Robust Image Deepfake Detection with Perceptual Hashing

Chun-Shien Lu and Chao-Hsuan Lin
Institute of Information Science, Academia Sinica, ROC, Taiwan

Keywords: AI Security, Deepfake, Image Forensics, Perceptual Hashing, Robustness.

Abstract: Owing to advert of deep learning, deepfake has received considerable attention in this deep learning era. The challenging problem of deepfake detection has been identified to the generalization capability in two aspects: (1) Cross-dataset evaluation and (2) Robustness against content-preserving image manipulations. In this work, we study an image hashing scheme that can be plugged into the existing deepfake detection model to improve their generalization capability. Preliminary experimental results have demonstrates the effectiveness of our perceptual image hashing method.

1 INTRODUCTION

Due to the popularity of internet and social networks, and image editing or generation tools, the fake media or mixup of real and fake contents have been fleed hither and thither. Image tampering can be achieved through splicing, object removal, inpainting, copy-move, and so on (Cozzolino and Verdoliva, 2020). The powerful capability of deep learning technologies and models even worsen this problem as they can generate and synthesize a fake object/image that is indistinguishable from a true one to preserve the perceptually pleasing property. To deal with tampered images and verify the authenticity of images, studies of multimedia security such as data hiding and fingerprinting have given rise to a new wave of study a couple of decades ago. Until recently, AI generative models awaken us the challenges of deepfake detection in this deep learning era (Yan et al., 2023). In this paper, we will focus on image deepfake detection.

A basic principle of image classifier model-based deepfake detection is to use the classifier model as an external force to learn deep features. This principle leads a certain performance in deepfake detection. Nevertheless, the challenges in deepfake detection are recognized as (1) Cross-dataset evaluation: If the method is trained on dataset A, how is the detection performance on the datasets other than A? (2) Robustness: If the fake image is further gone through image manipulations like compression (*e.g.*, JPEG), will the fake clues be eliminated by the compression effect? In this paper, we will address the aforementioned problems.

1.1 Literature Review

Traditional image forensics contains three types of forgeries: Copy-move (copy one or more regions of an image and paste them in the same image with different locations), Splicing (copy one or more regions of an image and paste them on another image), and Inpainting (removal of undesired objects or creation of desired objects).

Nevertheless, in the deep learning era, a new type of forgery, Deepfake, appears. For example, Deepfake can substitute a face of a person with another person to create a fake political or pornography image or video with the fake image quality far better than those generated by non-learning techniques. These crimes cause a severe negative social impact. Hence, a series of studies try to detect Deepfake contents.

In the literature, Zhou *et al.* (Zhou et al., 2018) devised a two-stream Faster R-CNN in that one RGB stream is to find tampered regions like strong contrast difference and unnatural tampered boundaries. Kwon *et al.* (Kwon et al., 2021) proposed CAT-Net to detect and localize image splicing. An RGB and DCT stream is considered to trace the JPEG compression artifacts without losing helpful information from the original RGB view. Yang *et al.* (Yang et al., 2020) proposed Constrained RCNN, which adopts BayarConv (Bayar and Stamm, 2018) as the first convolution layer to create a unified feature representation of various content manipulations. Chen *et al.* (Chen et al., 2021) proposed an MVSS-Net model to detect and localize the tampered regions. Both the edge feature extraction and noise feature extraction modules, together

Lu, C. and Lin, C.
Robust Image Deepfake Detection with Perceptual Hashing.
DOI: 10.5220/0012317800003648

In *Proceedings of the 10th International Conference on Information Systems Security and Privacy (ICISSP 2024)*, pages 374-378
ISBN: 978-989-758-683-5; ISSN: 2184-4356

