Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Demonstrations Track

iFakeDetector: Real Time Integrated Web-based Deepfake Detection System

Kangjun Lee[1] , **Inho Jung**[1] , **Simon S. Woo**[1,2]
[1]Department of Computer Science & Engineering, Sungkyunkwan University, Suwon, South Korea
[2]FakeDetector Inc, Seoul, South Korea
{gkdl677, inhovation97, swoo}@g.skku.edu

Abstract

Recently, deepfake detection research has been actively conducted. While many deepfake detectors have been proposed, validating the practicality of such systems against real world settings has not been explored much. Indeed, there are some gaps and disparities when they are applied in the real world. In this work, we developed a real time integrated web-based deepfake detection system, iFakeDetector, which incorporates the recent high performing deepfake detectors, and enables easy access for non-expert users to evaluate deepfake videos. Our system takes a deepfake video as input, allowing users to upload videos and select different detectors, and provides detection results on whether the uploaded video is a deepfake or not. Also, we provide an analysis tool that enables the video to be analyzed on a frame-by-frame basis with the probability of each frame being manipulated. Finally, we tested and deployed iFakeDetector in a real world scenario to verify its practicality and feasibility.

1 Introduction

Currently, deepfake videos pose a serious threat to society, as they create fake information on sensitive topics, as well as defamation on individuals. They can create disbelief and misinformation, lowering the credibility and trust of digital information. Deepfake videos are characterized by their ability to superimpose the face of one individual onto another in a realistic manner. And they have been increasingly misused for nefarious and malicious purposes, such as creating adult content by synthesizing individuals' faces. It is reported that such deepfakes constitute 96% of all deepfake videos, with the sites that primarily post such content garnering billions of views [Misirlis and Munawar, 2023].

Therefore, the need for an accessible system that allows even non-experts, including news agencies, government organizations, etc., to verify the authenticity of deepfake videos easily is critical and important. Moreover, this need is particularly urgent given that deepfakes, having already a grave societal influence, can affect not only celebrities or politicians but also normal people to harm them in various aspects.

However, currently, the majority of deepfake detection works mainly have been on the research side. And there is limited work on developing and providing practical applications and services in a real world setting. That is, despite the extensive research on deepfake detection methods, there are not many systems that are practically deployed to identify deepfakes in the real world. Furthermore, there might be a gap and disparity in transforming research products into real world detection applications.

In this demo paper, we present iFakeDetector, a unified and real time web-based system specifically developed to address the escalating concerns over deepfakes. Our system can handle differing quality of deepfakes with unknown deepfake generation methods so that it can be used in a real world setting. Our system incorporates the state of the art deepfake detection methods. In addition, iFakeDetector is aimed to be used by non-technical users by providing a user-friendly interface that enables users to easily utilize various deepfake detection models and analysis tools. Moreover, our system is expandable and designed to easily integrate and incorporate a new detection method through standardized interfaces.

The main contributions of our system are as follows:

- We developed iFakeDetector, which incorporates a diverse range of advanced deepfake detectors and can report their comprehensive prediction results, where input videos can be varying size and quality, even with unseen deepfake generation methods.
- We provide the analysis tool, which is designed to offer intuitive insights, generating results for frame-by-frame analysis of the input videos and allowing users to identify the specific intervals of the video real vs fake.
- We evaluate the performance of our system against real world dataset and demonstrate the efficacy and feasibility of our system. Currently, our system is used for testing, supporting and identifying deepfakes for national election by the National Election Commission in South Korea. Finally, we present the demo of our system here [1]

2 Our System Overview

The overview and workflow of our system, iFakeDetector, is presented in Figure 1. In our system, we integrate the

