/3197517.3201283

[21] B. Chen, W. Tan, Y. Wang, G. Zhao, Distinguishing Between Natural and GAN-Generated Face Images by Combining Global and Local Features, *Chinese Journal of Electronics*, Vol. 31, No. 1, pp. 59-67, January, 2022. https://doi.org/10.1049/cje.2020.00.372

[22] H. Guo, S. Hu, X. Wang, M. C. Chang, S. Lyu, Eyes Tell All: Irregular Pupil Shapes Reveal GAN-generated Faces, *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 2904-2908. https://doi.org/10.1109/ICASSP43922.2022.9746597

[23] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free Generative Adversarial Networks, *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS)*, Virtual, 2021, pp. 852-863.

[24] F. J. Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, Y. Liu, Countering Malicious DeepFakes: Survey, Battleground, and Horizon, *International Journal of Computer Vision*, Vol. 130, No. 7, pp. 1678-1734, July, 2022. https://doi.org/10.1007/s11263-022-01606-8

[25] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 1800-1807. https://doi.org/10.1109/CVPR.2017.195

[26] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: A Compact Facial Video Forgery Detection Network, *IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, China, 2018, pp. 1-7. https://doi.org/10.1109/WIFS.2018.8630761

[27] H. H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos, *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Tampa, FL, USA, 2019, pp. 1-8. https://doi.org/10.1109/BTAS46853.2019.9185974

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[29] X. Yang, Y. Li, H. Qi, S. Lyu, Exposing GAN-synthesized Faces Using Landmark Locations, *IH&MMSec'19: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, Paris, France, 2019, pp. 113-118. https://doi.org/10.1145/3335203.3335724

[30] C. Szegedy, W. Liu, Y Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper with Convolutions, *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015,

pp. 1-9.
https://doi.org/10.1109/CVPR.2015.7298594

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[32] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, FaceForensics++: Learning to Detect Manipulated Facial Images, *IEEE International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 1-11. https://doi.org/10.1109/ICCV.2019.00009

[33] A. Munoz, M. Hernandez, *fakeVideoForensics*, GitHub, https://github.com/bbvanexttechnologies/fakeVideoForensics, 2019.

[34] Y. Li, M. Chang, S. Lyu, In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking, *IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, China, 2018, pp. 1-7. https://doi.org/10.1109/WIFS.2018.8630787

[35] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 3204-3213. https://doi.org/10.1109/CVPR42600.2020.00327

[36] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The DeepFake Detection Challenge (DFDC) Dataset, *arXiv preprint*, arXiv: 2006.07397, October, 2020. https://arxiv.org/abs/2006.07397

[37] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, J. Fierrez, *DeepFakes_Facial Regions*, https://github.com/BiDAlab/DeepFakes_FacialRegions, 2022.

[38] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L. P. Morency, OpenFace 2.0: FAcial Behavior Analysis Toolkit, *IEEE International Conference on Automatic Face & Gesture Recognition*, Xi'an, China, 2018, pp. 59-66. https://doi.org/10.1109/FG.2018.00019

[39] B. Chen, X. Ju, B. Xiao, W. Ding, Y. Zheng, V. H. C. de Albuquerque, Locally GAN-generated Face Detection based on an Improved Xception, *Information Sciences*, Vol. 572, pp. 16-28, September, 2021. https://doi.org/10.1016/j.ins.2021.05.006

[40] A. Ciamarra, R. Caldelli, A. D. Bimbo, Temporal Surface Frame Anomalies for Deepfake Video Detection, *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2024, pp. 3837-3844. https://doi.org/10.1109/CVPRW63382.2024.00388

[41] Z. Yan, Y. Luo, S. Lyu, Q. Liu, B. Wu, Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 8984-8994. https://doi.org/10.1109/CVPR52733.2024.00858

[42] D. A. Coccomini, G. K. Zilos, G. Amato, R. Caldelli, F. Falchi, S. Papadopoulos, C. Gennaro, MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection, *IEEE Transactions on Information Forensics and Security*, Vol. 19, pp. 6084-6096, 2024. https://doi.org/10.1109/TIFS.2024.3409054

[43] F. Song, X. Tan, X. Liu, S. Chen, Eyes Closeness Detection from Still Images with Multi-scale Histograms of principal oriented gradients, *Pattern Recognition*, Vol. 47, No. 9, pp. 2825-2838, September, 2014. https://doi.org/10.1016/j.patcog.2014.03.024

[44] D. Guera, E. J. Delp, Deepfake Video Detection Using Recurrent Neural Networks, *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, New Zealand, 2018, pp. 1-6. https://doi.org/10.1109/AVSS.2018.8639163

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929-1958, 2014.

Biographies

Chin-Yuan Lin received his Ph.D. degree from the Chung Cheng Institute of Technology, National Defense University. His research areas include deepfake forensics and information security.

Jen-Chun Lee received his Ph.D. degree from the Chung Cheng Institute of Technology, National Defense University. His research areas include deep learning and image processing.

Shuenn-Jyi Wang received his Ph.D. degree in Information Management from National Chiao Tung University. His research areas include image processing and multimedia information systems.

Chung-Shi Chiang received his Ph.D. degree from the Chung Cheng Institute of Technology, National Defense University, Taiwan. His research areas include image processing and information security.

Chao-Lung Chou received his Ph.D. degree from the Chung Cheng Institute of Technology, National Defense University, Taiwan. His research areas include information security, deep learning, and biometric recognition.