with a dual attention module, are integrated to build the MVSS-Net. In (Zhao et al., 2021), Zhao *et al.* proposed a multi-attentional deepfake detection mechanism, wherein texture features and global features are first extracted at the shallow and deep layers, respectively, and then fed into an attention module for real/fake image classification. In (Luo et al., 2021), Luo *et al.* proposed to generalize face forgery detection with high-frequency features. The authors observed that a generalizable forgery detector should consider texture-related and texture-irrelevant features and identify the discrepancy between the tampered face and pristine background. They used noise features (SRM) (Fridrich and Kodovsky, 2012) to extract high-frequency features and boost the generalization ability. Sun *et al.* (Sun et al., 2021) proposed meta learning for domain general face forgery detection. Their method, termed learning-to-weight (LTW), contains the meta-test set generated based on the meta-split strategy and meta-optimization for learning a domain-invariant model used in detecting unseen domains. More recently, Sun *et al.* (Sun et al., 2022) proposed dual contrastive learning for general face forgery detection. In their method, the training images are augmented via the data views generation module, and then the intra-instance contrastive learning module and inter-instance contrastive learning module are proposed to learn general features. Hu *et al.* (Hu et al., 2022) proposed the frame inference based detection framework (FInfer) by feeding into two branches of video frames, *i.e.*, source frames and target frames. There are three learning modules in FINfer: the faces representative learning module encodes both the source faces and target faces, the faces predictive learning module predicts the target face representations from the source face representations, and the correlation-based learning module utilizes a representation-prediction loss for training.

1.2 Our Contributions

In this paper, we propose a robust deepfake detection method via perceptual hashing.

- To our knowledge, we are the first to introduce perceptual hash and incorporate it into deepfake detection models. The goal is to achieve not only the detection of fake images but also the resilience against content-preserving image manipulations, including compressions, online social networking (OSN) (Wu et al., 2022) processing, etc.
- The proposed perceptual hashing mechanism is a plug and play module to enhance to robustness of existing deepfake detection methods.

2 PROBLEM SETUP

In this work, deepfake detection is casted as a binary classification problem, where the true image is labeled "0" and fake one is labeled "1". We define a "fake" image as the one obtained from the true image by content-changing manipulations to change the contents globally or locally, including face swapping, splicing, and so on. On the other hand, if the image is edited via content-preserving manipulations, including compressions, blurring, sharpening, and so on, without changing its meaning, the resulting version is still regarded as "authentic."

In addition to distinguishing from true and fake images, a more challenging problem is encountered when a fake image is further gone through content-preserving manipulations. Specifically, the problems are (1) if a JPEG compressed fake image can still be detected to be falsified? and (2) if a JPEG compressed real image can be detected to be authentic? In particular, for Problem (1), we concern if the fake clues will be eliminated or destroyed by the JPEG compression effect, while for Problem (2), we concern if the JPEG effect is regarded as the fake clue so as to judge the compressed image to be inauthentic.

Therefore, a practical deepfake detection method should satisfy two requirements in that content-changing modifications are detected to be fake and content-preserving manipulations are treated to be authentic.

In addition to robustness, it is highly possible that a detection model is trained on dataset A but will be later tested on datasets other than A. This is referred to as the cross-dataset evaluation problem.

The goal of this paper is to develop a robust deepfake detection method.

3 PROPOSED METHOD

In this paper, we propose an image perceptual hashing method together with a deepfake detection model to deal with robust deepfake detection wherein a fake image, even being involved with content-preserving modifications (*e.g.*, JPEG), still can be detected but a real image manipulated with content-preserving processing can be tolerated. This image perceptual hashing method can be further extended to deal with cross-dataset evaluation. Our method can be incorporated with any deep learning model for the deepfake detection task. Since deepfake detection is casted as a binary classification problem in this paper, in addition to the conventional cross-entropy loss, we will introduce the so-called hash-preserving loss in the following.

3.1 Image Perceptual Hashing

Image Perceptual Hashing (abbreviated as IPH hereafter) has been studied based on hand-crafted features (Lu and Hsu, 2005)(Swaminathan et al., 2006) a couple of decades ago. Here we will develop an IPH method in terms of deep learning features. For ease of descriptions, let F denote the feature vector obtained from the last convolution layer of a learning model from which a corresponding hash vector/code H will be generated. The hash is designed to be a bipolar vector as:

$$H(i) = torch.sign(F(i) - torch.median(F(i))), \quad (1)$$

where $1 \leq i \leq L$ and L denotes the length of F (and H), and $H \in \{-1,+1\}^L$. We select the feature map of the last convolution layer for hash generation as it is a kind of low-frequency features, which is considered to be a stable feature without being easily affected by noises.