[1]https://www.youtube.com/watch?v=a0v3gj4rtjk

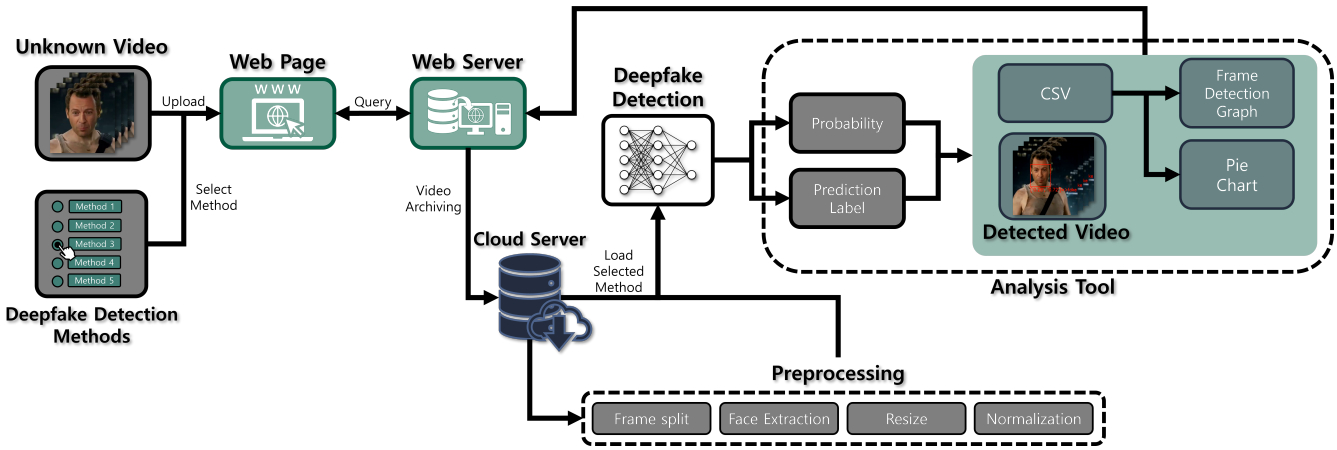

Figure 1: The overview and the detailed workflow of iFakeDetector

following five advanced deepfake detection methods: Xception [Chollet, 2017], ADD [Le and Woo, 2021], QAD [Le and Woo, 2023], CLRNet [Tariq *et al.*, 2021], and BZNet [Lee *et al.*, 2022] to demonstrate the feasibility of our approach. We develop our system as a web-based system to enable non-expert users to choose one out of five deepfake detection models to detect deepfake videos.

2.1 Description of Incorporated Deepfake Detectors

- **Xception.** Xception is based on a traditional Convolutional Neural Network (CNN) and it is widely utilized in deepfake detection.
- **BZNet.** It is a model that takes differing video quality into consideration. Since the quality of video varies widely in the real world, BZNet is trained on multi-scale images utilizing the Super Resolution technique.
- **CLRNet.** When a deepfake video is created, intra-frame and inter-frame artifacts can be smeared on the video. To detect both artifacts, CLRNet is designed by stacking Convolutional LSTM cells. Thus, CLRNet can detect the deepfake video utilizing spatio-temporal information.
- **ADD.** Since low-quality compressed deepfake images are widely used in social networking sites (SNS) in a compressed format, videos have much limited information to discern real vs. fake. To handle low-quality and compressed deepfakes, ADD employs frequency and multi-view attention distillation to effectively utilize the frequency information and correlated information from the teacher network.
- **QAD.** QAD is designated to detect both low- and high-quality deepfakes simultaneously, being the quality-agnostic approach. It employs collaborative learning to utilize raw (high-quality) and compressed (low-quality) images during training simultaneously.

Note that each detection is implemented in a separate container with a standardized interface in a plug and play manner. Hence, it is easy to integrate a new method into our system. Through this flexibility, our system can evolve continuously by integrating the latest deepfake detection models as they develop from ongoing research efforts. This expandability is expected to further enhance the accessibility of researched deepfake detection models in the future.

