Manipulated Face Detector: Joint Spatial and Frequency Domain Attention Network

Zehao Chen and Hua Yang

Shanghai Jiao Tong University, China
{ChenZehao0915, hyang}@sjtu.edu.cn

Abstract. Face manipulation methods develop rapidly in recent years, which can generate high quality manipulated face images. However, detection methods perform not well on data produced by state-of-the-art manipulation methods, and they lack of generalization ability. In this paper, we propose a novel manipulated face detector, which is based on spatial and frequency domain combination and attention mechanism. Spatial domain features are extracted by facial semantic segmentation, and frequency domain features are extracted by Discrete Fourier Transform. We use features both in spatial domain and frequency domain as inputs in proposed model. And we add attention-based layers to backbone networks, in order to improve its generalization ability. We evaluate proposed model on several datasets and compare it with other state-of-the-art manipulated face detection methods. The results show our model performs best on both seen and unseen data.

Keywords: Manipulated Face Detection · Spatial and Frequency Domain · Attention Mechanism

1 Introduction

With the rapid development of digital image technology, computer vision and deep learning, face manipulation methods have made a great progress. The quality of manipulated face images is being improved amazingly. People feel more and more difficult to distinguish between real face images and manipulated ones, so do computers. Including gender, age, skin color and other appearance features, face images are the most discernible personal information, which can prove the identity of people. As the result, face recognition system are becoming popular from mobile screen lock to face scan payment, and face images play a more and more important role in the society. Therefor, the advance and popularity of manipulated face methods results in the wide spread of fake news [2], and the rising risk of privacy and identity safety.

Face manipulation methods can be divided into three categories by function: whole face forgery, face swap and facial features manipulation. Whole face forgery is using Generative Adversarial Networks(GANs) [10] to generate face images by

noise vector directly. PGGAN [17] and StyleGAN [18] can generate high quality face images with the resolution of 1024×1024. Face swap be separated into two varieties: identity swap and expression swap. Identity swap replace target person's entire face by source's, so the identity of target is changed to source's. Many popular applications can achieve this function, such as DeepFakes [9], ZAO [4] and so on. Expression swap just change target's expression by soucer's, but will not change target's identity. As for facial features manipulation, it change facial attributes on real face images, like hair color, hair style, gender, expression and others. StarGAN [5] can change facial features automaticallyy after setting parameters, and SC-FEGAN [15] can achieve this function through drawing masks by users.

Due to the progress of face manipulation methods, researchers pay more attention on manipulated face detection. Several detection methods [46,1,28,48,44] and [43,12,30,23] aim to one category of manipulated faces. Thus they perform poorly on the other categories. To solve this problem, Stehouwer *et al.*[38] put forward an attention-based Convolutional Nerual Network(CNN), not only can detect manipulated faces of all categories, but also can locate the manipulated area of the images. Moreover, methods proposed in [37,31,32,14] also have good performance on detecting manipulated faces of all categories. Because majority of manipulation methods are GAN-based, some detection methods [29,27,45] for GAN-generated images also make sense for detecting manipulated faces. However, methods mentioned above have common drawbacks, that are performing not so good on the datasets generated by state-of-the-art manipulation methods and lack of generalization ability.

In this paper, we propose a joint spatial and frequency domain attention network for all kinds of manipulated face detection. The overall structure of proposed model is shown as Fig.1. The spatial domain features are acquired by facial semantic segmentation, which divide a face image into five parts: background, face, eyes, nose and mouth. And we use Discrete Fourier Transform(DFT) to extract the frequency domain features. These six parts of a face image with the original image are used to inference whether the face image is real or fake by pre-trained VGG-19 [36], a well-known CNN, which is the backbone network of proposed method. On the input terminal of VGG-19, we add an attention-based layer proposed by us. The attention-based layer can produce a attention heatmap, which make backbone network pay more attention on those features that can distinguish real from manipulated images better, to improve the overall performance of the method. Our main contributions are as follows:

- We propose a novel manipulated face detector based on spatial and frequency domain combination, which shows great ability of generalization especially on those unseen data, comparing to the other state-of-the-art manipulated face detection methods.
- We put forward a attention-based layer for our method to focus more on informative features, which can effectively improve the performance of the baseline model.

