Efficient-Frequency: a hybrid visual forensic framework for facial forgery detection

Chau Xuan Truong Du
Griffith University
Gold Coast, Australia
x.chau@griffith.edu.au

Le Hoang Duong
Hanoi University of Science and Technology
Hanoi, Vietnam
duonglhbk@gmail.com

Huynh Thanh Trung
Griffith University
Gold Coast, Australia
thanhtrunghuynh93@gmail.com

Pham Minh Tam
Hanoi University of Science and Technology
Hanoi, Vietnam
pminhtamnb@gmail.com

Nguyen Quoc Viet Hung
Griffith University
Gold Coast, Australia
quocviethung1@gmail.com

Jun Jo
Griffith University
Gold Coast, Australia
j.jo@griffith.edu.au

Tam Nguyen
HCMUT
Hochiminh, Vietnam
nt.tam88@hutech.edu.vu

***Abstract*—The recent years have witnessed the significant development of visual forgery techniques and their malicious applications such as spreading of fake news and rumours, defamation or blackmailing of politicians and celebrities, manipulation of election result in political warfare. The manipulated contents have reached to such sophisticated level that human cannot tell apart whether a given content is real or fake. To deal with this serious threat, a rich body of visual forensic techniques has been proposed for detecting forged video and images. However, existing techniques either rely solely on engineered features or require a complex deep learning model to extract the underlying patterns. In this paper, we propose a novel end-to-end visual forensic framework that can incorporate different modalities to efficiently classify real and forged contents. The model employs both the original content and its frequency domain analysis to fully exploit the richness of the image latent patterns. They are forwarded into two separated EfficientNet, a light yet efficient neural network architecture specialized for image classification, for pattern extraction. Then, we design a late-fusion mechanism to fuse the learnt features in original and frequency domain based on the importance of the underlying information. Our experimental results show that our proposed technique outperforms other state-of-the-art forensic approaches in many datasets and being robust to various visual forgery techniques.**

***Index Terms*—visual forensic, visual forgery, fake image detection**

I. Introduction

Forged facial visual content including images and videos has seen the significant improvement in quality with the development of advanced forgery techniques using sophisticated neural network architecture along with large amount of training data. The quality of these forged contents has reached to such level that even people with ideal vision in good lighting condition cannot tell apart whether the given content is real or fake [1]. Furthermore, the direct usage of deep learning based forgery techniques as a blackbox with the support of pretrained model significantly simplifies the creation of facial forgeries [2]. For example, a person without expert knowledge can produce a forged facial image using several images or one short video [3] of the target person. The high quality and straightforward usage of facial forgery techniques leads to the emerge of various malicious applications such as fake news and rumours for click-baits, defaming and impersonation of celebrities and politicians, fraud transactions and activities [4]–[12].

Given the dangerous threat of visual forgery, there is a plethora of emerging studies on visual forensics. *Traditional techniques* rely on handcrafted features to detect anomaly patterns of visual contents such as frequency analysis, head pose, facial details (a.k.a artifacts) [13]–[16]. Despite the simplicity yet effectiveness of such techniques, they often focus on a single or limited number of features. Thus, their performance is unstable and susceptible to several type of attacks, especially the graphic-based forgery techniques (e.g. FaceSwap [17], [18], 3D-Morphable face model [19]) which prioritize the maintenance of facial features. Also, the design of such techniques require a considerable workload of experts in handcrafting the features. For example, Visual-Artifacts [14] needs the tailoring of three different features for each type of common attacks such as Deepfake [20] and Face2Face [3].

Recently, the advance of modern deep neural network and its usage in visual forgery techniques also leads to the emerge of *deep learning forensic approach*. These techniques leverage deep neural models to automatically extract latent features that go beyond human perception to classify real and forged visual contents. These techniques vary in the network architectures, such as *Mesonet* [21], *GAN-fingerprint* [22], *Capsule* [2] and *Xception* [23]. However, in overall, existing deep learning forensic techniques require a sophisticated model to efficiently learn the underlying patterns. For example, the state-of-the-art Xception model contains around 21 millions parameters. Also, these techniques fail to utilize the expert knowledge features.