2.2 Detailed Workflow of Our System

Figure 1 presents the entire workflow of iFakeDetector. First, when a user (client) wants to check if a video contains deepfake content, she uploads the video. Then, the web server stores the uploaded video to the cloud server. Then, the video is divided into frames [King, 2009]. Next, the extracted face images are analyzed using a deepfake detection method chosen by the user to determine if the video contains deepfake contents. Finally, the results are distributed to the user (client) through an analysis tool containing detection result, a CSV file, a frame detection graph, and a pie chart.

In particular, we report the detection result, containing the probability of each face frame being real vs. fake, highlighting the bounding boxes. And the CSV file not only logs the probability of each frame being real or fake but also contains the final prediction of whether the video is deepfake and the progress/status information, as shown in Figure 2. Then, the results are saved into a CSV file, and they are used to visualize the final inference results on the web. This enables the analysis tool to automatically draw a pie chart and frame detection graph on the web. These results provide insights, enabling the user to analyze the video they want to.

2.3 Demonstration System

Figure 2 showcases our user interface, which is designed to allow users to select their desired video and deepfake detection method easily. While iFakeDetector conducts the analysis, users can track the progress of the analysis through a progress bar. For the real-time system, users can download intermediate results, including the analyzed video and the CSV file, during the analysis process. Moreover, since our system is a web-based approach, it can run on the popular web browsers and operating systems seamlessly.

For analysis, Figure 3 is provided to showcase our visualization tool interface, which includes a pie chart and a frame

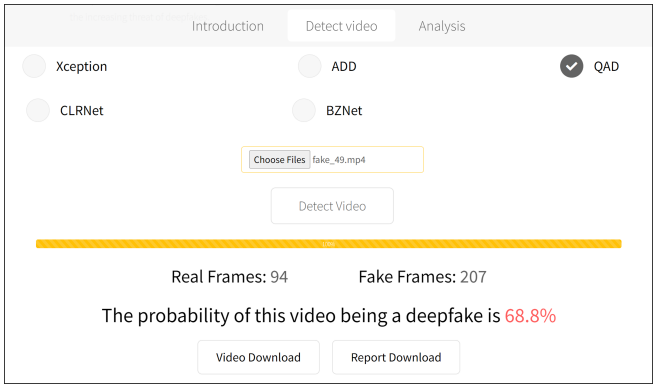

Figure 2: Web interface for deepfake detection results

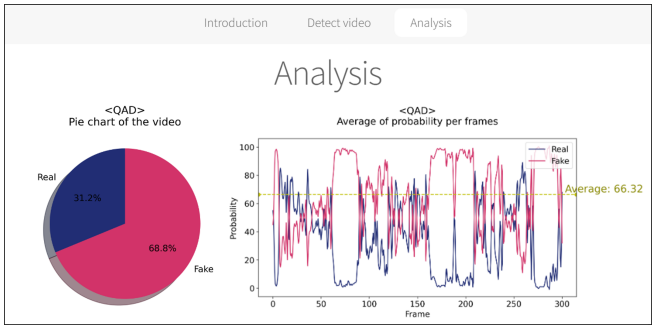

Figure 3: Web interface for visualization and analysis

detection graph for an intuitive assessment of deepfake detection outcomes for non-experts. The pie chart quickly conveys the ratio of frames classified as real vs fake, providing an immediate understanding of the overall detection performance.

Meanwhile, the frame detection graph has the X-axis for video frames and the Y-axis for probabilities of being real or fake for each frame. As illustrated in Figure 3, the red line represents the probability of being fake, the blue line represents being real, and the yellow line represents the average of both. This offers a straightforward and intuitive way to grasp the detection nuances and time-varying detection performance on a frame basis to identify the specific frame intervals that are real vs. fake.

3 Real World Performance Evaluation

Dataset. To validate our system in a real world scenario and demonstrate our system, we collected a real world dataset, which consists of 50 real human face videos from YouTube. For real world deepfake dataset, we used the real world deepfake dataset (RWDF-23) [Cho *et al.*, 2023] and selected 50 deepfake videos, where these deepfake videos are collected from four SNS platforms (BiliBili, Reddit, TikTok, and YouTube) created from 21 countries with various manipulation techniques by online users. Due to these characteristics, it is suitable to construct the real world setting for evaluating our system.

