Pattern Recognition Letters 172 (2023) 121–127

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Hierarchical supervisions with two-stream network for Deepfake detection

Yufei Liang[a], Mengmeng Wang[a], Yining Jin[b], Shuwen Pan[c], Yong Liu[a,*]

[a] Laboratory of Advanced Perception on Robotics and Intelligent Learning, College of Control Science and Engineering, Zhejiang University, Hangzhou, China
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada
[c] School of Information and Electrical Engineering, Zhejiang University City College, Hangzhou, China

ARTICLE INFO

Article history:
Received 9 October 2021
Revised 29 March 2023
Accepted 28 May 2023
Available online 5 June 2023

Edited by: Prof. S. Sarkar

Keywords:
Deepfake detection
Frequency domain
Two stream
Coarse to fine

ABSTRACT

Recently, the quality of face generation and manipulation has reached impressive levels, making it difficult even for humans to distinguish real and fake faces. At the same time, methods to distinguish fake faces from reals came out, such as Deepfake detection. However, the task of Deepfake detection remains challenging, especially the low-quality fake images circulating on the Internet and the diversity of face generation methods. In this work, we propose a new Deepfake detection network that could effectively distinguish both high-quality and low-quality faces generated by various generation methods. First, we design a two-stream framework that incorporates a regular spatial stream and a frequency stream to handle the low-quality problem since we find that the frequency domain artifacts of low-quality images will be preserved. Second, we introduce hierarchical supervisions in a coarse-to-fine manner, which consists of a coarse binary classification branch to classify reals and fakes and a five-category classification branch to classify reals and four different types of fakes. Extensive experiments have proved the effectiveness of our framework on several widely used datasets.

1. Introduction

With the rapid development of deep learning, more and more face forgery algorithms [1–5] have been proposed. It is now possible to generate fake videos and images that are difficult for human eyes to distinguish. Unfortunately, sometimes technologies may be misused maliciously, damaging an individual's reputation and causing political threats. In order to reduce this risk, it is particularly important to develop an effective Deepfake detection algorithm.

Early research of Deepfake detection try to use handcrafted features or to modify the existing neural network structure to detect fake images [6–8]. However, with the significant advancement of face forgery technologies [9–11], these methods are no longer reliable for Deepfake detection. Since then, the research method has gradually shifted to introducing prior knowledge into the backbone network [12,13]. Though there are some detection methods that have achieved better detection accuracy on public datasets, there are still some problems. First, low-quality images are difficult to detect. Second, the diversity of forgery algorithms also causes problems.

At present, both human eyes and Deepfake detection methods could not distinguish the compressed low-quality fake images well. At the same time, most fake images propagated on the Internet are compressed images. In the spatial domain, since the artifacts of fake images have been compressed, no artifacts can be captured. However, we find the artifacts in the frequency domain are preserved. Some examples of low-quality fake images are shown in Fig. 1. As shown in the figures, the low-quality fake images are very similar to real images in the spatial domain. It is difficult for us to distinguish fake images like Fig. 1(a), but by visualizing the corresponding spectrum of the images (Fig. 1(b)), we can clearly see the artifacts in the spectrum of the fake images (Fig. 1(c)). Some studies [14–16] have shown that, fake images and real images have clear distinctions in the frequency domain. These methods only use information in the frequency domain of the image to detect fake images. However, the spatial domain still contains a lot of appearance representations about the image that should not be discarded. In this paper, we strive to use information both in the spatial domain and in the frequency domain to detect fake images. We are the first attempt to use image frequency domain information and spatial domain information in a two-stream network form for the Deepfake detection task.

* Corresponding author.
E-mail addresses: yufeiliang@zju.edu.cn (Y. Liang), mengmengwang@zju.edu.cn (M. Wang), jyining@ualberta.ca (Y. Jin), pansw@zucc.edu.cn (S. Pan), yongliu@iipc.zju.edu.cn (Y. Liu).