To address the challenges above, in this work we propose a novel end-to-end visual forensic framework. Our main idea is to go beyond the existing forensic methods by simultaneously analyze the input content (image or video frame) in both original and frequency domain. Our framework then utilizes two separated EfficientNet, a state-of-the-art neural network model that balances the complexity of the architecture and the efficiency in pattern extraction. The learnt features from original and frequency domain are then unified by a late-fusion mechanism which considers the importance of the underlying information. The salient contributions of our work are summarized as follows:

- We propose **Efficient-Frequency**, an end-to-end neural network based visual forensic framework that investigates at the same time the visual content in original and frequency domain. We are the first to integrate handcrafted features into a neural network based forensic model. This enables us to incorporate rich information available in visual contents regardless of its modality.
- We utilize the Discrete Fourier Transformation to learn the frequency-domain representation of the input visual contents in an unsupervised and automatic manner. Then our proposed model employs two separated EfficientNet [24] to capture the underlying patterns in original contents and their frequency-based representation. Our work is the pioneer to use this simple yet powerful network architecture in visual forensic.
- We design a late-fusion mechanism to combine learnt features based on the importance of the information type to the visual contents. This makes our model adaptive to various forged datasets as well as forgery techniques.
- We evaluate our model in several experiments several collected real-world datasets and synthesized datasets using different visual forgery techniques. The empirical result shows that our framework not only achieves better results compared to baseline techniques but also exhibits robustness to various type of visual forgeries.

The rest of the paper is organised as follows. Section II introduces the related works and the formal problem formulation. Section III presents the details of the proposed visual forensic model, including the overview and its main components. The experimental setup and results are provided in Section IV while Section V concludes the paper.

II. Related Work

We here introduce briefly the relevant works, including the typical visual forgery and visual forensic techniques.

A. Visual Forgery Techniques

The visual forgery techniques often manipulate the visual content by injecting false information into an original image, such as the identity of the person or his emotion. The forgery techniques can be organized into two categories, namely *graphic-based techniques* and *neural network based techniques* [25]–[32].

The graphic-based techniques often leverage handcrafted features such as landmark points of human face to swap the face of the target person into the source image, following by a color adjustment step. For example, FaceSwap-2D [17] first performs the histogram matching of the input images, then employs the system of landmark points in two-dimensional setting to fit the face of the target person into the source image. FaceSwap-3D [18] goes beyond Faceswap-2D by leveraging the landmark points in three-dimensional setting instead. 3D Morphable Model (3DMM) [19] also learns the shape and the texture of the person face in three-dimensional scale, but uses a nonlinear mapping to project the 3D face to 2D plane learnt by a encoder-decoder deep neural network.

On the other hand, the neural network based technique leverages advanced neural network architectures (e.g. generative adversarial network (GAN) [33], variational autoencoder (VAE) [34]) to produce high quality forged visual contents. For instance, Deepfakes [20] leverages the autoencoder network to replace the face of person in the original image by the target person. The recent Deepfake variants such as Faceswap-GAN [35] inspires from GAN model to treat the decoder-encoder network as a generator and add two discriminator to enhance the quality of generated contents. ReenactGAN [36] is another GAN-based model that aims to transfer the facial movement and expression of the target person into the source image using facial contours aware latent space. StarGAN [37] is a GAN-based model that generate the forged contents by modifying the attributes (e.g. hair color, gender, facial expression) of the person in the original image.

B. Visual Forensic Techniques

As the rival of visual forgery techniques, the visual forensic techniques aim to detect the manipulated contents. The forensic techniques can also be organized into two categories: the *traditional techniques* and the *neural network based techniques* [38], [39].

The traditional techniques often rely on the engineered features (e.g. frequency analysis, head pose, facial details) to detect the anomaly patterns of the forged contents. Frequency domain-based detector (FDBD) [13] exploits the frequency characteristics of the input image to discover anomaly patterns. Visual-Artifacts [14] relies on several visual artifacts (e.g. eyes, teeth, facial contours) that appears in the forged contents generated by common facial forgery techniques such as DeepFake and FaceSwap. Head pose-based detector (HPBD) [15] focuses on the unstable head pose of the forged facial contents.