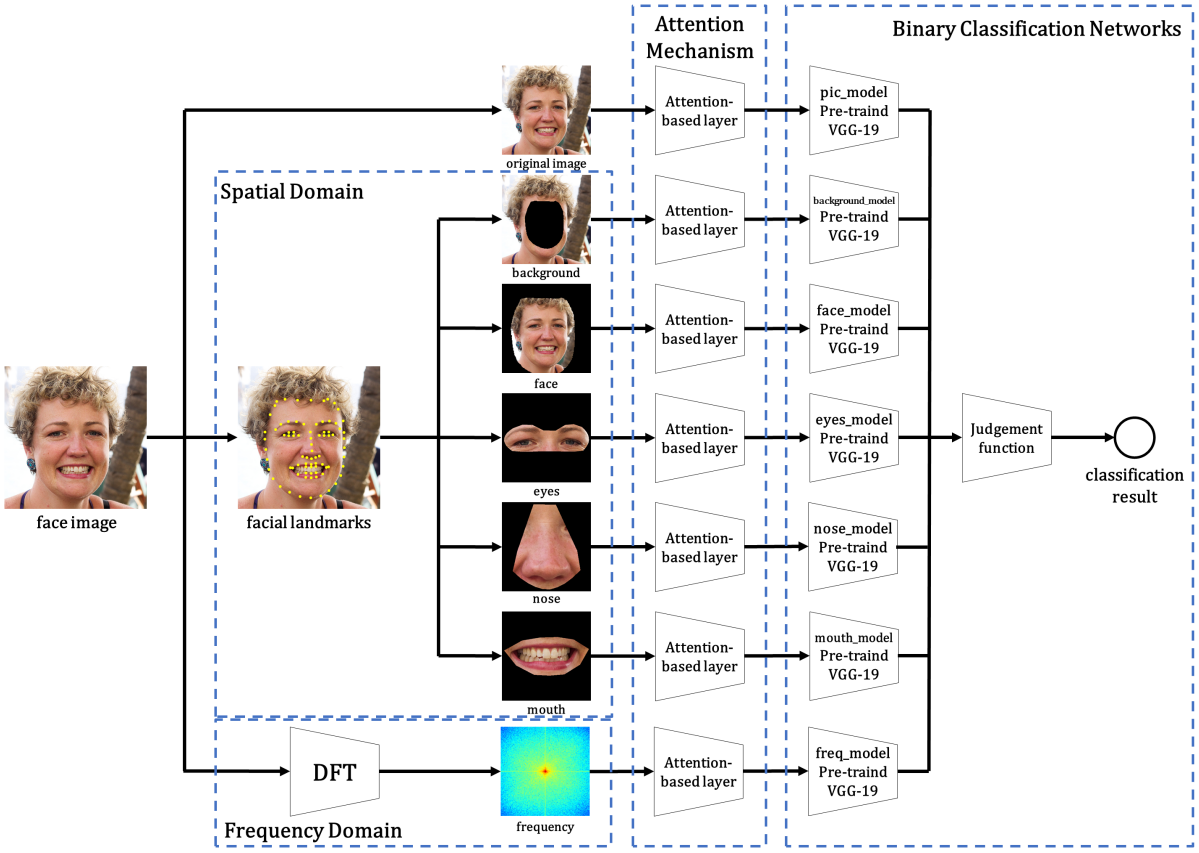

Fig. 1. The stucture of proposed model

2 Related Works

2.1 Face Manipulation Methods

Face manipulation methods have made a tremendous advance in recent years under the promotion of computer vision and deep learning. It's more and more difficult to tell difference between real and fake face images. At present, face manipulation methods can be divided into three categories by function: whole face forgery, face swap and facial features manipulation.

Whole face forgery This category of face manipulation methods is generating face images directly using GANs[10]. GAN is a classical framework for generating tasks, which mainly consists of two parts, a generator and a discriminator. Generator produce images by input noise vector, and the goal of discriminator is to distinguish the real images and the generated images. The training process

of a GAN is actually a dynamic game between the generator and the discriminator. They are optimized by each other, so finally the quality of the generated images are improved to a high level. Based on this framework, several modified GANs [34,3,11,26] with CNN-based generators and discriminators are proposed. However, these GANs can only generate low resolution face images with obvious evidence of manipulation.

Karras *et al.* present PGGAN[17], which significantly improves the quality of images generated by GANs. PGGAN uses the idea of starting from a low resolution and growing both the generator and discriminator progressively, and it can produce face images with the resolution of 1024×1024. Based on this, Karras *et al.* propose StyleGAN[18], which can control attributes (*e.g.* hair color and pose) of generated images on various level by controlling the latent code. In order to improve generated images quality and eliminate water droplet-like artifacts in generated images, the authors of StyleGAN modify the model greatly and propose StyleGAN2[19]. By using the technique of multi-scale gradients[16], MSGGAN produce higher quality face images.

Face Swap Face swap is to replace target person's face by source's, which has two varieties. Identity swap changes target's entire face except expression by source's, that can change the identity. And expression swap is just changing expression but not identity. There are kinds of face swap methods. Some of them are based on CNNs and GANs, the others are just using traditional ways of digital image technique.

DeepFakes[9] is the most popular and widely known identity swap algorithm. It is based on auto-encoders and CNNs. For each pair of target and source person, users need to train a specialized model with a large amount of face images. Thus it's a time-consuming method. FaceSwap[22] is a tradition-based method, which extracts face regions and transform them to realize face swap. It's more lightweight than DeepFakes. Based on Recursive Neural Network to reconstruct human faces, FSGAN[33] can swap target's face by source's on the pre-trained model. It does not need to train a one-on-one model for each pair. It remarkably improve the efficiency and conveniences of identity swap methods.