https://doi.org/10.1016/j.patrec.2023.05.029
0167-8655/

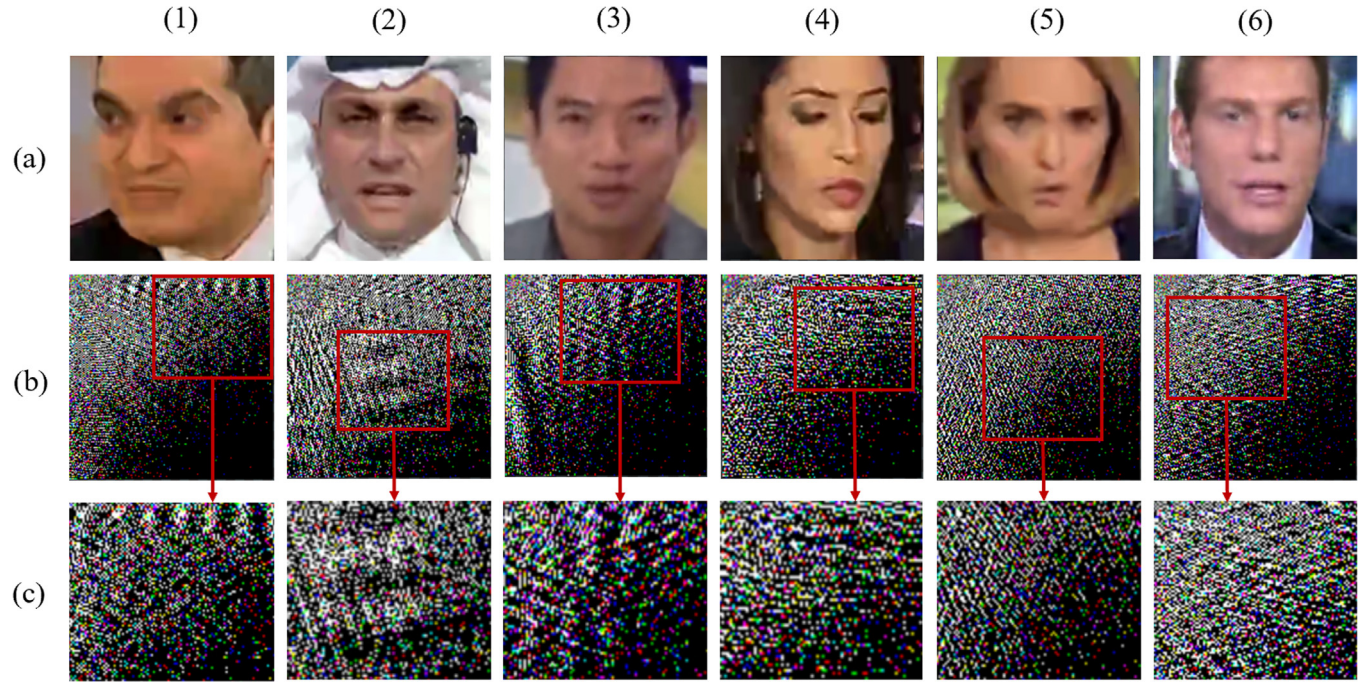

Fig. 1. Examples of real images and fake images. (a) is the images, (b) is the spectrum corresponding to the images, (c) is the artifacts in the spectrum, (1), (2) and (3) are fake images, (4), (5) and (6) are real images.

Moreover, most approaches turn the task of Deepfake detection into a binary classification problem. Considering that different GANs have different fingerprints, which means the feature distributions of fake faces generated by different operation methods are different. Therefore, it is unreasonable to simply classify all fake faces into one category. Since the constraints of multiple classifications are helpful to Deepfake detection tasks, in this paper, we use hierarchical classification in a coarse-to-fine way to implement the additional constraints of the multi-classification task after the binary classification task, which can further improve the performance.

In this paper, we propose a **H**ierarchical supervised **T**wo-stream **Net**work (HTNet) for Deepfake detection. Our main contributions can be summarized as follows:

- We introduce a two-stream network for Deepfake detection, which combines a spatial stream to encode semantic representations and a frequency stream to present frequency features.
- We use hierarchical supervisions in a coarse-to-fine manner to achieve the reintegration of fine-grained labels and ordinary binary labels.
- Extensive experiments demonstrate that our method achieves favorable performance.

2. Related work

Deepfake generation Forging faces is not just a modern problem. There is an image stitching technology[17] that can combine multiple images into a composite image. The new generation of AI-based image synthesis algorithms is based on the latest development of new deep learning models (especially Generative Adversarial Networks (GAN) [18]). Liu et al. [19] proposed an unsupervised image to image translation framework based on coupled GANs, which aims to learn the joint representation of images in different domains. This algorithm is the basis for the Deepfakes algorithm. Deepfakes[20] can replace the face of the target person with other faces in the video. Face2Face [3] is a real-time face reenactment system that use only an RGB camera. Instead of manipulating expressions only, the extended work [21] transfers the full 3D head position, rotation, expression, and eye blinking from a source actor to a portrait video of a target actor. The most advanced high-resolution (1024×1024) GAN models-PGGAN [22] and StyleGAN[9] can generate high-quality face images that can even fool humans.