The neural network based techniques adopt the advanced neural network architectures to automatically extract the important hidden features which helps to distinguish the real and fake images. Mesonet [21] analyzes the contents at a mesoscopic level using a simple neural network with small number of layers. Capsule [2] leverages the "capsule" architecture [40] with expectation maximization routing algorithm to better capture the anomaly patterns in forged images and videos. XceptionNet [23] utilizes the Inception network [41] to extract the underlying features that helps to distinguish between fake and real images. GAN-fingerprint [22] detects the forged contents by finding the fingerprint of GAN [33], the key architecture at the heart of state-of-the-art forgery techniques such as DeepFake and StarGAN. Our work goes beyond the existing techniques by integrating the engineered feature (frequency analysis) along with original image into a

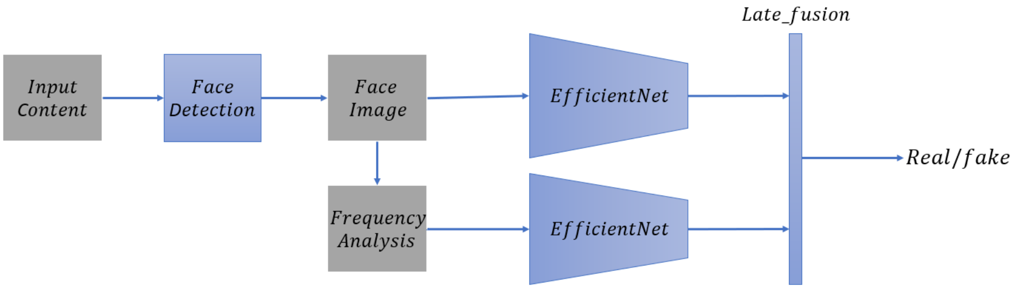

Fig. 1: Overview of Efficient-Frequency framework

neural network model that minimizes the complexity while guarantees the efficiency of the anomaly pattern extraction.

III. Efficient-Frequency forensic framework

Figure 1 depicts the overview of our framework. First, the face is detected from the input images or video frames using face detection module. The extracted face image then is analyzed in the frequency domain using Fourier Transform. The original image and its frequency-domain representation then are forwarded into two separated EfficientNet models to learn the underlying features. The learnt features are combined by a late-fusion mechanism which considers the importance of the information, then forwarded to a fully connected layer and the common cross entropy loss for binary classification.

Face Detection. We use Multi-task Cascaded Convolutional Networks (MTCNN) [42] to detect face from given images or video frames. The framework pipeline contains three main steps. In the first step, the given image is rescaled to a range of different sizes (a.k.a image pyramid), then a shallow fully convolutional network (so called P-Net) is employed to produce the candidate windows. In the second step, a more complex CNN model namely R-Net is adopted to refine the window candidates and keep only the high potential ones. In the last step, a powerful CNN model namely O-Net further refines the candidates and locates the facial landmarks positions. Between the steps, non-maximum suppression (NMS) is used to filter the candidate bounding boxes. The detailed implementation of P-Net, R-Net and O-Net can be found in [42].

Frequency Analysis. We utilize the Discrete Fourier Transform (DFT) to obtain the frequency-domain presentation of the input image. The presentation can be considered as a spectral decomposition of the image that indicates the distribution of its energy over a range of frequencies given the spatial resolution. For 2-dimensional image data of size $M \times N$, the Fourier Transform can be computed as:

$$X_{k,l} = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} x_{n,m} . e^{-\frac{i2\pi}{N}kn} . e^{-\frac{i2\pi}{M}lm} \tag{1}$$

where $X_{k,l}$ is value at position (k, l) in spectrum image and $x_{n,m}$ is value at position (n, m) in original image. It is worth noting that the obtained representation inherits the same dimensionality from the input signal, which is 2-dimensional for image. Therefore, an azimuthal averaging [43] is applied to flatten the representation into 1-dimensional form. The transformation can be considered as a compression where similar frequency components are gathered and averaged into a feature vector. The compression helps to significantly reduce the amount of features with the minimized loss of information, results in a more robust representation of the input image.

Efficient Net. Our model employs EfficientNet [24], a convolutional network architecture that performs better pattern extraction while guarantees the efficiency of the model. The concept of the model is designed using a multi-objective neural architecture search that optimizes the two mentioned criteria, accuracy and efficiency. Our model leverages the original EfficientNet variant, namely EfficientNet-B0, as this variant can capture better facial detail information comparing to other scaled-up variants such as EfficientNet-B1, where the details can be washed out due to over deep architecture. The detailed implementation of EfficientNet is shown in Figure 2.