As for expression swap, Thies *et al.* [41] use RGB-D cameras tracking and reconstructing 3D model of two people's faces to realize facial reenactment. Based on this, the authors put forward a better expression swap algorithm, Face2Face [42], by combining 3D reconstruction and video re-render technique. Afterwards, the authors propose Neural Textures [40], which can use imperfect 3D content to produce high quality re-renderings. This work makes people look more natural after expression swap. A generative neural network with a novel space-time architecture, propose by Kim *et al.*[20], also can be used for expression swap.

In addition, several face swap datasets have been released. FaceForensics++[35] contains 1000 real videos and 4000 fake videos manipulated by four kinds of face swap algorithms. Celeb-DF[24] consists of 590 real videos and 5639 manipulated videos generated by modified DeepFakes algorithm[24]. And DFFC[7] contains over 5000 original and tampered DeepFakes videos in total.

Facial Features Manipulation Facial features manipulation is to change some attributes of real face images. FaceAPP[8] is a popular application for change facial features. Users can apply more than 28 modifications to real face images, such as changing age or adding smile.

Most of methods for facial features manipulation are GAN-based. They are originate from GANs used for style transfer. CycleGAN [47] is an excellent model for style transfer, which also can manipulate some basic facial features, like gender. However, it cannot deal with complicated features and needs to retrain a pair of generator and discriminator for each feature. StarGAN solve these problems by adding a mask vector. It can change several facial features among specific values through only one model. SC-FEGAN [15] can manipulate facial features by drawing masks, which really increase the flexibility of changing facial features, not limited to specific values any more.

Besides, StyleGAN [18] and StyleGAN2 [19] also can be used to manipulate facial features. If we use a decoder of the model to get the latent vector of a real face image, we can control facial features quantitatively by modifying the latent vector.

2.2 Manipulated Face Detection Methods

With the development of face manipulation methods, researchers gradually notice the necessity and importance of detecting manipulated faces, and put forward several detection methods. Methods proposed by early study are usually tested on low quality manipulated images. MesoNet [1] is a CNN-based model inspired by InceptionNet [39], which aims to face swap detection. [46,28] are also CNN-based models.

Afterwards, some more powerful methods are presented. Zhuang *et al.* [45] come up with a method based on two-step learning and triplet loss. It's trained by pairs of real and manipulated images to learn more difference between them. [43,44] both use Support Vector Machine(SVM), but the input is different. The input of [43] is the vector showing the number activated layers in the network, and the input of [44] is the vector of normalized facial landmarks' location. Nguyen *et al.* [30] propose a auto-encoder-based model to detect face swap images. And Face X-ray [23] applies noise analysis and error level analysis by self-supervised manner to detect face swap images.

However, methods mentioned above can only detect one category of manipulated face images, but cannot deal with other categories. Therefore, some methods that can detect all categories are put forward. Stehouwer *et al.*[38] put forward an attention-based CNN with a dateset called DFFD, which contains manipulated faces of all categories. Songsri-in *et al.* [37] propose a CNN-based model with combination of the original images and the location of facial landmarks as inputs. Capsule Network [31,32] uses dynamic routing algorithm to choose features extracted by several Capsule, and also performs well on several categories malipulated face images. FDFtNet [14], proposed by Jeon *et al.*, adds the self-attention-based architecture composed of attention modules and down-

samplers to pre-trained CNNs. And it will be fine tuned on various datasets to get better performance.

Besides, most of manipulation methods are GAN-based, we can learn from detection methods for GAN-generated images. Using co-occurrence matrices [29], frequency spectrum extracted by DFT[29] and colour saturation [27] to detect GAN-generated images are all effective. Inspired by frequency spectrum, Frank *et al.* use Discrete Cosine Transform(DCT) to get frequency spectrum. Using this, they detect GAN-generated face images by k-Nearest Neighbor(KNN) and CNN models.

Nevertheless, all of these manipulated face detection methods have common drawbacks, that are performing not so good on the datasets generated by state-of-the-art manipulation methods and lack of generalization ability. The varieties of manipulation methods is rapidly increasing and the quality of manipulated face images is continuously improving. So the performance on seen and unseen data both become challenges for manipulated face detection methods.

3 Proposed Model

We propose a manipulated face detection model based on joint spatial and frequency domain with attention mechanism. The structure of proposed model is shown as Fig.1. It mainly consists of four parts: spatial domain, frequency domain, attention mechanism and binary classification networks. Then, we start to introduce details of each part of proposed model.

3.1 Spatial Domain

We use facial semantic segmentation to acquire features of face images in spatial domain in our model. Firstly, we use the facial landmarks extractor from dlib [21] to get the location of 81 facial landmarks. Secondly, we adjust some landmarks' location for getting better segmentation results to make images of eyes, nose and mouth contain more region with effective information. Finally, we connect some of landmarks to get five parts of the a face image: background, face, eyes, nose and mouth, shown as Fig. When connecting landmarks, we do some cubic curve fitting to make the boundary of the five parts smoother. For background and face these two parts, we extend the area of the region to make sure they have enough effective information. Also, we remove the pixels without any information in the eyes, nose and mouth parts, and resize the images to the same size as the original images.