Deepfake detection With the continuous development of Deepfake algorithms, the fake images generated by these algorithms are becoming more and more difficult to distinguish. Facial fakes pose a considerable threat to social security, so it is crucial to formulate effective countermeasures. Many proposals have been proposed. Traditional forgery can be detected by methods such as [23,24]. Zhou et al. [23] proposed a two-stream CNN for face tampering detection. Some early works used the biological characteristics in the face, such as. Li et al. [25] observed that Deepfake face lacks true blinking because the training images obtained from the Internet usually do not include photos of subjects with their eyes closed. The non-blinking phenomenon is detected by the CNN/RNN model, thereby exposing the Deepfake video. However, this detection can be circumvented by deliberately combining images with closed eyes during training. Yang et al.[6] used the inconsistency of head posture to detect fake videos. These methods achieved good detection results at that time, but with the continuous development of Deepfake algorithms, these algorithms are no longer reliable. Therefore, with the development of deep learning, the mainstream of research has gradually turned to introduce different information and prior knowledge into the backbone network to detect face forgery. Zhao et al. [26] insert the attention map into the backbone to help the network learn important features better. Qi et al. [13] used bioinformatics research and found that due to the blood circulating in the face, the skin color will periodically show small changes. Face X-ray [27] innovatively uses self-generated data to train the network to locate the hybrid boundary, which greatly improves the generalization ability. Two branch [28] uses a fixed filter bank to extract frequency information, limiting the ability to extract recognition features.

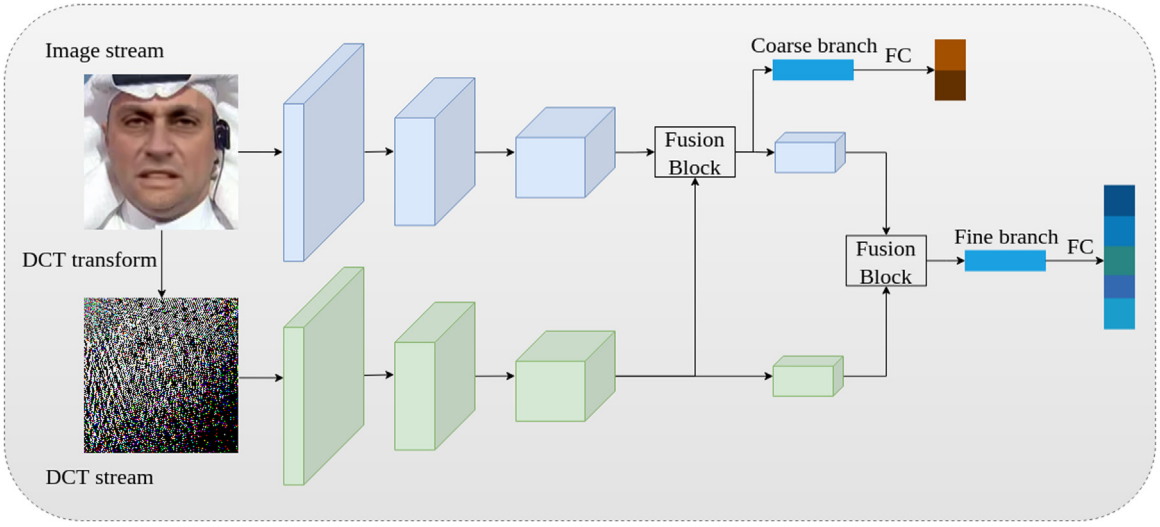

Fig. 2. Overview of the HTNet. Given an image, (1) use the DCT to obtain the image spectrogram, and send them to two networks respectively, where the frequency domain features and spatial domain features will be fused. (2) use the shallow output and deep output of the network to achieve hierarchical supervisions.

3. Method

We turn to introduce our HTNet. The architecture of the proposed HTNet is shown in Fig. 2. It consists of two streams, a spatial stream and a frequency stream. We take the spatial stream as the mainstream and integrate the frequency domain features into the spatial domain features via lateral connections in several layers. The spatial stream can focus on the spatial domain and semantics while the frequency stream can excavate imperceptible fake information due to the nature of frequency domain. By dealing raw input with different domains, our method allows the two streams to have their own expertise on Deepfake detection. Moreover, we get coarse classification results in the shallow layer of mainstream and get fine classification results in the deep layer to achieve hierarchical supervisions. The loss of coarse classification and fine classification is fused to get the final loss.

3.1. Two-stream framework

Spatial stream Spatial stream takes RGB images as input and encodes spatial features such as color features, texture features, etc. In Deepfake detection tasks, the spatial stream can capture forgery features such as facial position deviations, facial artifacts, and color artifacts. However, for more realistic forged images and compressed low-quality images, it is difficult to distinguish between forged images and real images only by information in the spatial domain.

As shown in Fig. 3(a), the feature maps of some forged faces extracted by the spatial stream are obviously different from real faces. However, some realistic forged faces are difficult to distinguish only by RGB image. For example, in Fig. 1(b), the fake face's feature map is similar to the real face's feature map. Thus, using only spatial domain information can not distinguish the fake image from the real image well.

Frequency stream Taking inspiration from traditional image forensics [29], we try to detect fake images with frequency domain information. Frank et al. [15] has shown that in the spatial domain, fake images look very similar to real images. However, in the frequency domain, multiple clearly visible artifacts can be easily found in frequency transformation of fake images. Therefore, in addition to the spatial domain information, the frequency domain information of the image can also assist Deepfake detection task. In order to utilize the frequency domain information in the Deepfake detection framework, we perform Discrete Cosine Transformation (DCT) on the image to obtain the corresponding spectrum of the image.

