Dynamic and Static Features Extraction for Deepfake Detection

Hao Teng, and Chia-Yu Lin
Department of Computer Science and Information Engineering
National Central University, Taoyuan, Taiwan
Corresponding Author: Chia-Yu Lin (sallylin0121@ncu.edu.tw)

2024 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan) | 979-8-3503-8684-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCE-Taiwan62264.2024.10674402

***Abstract*—Deepfake has emerged as a significant concern due to its ability to generate fake images and synthesize realistic videos. The increasing development of new techniques for deepfake creation raises concerns about the cross-forgery issue. Cross-forgery indicates that a model is initially trained to recognize a particular fake and must work against a different unknown forgery. Training a model needs substantial quantities of data, a challenge compounded if the deepfake generation technique is relatively new. Addressing cross-forgery is a critical and essential challenge that requires resolution. In order to solve cross-forgery, our effort presents a method that combines dynamic and static features to identify forgery. For the static component, we extract features similar to general deepfake detection techniques using a single RGB frame as input. Simultaneously, we utilize optical flow analysis to capture changes between consecutive frames for the dynamic part. Our experiments reveal a clear advantage in utilizing combined features, which is particularly evident in cross-forgery scenarios. Specifically, when encountering certain categories, the performance improvement is significant, demonstrating four times better than single-feature models.**

***Index Terms*—Deepfake detection, optical flow, vision transformer, multimodel**

I. Introduction

The generation and manipulation of media through machine learning methodologies has witnessed significant advancements in recent years. "Seeing is believing" no longer holds true in the realm of social media nowadays. Various techniques such as Face2Face [1], NeuralTextures [2], and StyleGAN [3] have emerged, contributing to the diversification of forgery creation methods.

As the technical sophistication of forgery creation continues to grow, a noteworthy challenge arises in the form of cross-forgery. This phenomenon necessitates the development of models trained on specific forgeries capable of discerning against other unknown methods.

Caldelli et al. [4] addressed the limitations of general methods in handling cross-forgery situations. To enhance model robustness, they incorporated optical flow features, thereby strength the overall detection capabilities. Two distinct networks were trained separately with identical architectures, one utilizing optical flow (OF) frames and the other employing spatial (RGB) frames. This approach may result in inefficient resource utilization.

In this paper, we propose a dynamic and static feature extractor for capturing distinctive features from video frames. Our approach leverages both general features and optical flow features to achieve the strongest possible model for deepfake detection. We used an easy way to capture the optical flow features and modified the model to a multimodel structure.

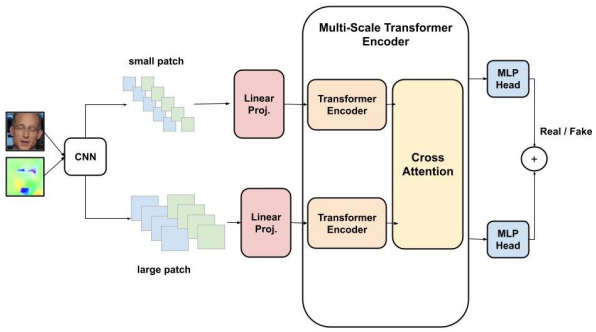

Fig. 1. Modal architecture. RGB feature and optical flow feature input in convolution neural network, respectively, and slice the output of CNN into small and large patches. The blue one is the RGB features patch, and the green one is the optical flow features patch. Then, combine them before inputting them into the transformer encoder.

II. Methods

Fig. 1 is the modal architecture. Combine the two different types of features as inputs for a hybrid CNN and Vision Transformer model. Utilize the benefits of both modalities by slicing the features into smaller and larger patches. This approach enables the model to capture both fine-grained and coarse-grained features, thereby enhancing its robustness.

The model can be separated into two parts: input features, CrossViT, and CvT model. Inspired by [5], we modify some parts based on this model.

A. Optical flow feature

Fig. 2 shows the example of the optical flow feature. Optical flow fields are used to figure out how pixels move between two frames, helping to spot patterns of motion in the image. If you have an image represented by $I = (x, y, t)$, where x and y are the coordinates and t is the time, and this image moves a distance (dx, dy) in a time interval dt, then the new position of the image will be at coordinates $I = (x + \Delta x + y + \Delta y + t + \Delta t)$ at time t + dt. According to [4], different from the RGB feature, optical flow features can capture the motion of consecutive frames. Notice the optical flow provides effective solutions for detecting forgery issues.

Fig. 2. The example of optical flow.

Different from [4], limited the optical flow values to align with the range of RGB frames, requiring meticulous adjustment efforts, we adopt a more straightforward approach. Specifically, we transform the flow file into a flow image straight away.

B. Cross Vision Transformer and CvT

CrossViT [6] captures global context information from input images, enabling the model to enhance its comprehension of the connections among various elements within the image. CrossViT is comprised of K multiscale transformer encoders, with each encoder featuring a large branch that employs a coarse-grained patch size and more transformer encoders alongside a small branch that functions at a fine-grained patch size with fewer encoders. CvT [7] combining the strengths of CNNs and Vision Transformers, CvT utilizes initial processing by a CNN to efficiently utilize inductive bias and translation equivariance. It differs from Vision Transformers, where the input is handled directly.

We have refined our approach. Instead of simply concatenating two images directly, we adopt a multimodel strategy. Here, we separately input distinct features into the CNN feature extractor and then concat the output extracted features. This technique helps prevent the feature extractor from being misled by the boundaries of the concatenated parts.

III. Experiments

In our experiments, we implemented multi-frame intervals for computing optical flow features, which yielded promising results for cross-forgery detection. We use the FaceForensics++ [8], the manipulation could be separated into three categories, one is identity swap (FaceSwap, FaceShifter, Deepfakes), another is face reenactment (Face2Face, NeutalTextures), the last is original. There are thousands of videos in each folder, separate them in train/test/valid in 8:1:1. We train on NeuralTextures and test on itself and FaceSwap, Face2Face. Table I observations revealed a significant improvement in addressing the cross-forgery issue when both training and testing data belonged to the same category. However, when the training and testing data were from different categories, the results were less satisfactory.

TABLE I
Result

Train	Test	Method	Original Acc.	Fake Acc.
NeurlTextures	Face2Face	RGB	0.99	0.07
		RGB + Optical flow	0.99	0.31
	FaceShifter	RGB	1.00	0.11
		RGB + Optical flow	0.99	0.03

IV. Conclusion

In this paper, our paper presents a novel approach to addressing the cross-forgery issue by combining RGB features and optical flow features. By integrating information from both modalities, we achieve a more comprehensive understanding of the visual content, enabling more effective detection of forged images or videos. Through experimental evaluation, we demonstrate the effectiveness of our proposed method in identifying cross-forgery instances with high accuracy and robustness. Thus, our approach is adaptable to various types of forgeries and resilient to common manipulation techniques. This makes it a valuable tool for applications such as digital forensics and multimedia security.

Acknowledgements

This work is sponsored by the National Science and Technology Council (NSTC) under the projects NSTC 110-2222-E-008-008-MY3 and NSTC 112-2622-8-A49-021.

References

[1] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[2] Justus Thies, Michael Zollhöfer, and Matthias Nießner, "Deferred neural rendering: Image synthesis using neural textures," *Acm Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[3] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[4] Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo, "Optical flow based cnn for detection of unlearnt deepfake manipulations," *Pattern Recognition Letters*, vol. 146, pp. 31–37, 2021.

[5] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *International conference on image analysis and processing*. Springer, 2022, pp. 219–229.

[6] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[7] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 22–31.

[8] 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*, 2019, pp. 1–11.