Explainability The essence of facial semantic segmentation used in our model to get features in spatial domain, is to cut a face image into five parts as individual inputs of binary classification networks. The way of segmentation, comparing to segmenting by block, is explainable semantically with specific meaning. And for each parts, they are the same regions of different faces. So it is easier for the corresponding binary classification network to learn more common difference between real and manipulated images from every part.

Re-use of Multi-scale Features For the five parts gotten from facial semantic segmentation, background and face are not resized. But for eyes, nose and mouth, the useless pixels in their images are removed. In order to make all input images same size, they're zoomed. So, the whole method get the multi-scale features of face images. Moreover, five parts of a face image are partly overlapped. Face part contains eyes, nose and mouth parts, and face part also overlaps background part. As a result, it realizes re-use of multi-scale spatial features for the whole model to make better performance.

High Efficiency The facial semantic segmentation used in our model, is based on facial landmarks. Connecting some of landmarks or fitting curve by them can cut face images into several parts. Comparing to using Fully Convolutional Network(FCN) [25] to segment face images directly, this way has a much lower algorithmic complexity. The time it consumes is much lower, so the efficiency of the whole model is increased. Besides, the way we use to segment face has higher flexibility. We can easily adjust the result of the segmentation by changing the location of landmarks. But if we use FCN, it must be more difficult, for we should retrain the model on re-labeled dataset.

Effectiveness Facial semantic segmentation make the model can learn multi-scale, more detailed and more decisive spatial features. Every part corresponds to its respective binary classification network, and produces a classification result. In the end, we take all classification results into account to produce the final classification result. The increase of classification results for each image, improve the error-tolerant rate of the model. For a real face image, even one of the part is inferred as manipulated, the final classification result won't be mistaken. And for a manipulated image, if two of classification results are '0', it will be considered as a manipulated image. Spatial features with this criterion improves the model's performance, especially its generalization ability.

3.2 Frequency Domain

For a 2D digital image with the width of M and the height of N, we can use Discrete Fourier Transform(DFT) to get its frequency spectrum, shown as Equation.1.

$$F(u,v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}. \tag{1}$$

The frequency spectrum of images show the intensity of the pixel value change in the image. The value of a pixel changing little from the adjacent pixel values corresponds to low frequency component. And if the value of a pixel changes a lot from the adjacent pixel values, it corresponds to high frequency component. Some details of images will be ignored in spatial domain, because we need

to resize images to a certain size. However, these details may be shown in frequency spectrum clearly, especially in high frequency component. It helps a lot for distinguishing real and manipulated face images.

For real face images are always photographed from real people and landscapes. So the contents of these images are usually natural, and we hardly can find violent change between adjacent pixels. As Fig.2 shows, frequency spectrum of real images are similar. The only have an orthogonal set of basic frequencies with narrow frequency band. There is few other frequency bands on both sides of basic frequency. And the values of high frequency are almost very low, with few high frequency noise.

But for manipulated face images, for generated by computers, have more or less traces of forgery in their spectrum. As Fig.2 shows, some frequency spectrum of manipulated images have much higher values of high frequency. Some have another basic frequency. And there are extra frequency bands on both sides of the basic frequency band in some manipulated spectrum. Although these manipulation characteristics in spectra are different, CNNs can learn the common difference between them and real spectra. That's the reason that frequency domain features works in proposed model.

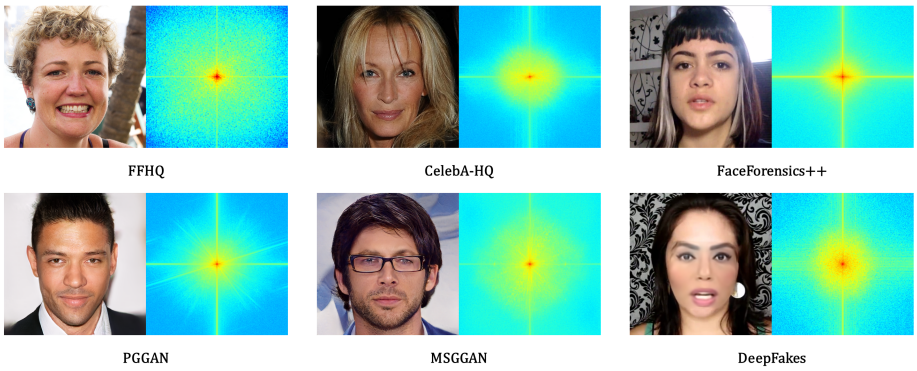

Fig. 2. Frequency spectrum of several kinds of face images. The first row are real images, and the second row are manipulated images.