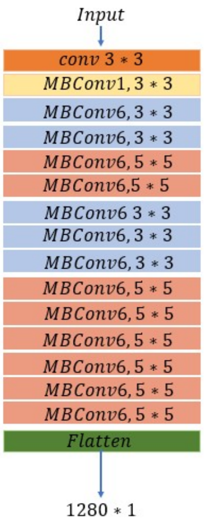

Fig. 2: EfficientNet architecture

The network starts with a convolutional layer of the size 3×3, which performs lightweight filtering. Then, the network continues with multiple stacked mobile inverted bottleneck (MBConv) layers [44]. The MBConv layer is the main building block of the EfficientNet, which contains a point-wise (1x1) convolution, a depth-wise convolution and a spatially-filtered feature map. Instead of standard convolution, MBConv uses a depth-wise separable layer to reduce computational cost while guarantees the quality of pattern extraction. The last layer of a MBConv block is a spatially-filtered feature map which projects information from previous layer back to a low-dimensional subspace using another point-wise convolution. The low-dimensional subspace helps to preserve the essential information while reduces the complexity of the model. For each MBConv block, a residual connection is added to aid gradient flow during backpropagation. After the stacked MBConv layers, the output feature is flatten into 1-dimensional vector using a fully connected layer at the end of the network.

Late-fusion mechanism. As discussed, we use two separated EfficientNet to extract the patterns from original image and its frequency-domain representation, results in two 1-dimensional feature vector X_o and X_f, respectively. We here combine the learnt features using following mechanism:

$$X = \Theta(\alpha . X_o, (1 - \alpha) . X_f) \tag{2}$$

where α is the learnable parameter that balances the importance between original and frequency-aware patterns; and Θ is the aggregate function. For our work, we choose Θ as concatenation over other common functions such as sum operator to avoid the information loss.

Loss Function. To train the whole network, we utilize binary cross entropy loss. This is a common loss function for binary classification problem, which is well-suited for our forgery detection problem. The loss function is computed as follows:

$$\mathbf{L} = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i) \tag{3}$$

where N is the number of images in the training set; $\hat{y}_i$ is the label of the i-th training image which is 1 for forged image and 0 for genuine image; $y_i = \sigma(X_i)$ is the scalar output of the model computed by the sigmoid function σ applying to the unified learnt feature X_i of the i-th training sample.

IV. Experiments

In this section, we conduct experiments with the aim of answering the following research questions:

(RQ1) Does our model outperform other forensic methods on real-world forged datasets?
(RQ2) How does our model performs against different types of facial forgery techniques?
(RQ3) How important is each design choice of our model?

A. Experimental Setting

Datasets. We employ the following datasets to assess the performance of visual forensic techniques:

- Deepfake-in-the-wild [45]. The Deepfake-in-the-wild dataset contains 7,314 face sequences extracted from 707 deepfake videos collected from the internet.
- Celeb-DF [46]. The Celeb-DF dataset is contains 590 real short interview videos and 5,639 DeepFake videos.
- DFDC [47]. The DFDC dataset contains 5,244 videos with diverse background, lighting and head poses.
- UADFV [15]. This dataset contains 49 real short videos 11.14 seconds average length and 49 DeepFake videos.
- DF-TIMIT [48]. The DF-TIMIT dataset contains 430 original videos and forged video generated using GAN-based face-swapping algorithm.
- FaceForensics++ [1]. The dataset includes 1,000 real YouTube videos and forged videos generated by Face2Face, FaceSwap, DeepFakes, NeuralTextures

Baselines. We compare the performance of our framework with the following representative baselines in the literature:

- *Spectrum1D* [13] investigates the frequency characteristics to discover anomaly patterns of forged contents.
- *Visual-Artifacts* [14] exploits anomaly patterns (e.g. eyes, facial contours) that often appear in the forged contents.
- *HPBD* [15] detects forged facial contents based on the unstable head pose.
- *Mesonet* [21] analyzes the contents at a mesoscopic level using a simplified convolution neural network.
- *Capsule* [2] leverages the ”capsule” architecture [40] to capture the anomaly patterns in forged contents.
- *XceptionNet* [23] utilizes the Inception network [41] to extract the underlying features to detect forged contents.
- *GAN-fingerprint* [22] detects the forged contents by finding the fingerprint of GAN [33].