It can be been from Eq. (1) that the hash vector H is designed to have half 1's and half -1's, where hash bit 1 is defined when $F(i)$ is larger than the median value; otherwise, hash bit is -1. Such a design is common and traditional in the literature and has been verified to provide a kind of robust feature in reflecting content-changing modifications and resisting content-preserving manipulations.

In the following, we will describe how to combine the perceptual hashing with a deepfake detection model to deal with the two challenges mentioned in Sec. 1.

3.2 Deepfake Detection with Perceptual Hashing

We denote the image I as an input to the network f, where f can be any backbone network model used in a deepfake detection method. Let F_I denote the feature map of the last convolution layer of f by feeding I to f, and let H_I denote its hash sequence.

In the deepfake detection scenario, it is assumed that we will have four kinds of images for training: true image (TI), fake image (FI), JPEG true image (JTI), and JPEG fake image (JFI). Since we cannot expect to have unlimited numbers or types of images for training, we will use JPEG to represent the content-preserving manipulations, and the JPEG compressed images are used as a kind of data augmentation. Thus, in addition to the true and fake images used for training as usual, both JPEG true and JPEG fake images will be used as the augmented data for calculating the hash-preserving loss during training.

Specifically, the true image and its corresponding JPEG compressed version are considered to be authentic and, thus, their hashes should be similar (Case 1). On the contrary, any pair of images, where one from true or JPEG true image and another from fake or JPEG fake image will be treated to be different as the latter contains fake parts. Thus, such a pair of images will have dissimilar hashes (Case 2). In the literature with non-learning paradigm, one usually employs the Hamming distance to measure the similar between two hash codes. Such a distance measure, however, is inconsistent with the form of cross-entropy loss for the classification task, and is not appropriate for training. In our method, we will transform the conventional hash loss in terms of Hamming distance to a probability form such that it can be jointly combined with cross-entropy loss for (true/fake) image classification. It is said that two images with similar hashes will be classified to the same class (either true or fake), and those with dissimilar hash codes will belong to different classes.

More specifically, let H_1 and H_2 be the hash codes with respect to images I_1 and I_2 obtained from Eq. (1), let I_1 and $I_2 \in \{TI, FI, JTI, JFI\}$, and let $< H_1, H_2 >$ denote their inner product. We refer to (Xia et al., 2021) to relate $< H_1, H_2 >$ with classification probability as:

$$p(S_{1,2}|H_1,H_2) = \begin{cases} \delta(< H_1,H_2 >), S_{1,2} = 1 \\ 1 - \delta(< H_1,H_2 >), S_{1,2} = 0, \end{cases} \quad (2)$$

where

$$\delta(< H_1,H_2 >) = \frac{1}{1 + exp(- < H_1,H_2 >)} \quad (3)$$

and $\delta(\cdot)$ denotes a sigmoid function. In Eq. (2), $S_{1,2} = 1$ indicates images I_1 and I_2 belong to the same class; otherwise $S_{1,2} = 0$. To apply the characteristic of hashing in deepfake detection, it should be noted that Eq. (2) is realized in a batch depending on the class labels of a pair of image. Actually, in addition to Case 1 and Case 2, there is one case that will be excluded in our implementation. This case contains the images in a batch that share the same label but are not from the same origin as in Case 1. This is because, despite having the same label, these images are fundamentally different and their hash values should not be forced to be similar.

Thus, the loss function of coding consistency between $\delta(< H_1,H_2 >)$ and $S_{1,2}$ is defined as (Xia et al., 2021):

$$\begin{aligned} L_{cc} = \; & - \; S_{1,2} log \delta(< H_1,H_2 >) \\ & - \; (1 - S_{1,2})(1 - log \delta(< H_1,H_2 >)). \end{aligned} \quad (4)$$

3.3 Overall Loss

Suppose an existing deepfake detection model is selected as the baseline model, which is incorporated with our proposed Image Perceptual Hashing for deepfake detection, with the loss function being denoted as L_M. Now, the total loss function here will be

$$L_{total} = (1-\gamma)L_M + \gamma L_{cc}, \quad (5)$$

where the weight γ is used to strike a balance, ensuring optimal improvement in the model's generalization capability while not significantly compromising model performance. As an example, please refer to (Zhao et al., 2021; Chen et al., 2022) for the detail of L_M.