Implementation Details We take the frequency spectrum got from DFT as frequency domain features in our proposed model. Generally, face images are often colorized, which have channels of R, G and B. So we need to apply DFT respectively to their each channel to get the frequency spectrum. After that, we apply the fftshift to shift the zero point of frequency to the middle of the spectrum. Moreover, we need to calculate modulus of every pixel in spectrum to change the spectrum from the complex number field to real number field. And

we also calculate logarithm of the spectrum. Finally, we normalize each spectrum to $[0, 1]$ to get the input of its corresponding binary classification network.

3.3 Attention Mechanism

The structure of our proposed attention-based layer is shown as Fig.3. The attention-based layer has two branches. For a face image with height of H, width of W and RGB three channels, $f \in \mathbb{R}^{H\times W\times 3}$, it goes through a convolutional layer to get the feature map 1, $F_1 \in \mathbb{R}^{H\times W\times 3}$. As for another branch, f goes through a convolutional layer block and be added by itself, to get feature map 2, $F_2 \in \mathbb{R}^{H\times W\times 3}$. After that, F_2 goes through anohter convolutional layer and be calculated by sigmoid function, to get the attentional heatmap $M \in \mathbb{R}^{H\times W\times 1}$. Finally, we multiply M and F_1, to acquire the attentional focused face image. It has the same size as original image. We add such attention-based layer to the input terminal of backbone network. And when training the model, they are optimized together with cross entropy loss function.

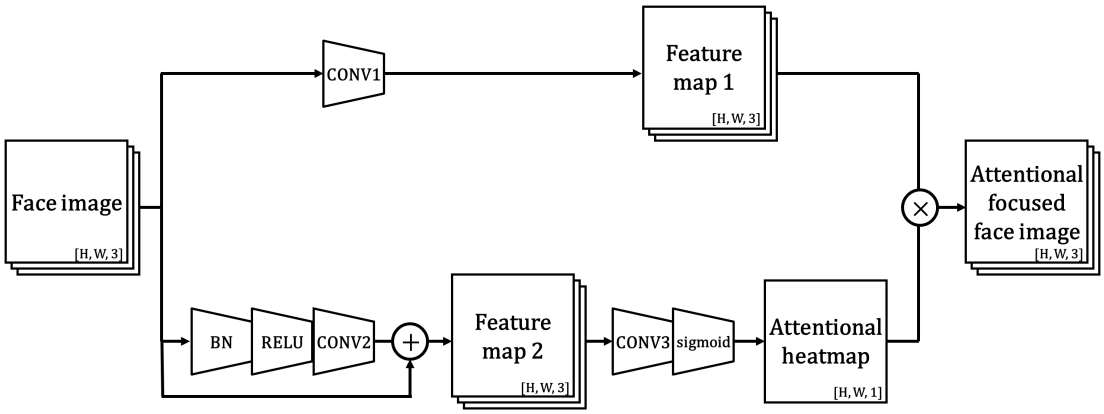

Fig. 3. The structure of proposed attention layer. CONV: convolutional layer, BN: batch normalization layer, RELU: Rectified Linear Unit.

Inspired by residual used in CNNs [13], F_2 is formed by letting the f go through a convolutional layer block and be added by f. Compared to f, F_2 has more effective features. Thus it can produce a better heatmap. The sigmoid function used in attention-based layer, is to normalize the heatmap from 0 to 1. If a pixel has higher attention, the corresponding value in heatmap is closer to 1. Attentional heatmaps of different inputs are shown in Fig.4.

As we know, CNNs can learn respective features of different kinds of images, which is the reason that CNNs classify them. However, distinguishing real and manipulated face images is not a easy task for CNN models. For real face images, the difference of shooting angle, light conditions and how much of the

image in focus reduce their feature similarity. And for manipulated images, the difference of manipulated area, image quality and manipulatin attributes also bring challenges to CNNs. Attention mechanism is used to slove sucn problem. It can make CNNs focus more on the common features, and reduce the interference of difference mentioned above. Thus, the proposed model can learn more common distinctions between real and manipulated images and strengthen its generalization ability.

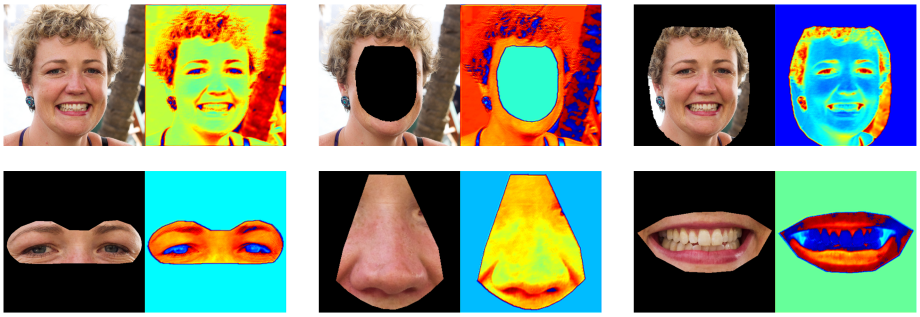

Fig. 4. Attentional heatmaps of different inputs.