On the other hand, we also evaluate the performance of the forensic techniques against following forgery techniques:

- *FaceSwap-2D* [17] employs the 2-D landmark points to inject the face of the target person into the source image.
- *FaceSwap-3D* [18] uses the 3-D landmark points to swap the face of the input images.
- *3DMM* [19] learns the shape and the texture of the input faces in 3-D setting using a nonlinear mapping.
- *Deepfakes* [20] leverages the autoencoder network, enhanced by GAN to impersonate the target person.
- *ReenactGAN* [36] transfers the facial movement and expression of the target person into the source image.
- *StarGAN* [37] is a GAN-based model that is able to modify the facial attributes (e.g. hair color, gender).
- *Monkey-Net* [49] is a GAN-based model that can inject the action of target person into the source image.

Reproducibility environment. The model is implemented in Python v3.6 using the Keras v2.2.4 API. The Keras API is a high-level neural networks API focusing on enabling fast

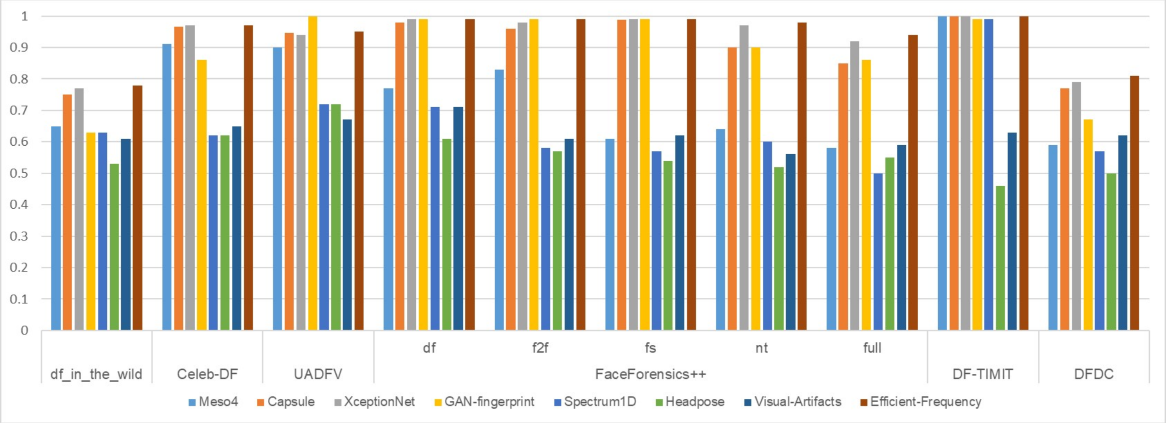

Fig. 3: End-to-end comparison on real-world datasets

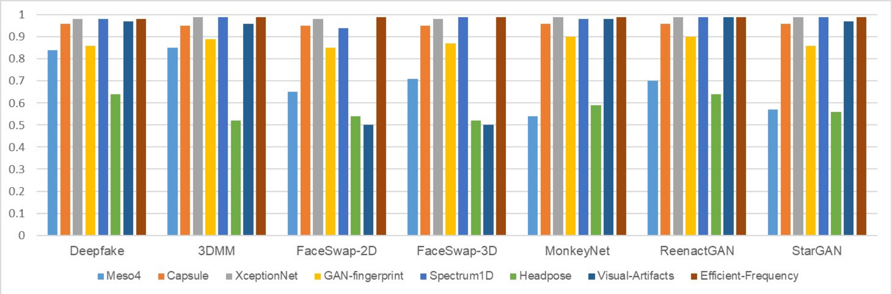

Fig. 4: End-to-end comparison against different forgery techniques

TABLE I: Model size and processing speed

	Number of parameter	Processing Speed (s/10,000 images)
Spectrum1D	N/A	245
HPBD	N/A	2400
Visual-Artifact	N/A	9100
Mesonet4	28073	57
Capsule	3895998	28
XceptionNet	21861673	24
GAN-fingerprint	14252563	243
Efficient-Frequency	8017081	27