4 EXPERIMENTS

To evaluate the effectiveness of our perceptual hashing method as a plug-and-play module in robustly detecting deepfake images, two state-of-the-art Deepfake detection models, namely MADD (Zhao et al., 2021) and SLADD (Chen et al., 2022), were selected for experiments as the authors released codes for fair comparison. Three popular datasets, including FaceForensics++ (Rossler et al., 2019), Celeb-DF (Li et al., 2020), and Deepfake Detection Challenge (DFDC) (Dolhansky et al., 2020), were used.

4.1 Datasets

FaceForensics++ is a dataset that comprises videos at three distinct compression levels: RAW, High Quality (HQ), and Low Quality (LQ). At each compression level, there is in total 1,000 videos, with 720, 140, and 140 videos being allocated for training, validation, and testing, respectively. For Celeb-DF, it consists of 408 real videos and 795 synthesized videos. For Deepfake Detection Challenge (DFDC), it is the most recently released large scale deepfake detection dataset, which includes over 1,000 real and 4,000 fake videos manipulated by multiple Deepfake, GAN-based, and non-learning methods.

4.2 Robust Deepfake Detection Against Content-Preserving Image Manipulations

So far, the state-of-the-art deepfake detection models have demonstrated their performance on standard deepfake datasets, including FaceForensics++, DFDC, and Celeb-DF. Nevertheless, when the detection models are subjected to testing using images that have undergone image content-preserving manipulations, such as compression or transmission through online social networks, the detect performance will be degraded.

In this paper, in addition to conducting cross-dataset evaluation, we take both JPEG compression and online social network (OSN) processing (Wu et al., 2022) into consideration as representative of content-preserving image manipulations to verify the robustness of our proposed deepfake detection method.

4.3 Evaluation Results

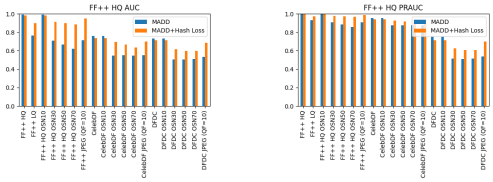

Figure 1: Trained on FF++ HQ and evaluated in terms of (a) AUC and (b) PRAUC, where the X-axis shows the (manipulated) datasets used for testing.

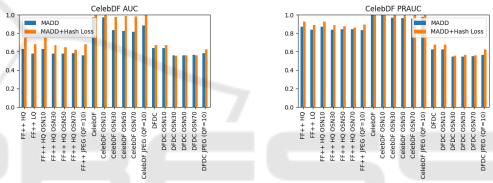

Figure 2: Trained on CelebDF and evaluated in terms of (a) AUC and (b) PRAUC, where the X-axis shows the (manipulated) datasets used for testing.

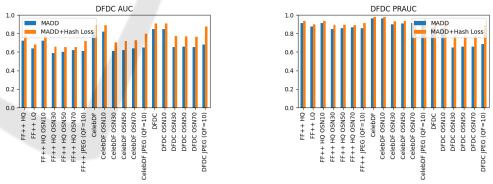

Figure 3: Trained on DFDC and evaluated in terms of (a) AUC and (b) PRAUC, where the X-axis shows the (manipulated) datasets used for testing.

Our method was evaluated on three cases, that is, the model was trained individually on one of the datasets, FF++, Celeb-DF, and DFDC, but tested on each of them. During testing on deepfake detection, each dataset will also be processed through JPEG compression with Quality Factor (QF=10) and OSN (Online Social Network) processing with a few different parameters, denoted as OSN10, OSN30, OSN50, and OSN70, to imply different degrees of manipulations.

Both AUC (Area Under the Curve) and PRAUC (Precision-Recall Area Under the Curve) were used as the evaluation metrics. These metrics are known for their objectivity and reliability in assessing model per-

formance, whether in typical scenarios or in situations involving imbalanced datasets.