3.4 Binary Classification Networks

Backbone Network For original image, background, face, eyes, nose, mouth and frequency parts, they have respective binary classification networks. We select VGG-19 [36] pre-trained on ImageNet [6] as backbone networks in proposed model. The proposed attention-based layers are added to input terminal of VGG-19. And the pre-trained VGG-19 is used for classify 1000 classes of images, the output length is 1000. We add a fully connected layer with 1000 of input length and 2 of output length to the output terminal of VGG-19. The output length of VGG-19 is changed to 2. Hence, we get the VGG-19 with attention-based layer as our binary classification network.

Classification Result Through respective VGG-19 with attention-based layer, seven classification results, $c_p, c_b, c_f, c_e, c_n, c_m, c_F$, are produced. These results have two values, '0' and '1'. '0' means manipulated and '1' means real. We need to draw the final conclusion, c, using these seven results. The judgement function is shown as follow:

$$c = J(c_p, c_b, c_f, c_e, c_n, c_m, c_F) = \begin{cases} 0, c_p + c_b + c_f + c_e + c_n + c_m + c_F < 6, \\ 1, c_p + c_b + c_f + c_e + c_n + c_m + c_F \geqslant 6. \end{cases} \quad (2)$$

We can know that if there are two or more results in seven results are 'manipulated', the final result is 'manipulated'. This judgement standard helps a lot

to improve detection accuracy on those unseen manipulated face images. Cause unseen manipulated face images may have similarity with seen data in some varieties of inputs. So it makes the generalization ability of the model stronger. As for real images, we also make sure the model's fault-tolerant ability, to reduce misjudgement as much as possible. And we can learn from the test result, that this criterion is reasonable and effective.

4 Experiments

To evaluate the ability of our proposed method, we test it on several datasets and compare the experiment results with other state-of-the-art manipulated face detection methods.

4.1 Datasets

We collect several manipulated faces datasets according to three categories: whole face forgery, face swap and facial features manipulation.

Whole Face Forgery Dataset This dataset contains two real face datasets, CelebA-HQ[17] and FFHQ[18]. As for manipulated faces, they are respectively generated by four state-of-the-art GANs: PGGAN[17], StyleGAN[18], StyleGAN2[19] and MSGGAN[16]. Training set has 10000 real face images from CelebA-HQ, 10000 from FFHQ, and 5000 manipulated face images geneated by PGGAN, 5000 by StyleGAN, 5000 by StyleGAN2, 5000 by MSGGAN, so does the test set. And the size of the validation set is a tenth of test set's.

FaceForensics++: This dataset consists of 1000 real videos grabbed from *YouTube* and 4000 manipulated videos. The authors used four kinds of face swap algorithms, which are FaceSwap(FS)[22], DeepFakes(DF)[9], Face2Face(F2F)[42] and NerualTexture(NT)[40], to respectively generate 1000 manipulated videos. We split it into a training, test and validation set, respectively consisting of 750, 225 and 25 real videos and corresponding manipulated videos produced by each face swap algorithm. After that, we extract some frames from videos, and use face detector in Dlib[21] to get the images of face region. We utilize 20000 real face images and 20000 manipulated images, 5000 for each face swap algorithm, to train our model. The test set has the same size as the training set, and the size of the validation set is a tenth of the test set's.

4.2 Implementation Details

There are seven VGG-19 binary classification models with attention-based layers in proposed method. For each binary classification model, it corresponds to a kind of input images. So we train these seven binary classification models respectively. The input size of all models is 224×224. We use SGD as the optimizer,

with initial learning rate of 10^{-3}. After every 5 epochs on the training set, the learning rate decays to a tenth of original. And each model is trained 15 epochs on the training set totally. We choose the one which has best performance on the validation set.

As for evaluating the method, we combine seven trained binary classification networks together as our model structure. We choose accuracy to evaluate the performance of methods.

4.3 Test on Seen Data

First, we verify overall performance of proposed method. We train and test the model on two datasets mentioned above. What's more, we compare the result to state-of-the-art methods and baseline model. All these models are trained and tested in the same way as our proposed model. The test result is shown as Tabel.1.

Table 1. The result of test on seen data

model \ dataset	Whole Face Forgery	FaceForensics++
VGG-19(baseline) [36]	99.48	99.69
Stehouwer *et al.*[38]	99.73	99.79
Capsule [32]	96.53	98.17
Proposed model	**99.94**	**99.93**

We can learn from the result that our proposed model performs best on all datasets, comparing to other state-of-the-art methods and baseline model. And the accuracy is really high, nearly to 100%. It means proposed model has great detection ability on those seen datasets. Because the final result of model is based on seven classification results, the probability of misjudgement is greatly reduced.

4.4 Test on Unseen Data

Then, we evaluate generalization ability of proposed model emphatically. There are four kinds of manipulation methods in whole face forgery dataseft. Each time we remove data generated by one of them in training set. And the test set is divided into two parts correspondingly. One part has data generated by same manipulation methods as training set, and the other part has data generated by the methods removed from training set. So the experiment has for groups, and each group contains one training set and two test set. For each group, we train proposed model and other contrastive methods on training set, and test them on two test sets respectively. The result is shown in Tabel.2.