We use type ii 2D-DCT, which is also used for JPEG compression [29]. Thus, using DCT will be more compatible with the description of compression artifacts out of the forgery patterns.

Mathematically, 2-dimensional DCT on an input **I** with dimension N to the output **D** is defined as:

$$\mathcal{D} \quad = C^N \cdot I \cdot (C^N)^T \tag{1}$$

where C^N is the coefficient of the transform matrix defined by:

$$C^N_{jk} \quad = \sqrt{\frac{\alpha_j}{N}} cos(\frac{\pi\,(2k+1)j}{2N}) \tag{2}$$

given $\alpha_j = 1$ for $j = 0$, and $\alpha_j = 2$ for $j > 0$.

Our HTNet directly uses the spectrogram obtained by DCT as the network's input. Even though previous approache [30] has attempted to introduce frequency domain information in the face swap detection task, our HTNet is the first attempt to directly extract spectrogram features utilizing the network in the Deepfake detection task.

Lateral connections

We believe that information in the spatial domain and the frequency domain is different but complementary. Therefore, we propose a spatial-frequency two-stream network to fuse information in the two domains of the image. Our HTNet is the first attempt of using image spatial domain information and frequency domain information in a two-stream network form for the Deepfake detection task.

In our HTNet, we take the spatial stream as the mainstream and integrate the frequency domain feature map with the spatial domain feature map in the network. There are many feature fusion methods. Here we have tried three fusion methods, i.e., summation, concatenate and channel shuffle. We have implemented these three feature fusion methods in our HTNet. And Concat can get the best Deepfake detection accuracy after experiments.

3.2. Hierarchical supervisions

Different GANs have different fingerprints. It is unreasonable to simply classify all fake faces into one category. Therefore, beyond binary classification, we use the supervision of different forgery methods. Specifically, we realizes fine-grained label reintegration

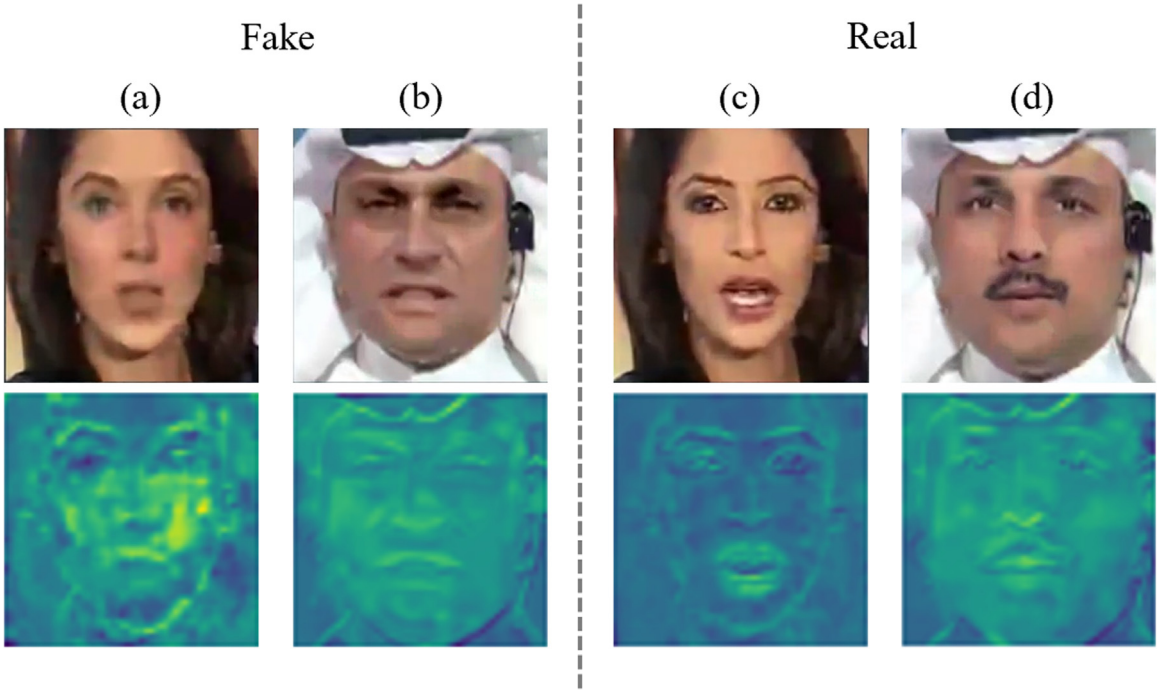

Fig. 3. Samples of faces and their feature maps. (a) and (b) are fake, (c) and (d) are real. The bottom is the feature maps.

first, then modifies the backbone network to have hierarchical outputs matched with new labels. Finally, we supervise multiple outputs respectively in a coarse-to-fine manner and fuse them in the overall loss function. The realization of hierarchical supervisions in the network is shown in Fig. 4.