TABLE II: Ablation test result on real-world datasets

	df_in_the_wild	Celeb-DF	UADFV	FaceForensics++					DF-TIMIT	DFDC
				df	f2f	fs	nt	full		
Efficient-Frequency-1	0.72	0.95	0.94	0.97	0.94	0.97	0.83	0.81	1	0.81
Efficient-Frequency-2	0.64	0.69	0.5	0.59	0.59	0.6	0.62	0.57	0.99	0.58
Full model	0.79	0.97	0.95	0.99	0.99	0.99	0.98	0.94	1	0.81

TABLE III: Ablation test result on visual forgery techniques

	Deepfake	3DMM	FaceSwap-2D	FaceSwap-3D	MonkeyNet	ReenactGAN	StarGAN
Efficient-Frequency-1	0.91	0.95	0.95	0.94	0.95	0.95	0.95
Efficient-Frequency-2	0.95	0.95	0.78	0.82	0.97	0.95	0.95
Full model	0.98	0.99	0.99	0.99	0.99	0.99	0.99

experimentation. The techniques are trained and tested on 2 × GPU GeForce GTX 2080Ti, CPU AMD Ryzen Threadripper 1900X 8-Core Processor and 64 GB RAM. Results are averaged over 10 runs to avoid the randomness.

B. End-to-end comparison on real-world datasets

We answer (RQ1) by comparing the performance of our proposed technique and other baselines on various real-world datasets. The experimental results are shown in Figure 3. In overall, our technique outperforms other techniques on all the datasets except *UADFV* comparing to other forensic techniques. Our technique achieves around 0.8 of accuracy for two challenging datasets *df_in_the_wild* and *DFDC*; and greater than 0.9 of accuracy for the rest 8 datasets. Among the baselines, Xception performs the best and even performs equally to our technique in some dataset such as *Celeb-DF* and *DF-TIMIT*, the datasets that focuses on DeepFake forgery technique. GAN-fingerprint shows considerable results overall and outperforms other techniques on *UADFV* dataset, a small dataset using GAN-based forgery technique. Mesonet shows similar level of performance to GAN-fingerprint, while Capsule gives slightly better result. The traditional techniques such as Spectrum1D, HPBD and Visual-Artifacts perform not as good as the above neural network based techniques, with the accuracy of just around 0.7.

When it comes to the processing speed and model size, the experiment results are shown in Table I. Our proposed technique is one of the fastest techiques along with Capsule and XceptionNet, with less than 30 seconds per 10,000 images. On the other hand, the processing speed of traditional techniques such as HPBD and Visual-Artifacts are much slower, with 2,400 and 9,100 seconds of processing time per 10,000 images. This is because these techniques require a large amount of time to extract the engineered features from the input images. For model size, our model uses significant less parameters than other neural network based model such as XceptionNet and GAN-fingerprint while achieves better accuracy.

C. End-to-end comparison against forgery techniques

We answer (RQ2) by comparing the performance of our proposed technique and other baselines against different type of forgery techniques. The experimental results are shown in Figure 4. It can be seen that our technique Efficient-Frequency performs the best among all forensic techniques. Efficient-Frequency achieves greater than 0.95 of accuracy against all type of forgery. For the baselines, in general, the neural network based techniques such as Xception, Capsule, Mesonet and GAN-fingerprint perform better than traditional techniques such as Spectrum1D, HPBD and Visual-Artifacts. XceptionNet performs slightly behind our technique, achieves greater than 90% of accuracy against all forgery techniques. The performance of Capsule is quite similar to XceptionNet, while GAN-fingerprint work surprisingly well even for non GAN-based techniques, thanks to the in-depth analysis on both image and model levels. Mesonet performs considerably inferior comparing to the above neural network based techniques. For traditional techniques, it is interesting that Visual-Artifacts, HPBD and Spectrum1D performs better on sophisticated neural network based visual forgery techniques such as Deepfake and StarGAN than the traditional techniques.

D. Ablation Test

We evaluate the design choices of our model by comparing it with various variants (RQ3):

- *Efficient-Frequency-1:* uses only the original image and removes the frequency-domain representation.
- *Efficient-Frequency-2:* uses only the frequency-domain representation and discards the original image.