Table 2. The result of test for seen generalization ability on whole face forgery dataset.

model										
	training set	PGGAN			✓		✓		✓	
		StyleGAN		✓			✓		✓	
		StyleGAN2		✓		✓			✓	
		MSGGAN		✓		✓		✓		
	test set	PGGAN		✓	✓		✓		✓	
		StyleGAN	✓			✓	✓		✓	
		StyleGAN2	✓		✓			✓	✓	
		MSGGAN	✓		✓		✓		✓	
VGG-19(baseline)[36]			99.12	80.68	99.32	98.04	99.67	60.94	99.37	89.72
Stehouwer *et al.*[38]			99.69	50.05	99.77	97.36	99.57	65.27	99.84	73.43
Proposed method			**99.69**	**93.80**	**99.89**	**99.85**	**99.91**	**94.47**	**99.91**	**98.44**

Also, we do the unseen data experiment on the FaceForensics++ dataset. The implementation details are same as unseen data experiment on whole face forgery dataset. The result in shown as Tabel.3.

Table 3. The result of test for seen generalization ability on FaceForensics++.

model										
	training set	FS			✓		✓		✓	
		DF		✓			✓		✓	
		F2F		✓		✓			✓	
		NT		✓		✓		✓		
	test set	FS		✓	✓		✓		✓	
		DF	✓			✓	✓		✓	
		F2F	✓		✓			✓	✓	
		NT	✓		✓		✓			✓
VGG-19(baseline)[36]			99.80	49.95	99.57	98.46	99.52	95.13	99.70	85.58
Stehouwer *et al.*[38]			99.70	53.79	99.68	97.95	99.66	97.93	99.58	88.68
Proposed method			**99.85**	**96.82**	**99.91**	**99.93**	**99.93**	**99.94**	**99.91**	**99.95**

The results of two datasets show that proposed model has best performance on all test indices. We focus on the analysis of generalization ability. It's clear to see accuracy of proposed model on unseen data is higher than other methods. All accuracy of proposed model on unseen data is more than 90%. And for some groups of datasets, like whole face forgery dataset without StyleGAN2 or FaceForensics++ without FaceSwap, proposed model performs much better than other methods on unseen data parts. Proposed model has seven varieties of inputs for one image. And for unseen manipulated images, seven inputs of them may not be all judged as 'manipulated'. But if only two or three varieties of inputs have similarity with seen manipulated data, the image will be judged as 'manipulated'. That's why proposed model has such strong generalization ability.

5 Ablation Study

In order to illustrate the effect of each parts of the proposed model, we do the ablation study. We use the whole face forgery dataset. The specific composition of the dataset is shown as Table.4. The test set 1 contains seen data to test overall ability. And the test set 2 consists of unseen manipulated face images, which is used to test generalization ability.

Table 4. The dataset for ablation study

	Real Face Images		Manipulated Face Images			
	CelebA-HQ	FFHQ	PGGAN	StyleGAN	StyleGAN2	MSGGAN
Training Set	7500	7500	5000	5000	0	5000
Test Set 1	7500	7500	5000	5000	0	5000
Test Set 2	2500	2500	0	0	5000	0
Validation Set	750	750	500	500	0	500

5.1 Spatial Domain

We remove the spatial domain features from the model to evaluate their function. We test the model with and without spatial domain features on the dataset used for ablation study. As for model without spatial domain features, we adjust judgement function, which is to produce the final classification result. The result is shown as Tabel.5.

Table 5. The result of ablation study on spatial domain.

	Test Set 1	Test Set 2
with	**99.91**	**94.47**
without	99.74	84.79

The result shows that model with spatial domain features performs better on both test set. It means using facial semantic segmentation to get five parts of face images and regarding them as input images, can improve model performance overall. Especially, generalization ability is ascended by spatial domain features.

5.2 Frequency Domain

To evaluated the function of frequency domain in proposed model, we remove it from the model. We evaluate the model with and without frequency domain features on the dataset used for ablation study. The result is shown as Tabel.6.

Table 6. The result of ablation study on frequency domain.

	Test Set 1	Test Set 2
with	99.91	**94.47**
without	99.91	91.52

The result shows that model with frequency domain features has same performance as the one without on test set 1, but has a better performance on test set 2. That's to say, adding frequency domain features can obviously improve generalization ability of proposed model, which can let model detect more unseen manipulated images.

5.3 Attention Mechanism

To evaluated the importance of attention mechanism in proposed model, we remove attention-based layers from the model. We train the model on the dataset used for ablation study, and contrast the result with the model with attention-based layers. To specify the effect of attention-based layers, we also show specific test accuracy of all inputs respectively in Tabel.7.