3.2.1. Label reintegration

At first, we transform binary labels into multi-class labels according to fine-grained label information. One class for real images and multi-classes for fake images, which takes different forgery methods into consideration. Moreover, we can find that a hierarchical structure exists in the aforementioned multi-class labels when the whole dataset is divided into two classes, real and fake. Meanwhile, diverse fakes belong to the overall fake class. Inspired by Zhu and Bain [35], we reintegrate dataset label to its naturally hierarchical structure, where each sample has labels of both coarse level and fine level.

3.2.2. Hierarchical loss

Based on hierarchical labels, we modified the backbone network as shown in Fig. 4. CNNs extract image features hierarchically, which is one of their natural attributes. We combine it with hierarchical labels to further explore this property meanwhile improve the interpretability of features from various layers. Take HRNet [36] as an example. The original HRNet [36] includes four stages, where the unique output comes from the fourth stage and is supervised by binary cross-entropy loss after processed by the classification head. At this point, corresponding to hierarchical labels with two levels, we choose outputs from the third and the fourth stage to compute classification loss with labels. Respectively, for coarse classification, we employ output from the third stage and the first level label (real or fake) to compute binary cross-entropy loss. For fine classification, we employ output from the fourth stage and the second level label to compute multi-class cross-entropy loss. Both outputs are processed by classification head which is global average pooling to obtain feature vector from feature map. The following overall loss supervises the network training in a coarse-to-fine manner:

$$\begin{aligned}\mathcal{L}_C &= \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\alpha_h \mathrm{CE}(\mathbf{f}^h, y_i) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\alpha_h \log\left(\frac{e^{f^h_{y_i}}}{\sum_j e^{f^h_j}}\right)\end{aligned} \tag{3}$$

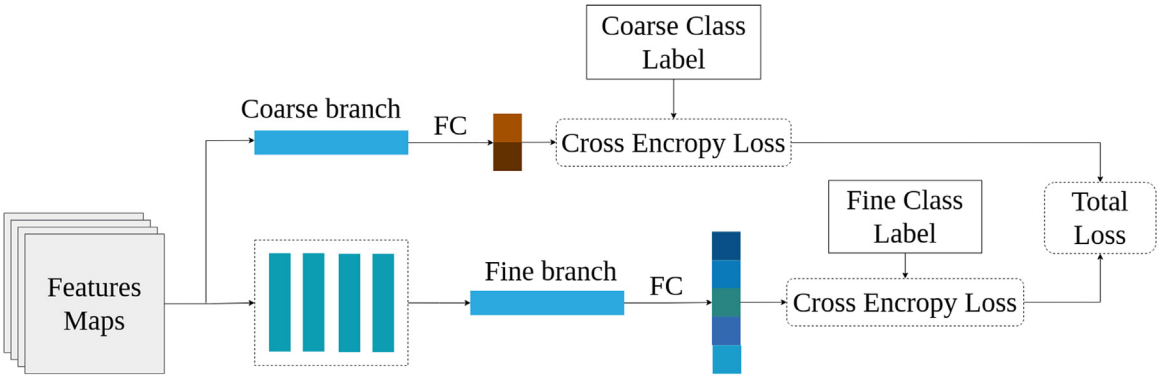

Fig. 4. Hierarchical supervisions diagram. The shallow feature maps pass through the coarse branch to get the coarse classification result, and the deep feature maps pass the fine branch to get the fine classification result.

Table 1
Acc and AUC on FaceForensics++ dataset .

Methods	Param.	GFLOPs	Acc (LQ)	AUC (LQ)	Acc (HQ)	AUC (HQ)	Acc (RAW)	AUC (RAW)
Steg.Features [29]	–	–	55.98%	–	70.97%	–	97.63%	–
LD-CNN [31]	–	–	58.69%	–	78.45%	–	98.57%	–
MesoNet [7]	0.3 M	–	70.47%	–	83.10%	–	95.23%	–
Face X-ray [27]	–	–	–	0.616	–	0.874	–	–
Xception [32]	20.8 M	4.6	82.71%	0.893	95.04%	0.963	98.77%	0.992
Xception-ELA [33]	20.8 M	4.7	73.69%	0.829	92.09%	0.948	97.13%	0.984
Xception-PAFilters [34]	20.8 M	4.9	83.24%	0.902	–	–	–	–
F^3-Net [30]	57.3 M	–	86.43%	0.914	96.63%	0.971	99.40%	0.996
Two Branch [28]	–	–	–	0.866	–	**0.987**	–	–
Multi-attention [26]	49.5 M	–	**86.95%**	0.873	96.37%	0.970	–	–
HTNet (Xception)	41.6 M	9.3	86.42%	0.921	96.89%	0.973	99.53%	0.994
HTNet (HRNet)	125.2 M	26.1	86.68%	**0.929**	**97.00%**	0.978	**99.60%**	**0.996**

The best scores are marked in bold and the second best scores are underlined.

where N is the total number of images, H is the total number of levels in the hierarchical classification, α_h is the hth level classification weight, f^h is the classification result of the hth lavel, y_i is the classification label of the hth lavel, and CE is the cross-entropy loss.