Experiment Settings. For training models, we utilized the FaceForensics++ dataset [Rossler *et al.*, 2019] for training.

Methods	F1-Score	AUROC	Accuracy		
			Real	Fake	Avg
Xception	57.77	70.30	100.00	36.00	*68.00*
BZNet	62.82	72.23	100.00	44.00	*72.00*
CLRNet	61.55	68.76	98.00	48.00	*73.00*
ADD	**85.02**	**86.09**	**98.00**	**78.00**	***88.00***
QAD	**66.71**	**71.60**	**94.00**	**68.00**	***81.00***

Table 1: Performance evaluation of methods with real world dataset

For a quantitative evaluation, we use F1-score, Area Under the Receiver Operating Characteristics (AUROC), and accuracy. The F1-score and AUROC were evaluated at the frame level, whereas accuracy was evaluated separately for real and fake videos.

Results. For F1-score and AUROC, ADD achieves the highest performance, while Xception shows the lowest. In accuracy, all models generally perform well on real videos. However, a significant drop is observed in fake video detection performance. Notably, models, except for ADD and QAD, other models achieve an accuracy of less than 50%. In summary, across all evaluation metrics, ADD consistently displays superior performance, whereas Xception performed poorly, achieving the lowest performance. Also, we observed and learned that real world deepfakes are created with unknown methods, and they are quite different from the training dataset. Hence, we believe this contributes to the generally lower results than reported accuracy in the its respective research papers.

The average end-to-end time to process and generate the final prediction result for a 100Mb video is around 55 seconds, which is less than 1 minute. Hence, it is fast and fully usable in a web-based settings.

4 Deployment and Applicability

Currently, our system is used to help detect harmful and defaming deepfakes related with national election in South Korea. We provide the consultation from our detection results to the National Election Commission in South Korea [ele, 2024].

5 Conclusion and Future Work

We developed a real time integrated web-based system deepfake detection system, iFakeDetector, to be easily used for non-technical users. In the future, we plan to continuously incorporate additional detection methods and provide a better strategy to combine ensemble results. We hope that our tool can be used to counteract the rapidly evolving landscape of deepfake manipulations.

Acknowledgments

We thank Eun-Ju Park for the contribution of providing the real world datasets. This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government

MSIT: (No. RS-2022-II221199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2021-0-02068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. RS-2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes).

References

[Cho *et al.*, 2023] Beomsang Cho, Binh M Le, Jiwon Kim, Simon Woo, Shahroz Tariq, Alsharif Abuadbba, and Kristen Moore. Towards understanding of deepfake videos in the wild. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4530–4537, 2023.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[ele, 2024] The national election commission. https://www.nec.go.kr/site/avt/main.do, 2024. Accessed: 2024-02-10.

[King, 2009] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[Le and Woo, 2021] Binh M Le and Simon S Woo. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. *arXiv preprint arXiv:2112.03553*, 2021.

[Le and Woo, 2023] Binh M Le and Simon S Woo. Quality-agnostic deepfake detection with intra-model collaborative learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22378–22389, 2023.

[Lee *et al.*, 2022] Sangyup Lee, Jaeju An, and Simon S Woo. Bznet: Unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos. In *Proceedings of the ACM Web Conference 2022*, pages 3500–3510, 2022.

[Misirlis and Munawar, 2023] Nikolaos Misirlis and Harris Bin Munawar. From deepfake to deep useful: risks and opportunities through a systematic literature review. *arXiv preprint arXiv:2311.15809*, 2023.

[Rossler *et al.*, 2019] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[Tariq *et al.*, 2021] Shahroz Tariq, Sangyup Lee, and Simon Woo. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of the web conference 2021*, pages 3625–3637, 2021.