The deepfake detection results were shown in Fig. 1, Fig. 2, and Fig. 3 (best viewed in a color display), where the blue and orange bars indicate the results obtained from MADD and our method (MADD+hash loss), respectively. It can be observed that the orange bars are generally higher than blue bars, indicating that the proposed hashing is efficient in improving the generalization capability of MADD not only in resisting content-preserving manipulations, including JPEG and OSN attacks, but also in dealing with cross-dataset detection. In particular, the performance gap between the original MADD and MADD+our hashing is large remarkably in several cases. Although it is not shown here, we have also observed similar results for SLADD trained on FF++.

5 CONCLUSIONS

In this paper, we have presented a perceptual image hashing method that can be plugged into the existing deepfake detection models to boost their performance in resisting content-preserving image manipulations in that the fake clues can be properly reserved under JPEG compression and online social network processing. The preliminary experimental results demonstrate the effectiveness of proposed perceptual hashing. In the future, we will further study and apply the idea of perceptual hashing in other deepfake detection models.

ACKNOWLEDGEMENT

This work was supported by the National Science and Technology Council (NSTC), Taiwan, ROC, under Grants NSTC 112-2221-E-001-011-MY2 and 112-2634-F-001-002-MBK. We also thank Taiwan Cloud Computing (TWCC) for providing computational and storage resources.

REFERENCES

Bayar, B. and Stamm, M. C. (2018). Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*.

Chen, L., Zhang, Y., Song, Y., Liu, L., and Wang, J. (2022). Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*.

Chen, X., Dong, C., Ji, J., Cao, J., and Li, X. (2021). Image manipulation detection by multi-view multi-scale supervision. In *ICCV*.

Cozzolino, D. and Verdoliva, L. (2020). Noiseprint: A cnn-based camera model fingerprint. *IEEE Trans. on Information Forensics and Security*, 20.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. (2020). The deepfake detection challenge dataset. In *arXiv preprint arXiv:2006.07397*.

Fridrich, J. and Kodovsky, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on information Forensics and Security*.

Hu, J., Liao, X., Liang, J., Zhou, W., and Qin, Z. (2022). Finfer: Frame inference-based deepfake detection for high-visual-quality videos. In *AAAI*.

Kwon, M.-J., Yu, I.-J., Nam, S.-H., and Lee, H.-K. (2021). Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *WACV*.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*.

Lu, C.-S. and Hsu, C.-Y. (2005). Geometric distortion-resilient image hashing scheme and its applications on copy detection and authentication. *ACM Multimedia Systems Journal, special issue on Multimedia and Security*, 11(2).

Luo, Y., Zhang, Y., Yan, J., and Liu, W. (2021). Generalizing face forgery detection with high-frequency features. In *CVPR*.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *ICCV*.

Sun, K., Liu, H., Ye, Q., Gao, Y., Liu, J., Shao, L., and Ji, R. (2021). Domain general face forgery detection by learning to weight. In *AAAI*.

Sun, K., Yao, T., Chen, S., Ding, S., Li, J., and Ji, R. (2022). Dual contrastive learning for general face forgery detection. In *AAAI*.

Swaminathan, A., Mao, Y., and Wu, M. (2006). Robust and secure image hashing. *IEEE Trans. Information Forensics and Security*, 1(2).

Wu, H., Zhou, J., Tian, J., and Liu, J. (2022). Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xia, H., Jing, T., Chen, C., and Ding, Z. (2021). Semi-supervised domain adaptive retrieval via discriminative hashing learning. In *Proceedings of ACM Multimedia*.

Yan, Z., Zhang, Y., Fan, Y., and Wu, B. (2023). Ucf: Uncovering common features for generalizable deepfake detection. In *ICCV*.

Yang, C., Li, H., Lin, F., Jiang, B., and Zhao, H. (2020). Constrained r-cnn: A general image manipulation detection model. In *ICME*.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. (2021). Multi-attentional deepfake detection. In *CVPR*.

Zhou, P., Han, X., Morariu, V. I., and Davis, L. S. (2018). Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.