4. Experiment

4.1. Datasets

As in Li et al. [27], Masi et al. [28], we mostly use FaceForensics++ (FF++) [37] for our experiments due to its forgery diversity. It contains 1000 original videos and 4000 fake videos generated using DeepFakes [20], Face2Face [3], FaceSwap [38], and NeuralTextures [39]. Each video in FF++ has three different qualities: RAW, High Quality (HQ) and Low Quality (LQ), respectively. To evaluate the robustness of our HTNet, we also conduct experiments on the recent proposed large-scale face manipulated dataset, i.e., Celeb-DF [40] and Deepfake Detection Challenge (DFDC) [41] datasets.

4.2. Implementation details

Our framework is implemented by PyTorch. We conduct experiments on two backbones, Xception [32] and HRNet [36]. For Xception, we use Xception [32] pre-trained on ImageNet [42] as the backbone, during which the newly introduced layers and blocks are initialized randomly. For hierarchical supervisions, add coarse classification head after block7 for coarse classification, and add fine classification head after block12 for fine classification. For two-stream framework, spatial domain and frequency domain feature fusion is also performed on block7 and block12. We optimize the network through SGD. We set the benchmark learning rate to 0.05 and use the multistep learning rate scheduler to adjust the learning rate. Momentum is set to 0.9. Weight decay is set to 0.0001. For HRNet [36], add coarse classification head after stage3 for coarse classification, add fine classification head after stage4 for fine classification, perform feature fusion on stage3 and stage4, and the remaining settings are the same as Xception [32].

4.3. Comparing with previous methods

We compare our method with current state-of-the-art Deepfake detection methods. We evaluate the performance on FF++ and further evaluate the cross-dataset performance on Celeb-DF and DFDC.

4.3.1. Evalustion on FaceForensics++

The results are shown in Table 1. Our proposed HTNet performs better than them in all quality settings (LQ, HQ, and RAW) using Xception as the baseline. Especially for low quality (LQ) setting, the proposed HTNet achieves 86.42% in Acc and 0.921 in AUC with the baseline Xception. Compared to the second-best performing method (Xception-PAFilters [34]), the accuracy is improved by 3.18%, which is a significant improvement. These results show that our proposed method effectively solves the problem that low-quality fake images are difficult to distinguish. We conduct experiments on different backbones to verify that backbones do not restrict our proposed framework. When we use HRNet [36] as the backbone, the performance of our HTNet is further improved, reaching 86.68% in Acc and 0.929 in AUC on low-quality data.

4.3.2. Cross-dataset evaluation

Furthermore, we evaluate the cross-dataset performance of our HTNet, that is trained on FaceForensics++ but tested on Celeb-DF [40] and DFDC [41] datasets. The results are shown in Table 2. Our method shows better transferability than other existing methods.

4.4. Ablation study

The HTNet we proposed consists of two parts: hierarchical supervisions and spatial-frequency two-stream network. In order to evaluate the effectiveness of the proposed hierarchical supervisions and spatial-frequency two-stream of HTNet, we quantitatively evaluate HTNet and its variants in the following three ways: 1) baseline (Xception), 2) HTNet without spatial-frequency two-stream, 3) HTNet without hierarchical supervisions. We use Xception as the baseline for all ablation experiments and complete all experiments on low-quality data in Faceforensics++ dataset.

The quantitative results are shown in Table 3. By comparing Model 1 (baseline) and Model 2 (HTNet without spatial-frequency two-stream), the proposed hierarchical supervisions can improve the Acc and AUC scores of Deepfake detection tasks. Adding spatial-frequency two-stream based on Model 2, Acc and AUC scores are higher. These gradual improvements prove that the hierarchical supervisions and spatial-frequency two-stream in the proposed HTNet are indeed helpful for Deepfake detection tasks.

Table 2
Cross-dataset evaluation on Celeb-DF and DFDC by training on FaceForensics++.

Methods	Celeb DF	DFDC
MesoNet [7]	54.80	60.35
FWA [43]	56.90	67.30
Xception [32]	65.30	64.50
Multi-task [44]	54.30	–
Capsule [45]	57.50	53.30
Face X-ray [27]	66.40	65.50
Multi-attention [26]	67.44	69.87
HTNet (Xception)	67.83	71.98
HTNet (HRNet)	**69.82**	**73.38**

Table 3
Ablation study of the proposed HTNet on the low quality task(LQ). We compare HTNet and its variants by removing hierarchical supervisions and spatial-frequency two-stream network step by step .

hierarchical supervisions	Two-stream	Acc	AUC
		82.71%	0.893
√		85.659%	0.904
	√	85.95%	0.910
√	√	86.42%	0.921

Table 4
Results of different type of classifier .