Table 7. The result of ablation study on attention mechanism.

input	Test Set1		Test Set2	
	with	without	with	without
original image	**99.76**	99.69	**77.77**	58.82
background	**99.70**	99.59	**83.35**	65.72
face	**99.64**	99.46	**74.11**	58.30
eyes	**99.74**	99.67	**75.72**	60.44
nose	99.88	**99.94**	**76.67**	76.02
mouth	99.78	**99.88**	**68.69**	64.97
frequency	**99.90**	99.84	**65.78**	58.18
proposed model	**99.91**	99.87	**94.47**	80.17

For single input, VGG-19 with attention-based layer performs better in 12 of 14 items, and the lower 2 items are just slightly inferior. It shows significant effect of attention mechanism, especially on the unseen data. Adding a attention-based layer to binary classification model can make it more focus on the attentioned area, which show more difference between real and manipulated images. And for the final result, proposed model without attention-based layer has similar performance on test set 1, but performs much worse than the one attention-based layer on test set 2. That is to say, the model with attention mechanism has stronger generalization ability. In summary, attention mechanism increase the performance of the proposed model overall.

6 Conclusion

We propose a manipulated face detector, which is based on joint spatial and frequency domain with attention mechanism. We evaluate proposed model on several categories manipulated face images datasets, and test both on seen and unseen data. The results show that proposed model achieve higher accuracy than other state-of-the-art detection methods. It proves strong overall manipulated face detection ability and generalization ability on unseen data of proposed model. Our ablation study illustrates function of each part in proposed model. The further work may include reducing the complexity of the model and lifting efficiency.

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–7. IEEE (2018)
2. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
4. Changsha Shenduronghe Network Technology Co., Ltd.: Zao, `https://apps.apple.com/cn/app/zao/id1465199127`
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
6. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (June 2009)
7. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854 (2019)
8. FaceApp Inc: Faceapp, `https://www.faceapp.com`
9. FaceSwapDevs: Deepfake, `https://github.com/deepfakes/faceswap`
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27, pp. 2672–2680 (2014)
11. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. pp. 5767–5777 (2017)
12. Gera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6 (2018)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
14. Jeon, H., Bang, Y., Woo, S.S.: Fdftnet: Facing off fake images using fake detection fine-tuning network. arXiv preprint arXiv:2001.01265 (2020)

15. Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
16. Karnewar, A., Wang, O., Iyengar, R.S.: MSG-GAN: multi-scale gradient GAN for stable image synthesis. arXiv preprint arXiv:1903.06048 (2019)
17. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
19. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. arXiv preprint arXiv:1912.04958 (2019)
20. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (TOG) **37**(4), 1–14 (2018)
21. King, D.E.: Dlib, `https://dlib.net/`
22. Kowalski, M.: Faceswap, `https://github.com/MarekKowalski/FaceSwap/`
23. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. arXiv preprint arXiv:1912.13458 (2019)
24. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A new dataset for deepfake forensics. arXiv preprint arXiv:1909.12962 (2019)
25. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
26. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802 (2017)
27. McCloskey, S., Albright, M.: Detecting gan-generated imagery using saturation cues. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 4584–4588. IEEE (2019)
28. Mo, H., Chen, B., Luo, W.: Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. pp. 43–47 (2018)
29. Nataraj, L., Mohammed, T.M., Manjunath, B., Chandrasekaran, S., Flenner, A., Bappy, J.H., Roy-Chowdhury, A.K.: Detecting gan generated fake images using co-occurrence matrices. Electronic Imaging **2019**(5), 532–1 (2019)
30. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876 (2019)
31. Nguyen, H.H., Yamagishi, J., Echizen, I.: 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). pp. 2307–2311. IEEE (2019)
32. Nguyen, H.H., Yamagishi, J., Echizen, I.: Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467 (2019)
33. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
34. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

35. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1–11 (2019)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Songsri-in, K., Zafeiriou, S.: Complement face forensic detection and localization with faciallandmarks. arXiv preprint arXiv:1910.05455 (2019)
38. Stehouwer, J., Dang, H., Liu, F., Liu, X., Jain, A.: On the detection of digital face manipulation. arXiv preprint arXiv:1910.01717 (2019)
39. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
40. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG) **38**(4), 1–12 (2019)
41. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. **34**(6), 183–1 (2015)
42. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2387–2395 (2016)
43. Wang, R., Ma, L., Juefei-Xu, F., Xie, X., Wang, J., Liu, Y.: Fakespotter: A simple baseline for spotting ai-synthesized fake faces. arXiv preprint arXiv:1909.06122 (2019)
44. Yang, X., Li, Y., Qi, H., Lyu, S.: Exposing gan-synthesized faces using landmark locations. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. p. 113118. Association for Computing Machinery (2019). https://doi.org/10.1145/3335203.3335724
45. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. ArXiv Preprint (2019)
46. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1831–1839. IEEE (2017)
47. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
48. Zhuang, Y., Hsu, C.: Detecting generated image based on a coupled network with two-step pairwise learning. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 3212–3216 (2019)