Table II and table III compares the performance of the variants and the full model on forged datasets and against visual forgery techniques. It can be seen from the tables that the full model outperforms the variants on all the datasets and against all forgery type, which confirms the efficiency of the simultaneous usage of both original image and its frequency analysis attribute information.

V. Conclusion

We propose a novel visual forensic framework that aims to detect the forged facial visual contents including images and videos. The key idea of the framework is to analyze simultaneously the raw content and its frequency-domain representation using two separated EfficientNet, a neural network architecture that well balances between accuracy and efficiency. The learnt features then are combined using a late-fusion mechanism, which considers the importance of the underlying information. Extensive experiments reveal that our proposed technique not only outperforms other state-of-the-art forensic approaches in many real-world forged datasets but also show significant robustness against different visual forgery techniques.

References

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: ICCV, 2019, pp. 1–11.

[2] H. H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: ICASSP, IEEE, 2019, pp. 2307–2311.

[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. 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.

[4] W. Wang, H. Yin, Z. Huang, Q. Wang, X. Du, Q. V. H. Nguyen, Streaming ranking based recommender systems, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 525–534.

[5] N. T. Tam, M. Weidlich, D. C. Thang, H. Yin, N. Q. V. Hung, Retaining data from streams of social platforms with minimal regret, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 2850–2856.

[6] N. T. Toan, P. T. Cong, N. Q. V. Hung, J. Jo, A deep learning approach for early wildfire detection from hyperspectral satellite images, in: 2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA), 2019, pp. 38–45.

[7] Q. Wang, H. Yin, H. Wang, Q. V. H. Nguyen, Z. Huang, L. Cui, Enhancing collaborative filtering with generative augmentation, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 548–556.

[8] N. T. Tam, M. Weidlich, B. Zheng, H. Yin, N. Q. V. Hung, B. Stantic, From anomaly detection to rumour detection using data streams of social platforms, Proceedings of the VLDB Endowment 12 (9) (2019) 1016–1029.

[9] C. T. Duong, T. D. Hoang, H. T. H. Dang, Q. V. H. Nguyen, K. Aberer, On node features for graph neural networks, arXiv preprint arXiv:1911.08795 (2019).

[10] K. Sun, T. Qian, T. Chen, Y. Liang, Q. V. H. Nguyen, H. Yin, Where to go next: Modeling long-and short-term user preferences for point-of-interest recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 214–221.

[11] S. Zhang, H. Yin, Q. Wang, T. Chen, H. Chen, Q. V. H. Nguyen, Inferring substitutable products with deep network embedding., in: IJCAI, 2019, pp. 4306–4312.

[12] T. Chen, H. Yin, Q. V. H. Nguyen, W.-C. Peng, X. Li, X. Zhou, Sequence-aware factorization machines for temporal predictive analytics, in: 2020 IEEE 36th International Conference on Data Engineering (ICDE), 2020, pp. 1405–1416.

[13] R. Durall, M. Keuper, F.-J. Pfreundt, J. Keuper, Unmasking deepfakes with simple features, arXiv preprint arXiv:1911.00686 (2019).

[14] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: WACVW, IEEE, 2019, pp. 83–92.

[15] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP, IEEE, 2019, pp. 8261–8265.

[16] R. M. D'Addio, R. S. Marinho, M. G. Manzato, Combining different metadata views for better recommendation accuracy, Information Systems 83 (2019) 1–12.

[17] matthewearl, faceswap, [Online; accessed 23. Jul. 2020] (Jul 2020). URL https://github.com/matthewearl/faceswap

[18] MarekKowalski, FaceSwap, [Online; accessed 23. Jul. 2020] (Jul 2020). URL https://github.com/MarekKowalski/FaceSwap

[19] L. Tran, X. Liu, Nonlinear 3d face morphable model, in: CVPR, Salt Lake City, UT, 2018.

[20] iperov, DeepFaceLab, [Online; accessed 1. Aug. 2020] (Aug 2020). URL https://github.com/iperov/DeepFaceLab

[21] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: WIFS, IEEE, 2018, pp. 1–7.

[22] N. Yu, L. S. Davis, M. Fritz, Attributing fake images to gans: Learning and analyzing gan fingerprints, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7556–7566.

[23] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[24] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, ICML (2019).