Model	Acc	AUC
Binary classifier after block7	82.93%	0.897
Binary classifier after block12	82.71%	0.893
Multi-classifier after block12	83.59%	0.899
hierarchical supervisions	85.82%	0.915

Table 5
Results of different loss ratios of coarse classification and fine classification in hierarchical supervisions .

Loss Ratio	Acc	AUC
0.01	82.93%	0.897
0.1	85.18%	0.901
1	85.66%	0.914
5	85.82%	0.915
10	85.37%	0.904

4.4.1. Hierarchical supervisions

We separately prove the role of hierarchical supervisions in this subsection. There are two different one-hot labels from coarse to fine. In order to verify the effect of hierarchical classification, we conduct experiment. And the results are shown in Table 4. From the results, we can see that the results obtained by using hierarchical classification are better than the ones using binary classification and multi-classification. Experiments verify the effectiveness of hierarchical supervisions in Deepfake detection tasks.

Hierarchical classification involves coarse classification and fine classification loss ratio in the loss function. Here we choose different loss ratios for experiments. The results are shown in Table 5, and the accuracy versus loss ratio is shown in Fig. 5. From Table 5, the accuracy is the highest when the loss ratio is 5:1. The reason is that the coarse classification and the fine classification can be used in a balanced manner. The ratio of 5:1 can ensure that the two-classification task is the leading task while using the multi-classification supervision information.

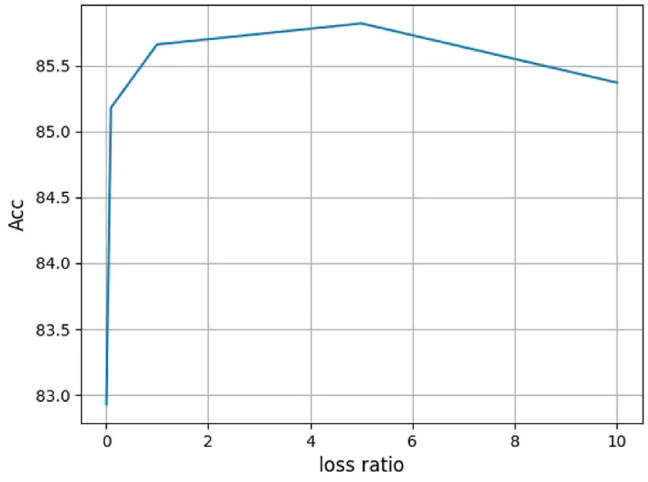

Fig. 5. The variation trend of accuracy with the loss ratio.

Table 6
Results of different fusion methods and baseline .

Fusion	Acc	AUC
image basline	82.71%	0.893
dct baseline	80.76%	0.862
Sum	85.82%	0.910
Concatenate	85.95%	0.914
Channel Shuffle	85.12%	0.897

4.4.2. Two-stream network

We explore the influence of frequency domain information and fusion modules of different structures on the performance of Deepfake detection tasks. In order to prove that the two-stream network that introduces frequency domain information (DCT) can improve the performance of the network in Deepfake detection tasks, we conduct experiments to compare the performance of 1) baseline (Xception), using only spatial domain information. 2) baseline(Xception), using only frequency domain information obtained by DCT. 3), 4), 5) two stream with three fusion methods. The results are shown in Table 6.

From the results, only using the frequency domain information can not get a good detection accuracy. The reason is that although the frequency domain contains additional information compared with the spatial domain, there is a lot of information in the spatial domain that the frequency domain is unable to express. In the HTNet, the spatial stream is the mainstream. The frequency domain information is integrated into the spatial information at specific layers, bringing auxiliary information to the monitoring model, and the detection accuracy is improved. From the results, we can see that simple concatenate and addition operations can achieve better performance. This fully shows the effectiveness of our two-stream network that integrates frequency domain information and spatial domain information.

5. Conclusion

In this article, we propose a new network for the Deepfake detection task—HTNet. HTNet is mainly composed of two-stream network and hierarchical supervisions. The two-stream network introduces frequency domain information into the model to assist in the Deepfake detection task. Compared with only using spatial domain or frequency domain information, the fusion of two domains feature map in the network can obtain good result. And the hierarchical classification innovatively transforms Deepfake detection task from a simple binary classification task to a coarse-to-fine multi-level classification task. It uses a more detailed classification task to assist the accuracy of the binary classification task. We also prove the effectiveness of our HTNet through extensive experiments.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Ackonwledgment

We thank all reviewers and the editor for excellent contributions. This work is supported by the National Key R&D Program of China (Grant No: 2021YFB2012301).

References

[1] I. Kemelmacher-Shlizerman, Transfiguring portraits, ACM Trans. Graph. (TOG) 35 (4) (2016) 1–8.
[2] M. R. Koujan, M. C. Doukas, A. Roussos, S. Zafeiriou, Head2head: video-based neural head synthesis, arXiv preprint arXiv:2005.10954(2020).
[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] Y. Nirkin, Y. Keller, T. Hassner, FSGAN: subject agnostic face swapping and reenactment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7184–7193.
[5] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, F. Moreno-Noguer, Ganimation: anatomically-aware facial animation from a single image, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 818–833.
[6] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8261–8265.
[7] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, MesoNet: a compact facial video forgery detection network, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2018, pp. 1–7.
[8] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, Protecting world leaders against deep fakes, in: CVPR Workshops, 2019, pp. 38–45.
[9] T. Karras, S. Laine, T. 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.
[10] C. Yang, S.-N. Lim, One-shot domain adaptation for face generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5921–5930.
[11] Y. Deng, J. Yang, D. Chen, F. Wen, X. Tong, Disentangled and controllable face image generation via 3D imitative-contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5154–5163.
[12] H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5781–5790.
[13] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, J. Zhao, Deeprhythm: exposing Deepfakes with attentional visual heartbeat rhythms, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4318–4327.
[14] N. Yu, L.S. Davis, M. Fritz, Attributing fake images to GANs: learning and analyzing GAN fingerprints, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7556–7566.
[15] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, Leveraging frequency analysis for deep fake image recognition, in: International Conference on Machine Learning, PMLR, 2020, pp. 3247–3258.
[16] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A.A. Efros, CNN-generated images are surprisingly easy to spot...for now, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8695–8704.
[17] J. Hays, A.A. Efros, Scene completion using millions of photographs, ACM Trans. Graph. (ToG) 26 (3) (2007) 4-es.
[18] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, arXiv preprint arXiv:1406.2661(2014).
[19] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, arXiv preprint arXiv:1703.00848(2017).
[20] Deepfakes, https://github.com/deepfakes/faceswap.
[21] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, C. Theobalt, Deep video portraits, ACM Trans. Graph. (TOG) 37 (4) (2018) 1–14.
[22] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196(2017).
[23] P. Zhou, X. Han, V.I. Morariu, L.S. Davis, Two-stream neural networks for tampered face detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2017, pp. 1831–1839.
[24] D. Cozzolino, L. Verdoliva, Noiseprint: a CNN-based camera model fingerprint, arXiv preprint arXiv:1808.08396(2018).
[25] Y. Li, M.-C. Chang, S. Lyu, In ictu oculi: exposing ai created fake videos by detecting eye blinking, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2018, pp. 1–7.
[26] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, Multi-attentional Deepfake detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2185–2194.
[27] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face X-ray for more general face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5001–5010.
[28] I. Masi, A. Killekar, R.M. Mascarenhas, S.P. Gurudatt, W. AbdAlmageed, Two-branch recurrent network for isolating Deepfakes in videos, in: European Conference on Computer Vision, Springer, 2020, pp. 667–684.
[29] J. Fridrich, Digital image forensics, IEEE Signal Process. Mag. 26 (2) (2009) 26–37.
[30] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: face forgery detection by mining frequency-aware clues, in: European Conference on Computer Vision, Springer, 2020, pp. 86–103.
[31] D. Cozzolino, G. Poggi, L. Verdoliva, Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection, in: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, 2017, pp. 159–164.
[32] 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.
[33] T.S. Gunawan, S.A.M. Hanafiah, M. Kartiwi, N. Ismail, N.F. Za'bah, A.N. Nordin, Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis, Indonesian J. Electr. Eng. Comput. Sci. 7 (1) (2017) 131–137.
[34] M. Chen, V. Sedighi, M. Boroumand, J. Fridrich, JPEG-phase-aware convolutional neural network for steganalysis of JPEG images, in: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia security, 2017, pp. 75–84.
[35] X. Zhu, M. Bain, B-CNN: branch convolutional neural network for hierarchical classification, arXiv preprint arXiv:1709.09890(2017).
[36] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, J. Wang, High-resolution representations for labeling pixels and regions, arXiv preprint arXiv:1904.04514(2019).
[37] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, FaceForensics++: learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.
[38] Faceswap, https://github.com/MarekKowalski/FaceSwap.
[39] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: image synthesis using neural textures, ACM Trans. Graph. (TOG) 38 (4) (2019) 1–12.
[40] 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.
[41] 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(2020).
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
[43] Y. Li, S. Lyu, Exposing Deepfake videos by detecting face warping artifacts, arXiv preprint arXiv:1811.00656(2018).
[44] H.H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos, arXiv preprint arXiv:1906.06876(2019a).
[45] H.H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: using capsule networks to detect forged images and videos, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2307–2311.