[25] N. Q. V. Hung, N. T. Tam, L. N. Tran, K. Aberer, An evaluation of aggregation techniques in crowdsourcing, in: International Conference on Web Information Systems Engineering, 2013, pp. 1–15.

[26] H. Yin, X. Zhou, B. Cui, H. Wang, K. Zheng, Q. V. H. Nguyen, Adapting to user interest drift for poi recommendation, IEEE Transactions on Knowledge and Data Engineering 28 (10) (2016) 2566–2581.

[27] H. Yin, Z. Hu, X. Zhou, H. Wang, K. Zheng, Q. V. H. Nguyen, S. Sadiq, Discovering interpretable geo-social communities for user behavior prediction, in: 2016 IEEE 32nd International Conference on Data Engineering (ICDE), 2016, pp. 942–953.

[28] N. Q. V. Hung, D. C. Thang, M. Weidlich, K. Aberer, Minimizing efforts in validating crowd answers, in: Proceedings of the 2015 ACM SIGMOD international conference on management of data, 2015, pp. 999–1014.

[29] N. Q. V. Hung, H. Jeung, K. Aberer, An evaluation of model-based approaches to sensor data compression, IEEE Transactions on Knowledge and Data Engineering 25 (11) (2012) 2434–2447.

[30] H. Chen, H. Yin, W. Wang, H. Wang, Q. V. H. Nguyen, X. Li, Pme: projected metric embedding on heterogeneous networks for link prediction, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1177–1186.

[31] H. Yin, H. Chen, X. Sun, H. Wang, Y. Wang, Q. V. H. Nguyen, Sptf: a scalable probabilistic tensor factorization model for semantic-aware behavior prediction, in: 2017 IEEE International Conference on Data Mining (ICDM), 2017, pp. 585–594.

[32] H. Yin, L. Zou, Q. V. H. Nguyen, Z. Huang, X. Zhou, Joint event-partner recommendation in event-based social networks, in: 2018 IEEE 34th International Conference on Data Engineering (ICDE), 2018, pp. 929–940.

[33] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[34] D. P. Kingma, M. Welling, Auto-encoding variational bayes, ICLR (2014).

[35] shaoanlu, faceswap-gan: A denoising autoencoder + adversarial losses and attention mechanisms for face swapping, in: https://github.com/shaoanlu/faceswap-GAN, 2019.

[36] W. Wu, Y. Zhang, C. Li, C. Qian, C. Change Loy, Reenactgan: Learning to reenact faces via boundary transfer, in: ECCV, 2018, pp. 603–619.

[37] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: CVPR, 2018, pp. 8789–8797.

[38] T. T. Nguyen, M. Weidlich, H. Yin, B. Zheng, Q. H. Nguyen, Q. V. H. Nguyen, Factcatch: Incremental pay-as-you-go fact checking with minimal user effort, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 2165–2168.

[39] T. Chen, H. Yin, H. Chen, R. Yan, Q. V. H. Nguyen, X. Li, Air: Attentional intention-aware recommender systems, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019, pp. 304–315.

[40] G. E. Hinton, A. Krizhevsky, S. D. Wang, Transforming auto-encoders, in: AAAI, Springer, 2011, pp. 44–51.

[41] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.

[42] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (10) (2016) 1499–1503.

[43] J. Boulanger, C. Gueudry, D. Münch, B. Cinquin, P. Paul-Gilloteaux, S. Bardin, C. Guérin, F. Senger, L. Blanchoin, J. Salamero, Fast high-resolution 3d total internal reflection fluorescence microscopy by incidence angle scanning and azimuthal averaging, Proceedings of the National Academy of Sciences 111 (48) (2014) 17164–17169.

[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: ICCV, 2018, pp. 4510–4520.

[45] deepfake_in_the_wild dataset.
URL `https://github.com/deepfakeinthewild/deepfake_in_the_wild`

[46] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.

[47] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, C. Ferrer, The deepfake detection challenge (dfdc) preview dataset (10 2019).

[48] P. Korshunov, S. Marcel, Deepfakes: a new threat to face recognition? assessment and detection, arXiv preprint arXiv:1812.08685 (2018).

[49] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe, Animating arbitrary objects via deep motion transfer, in: CVPR, 2019, pp. 2377–2386.