Multi-attentional Deepfake Detection

Hanqing Zhao[1] Wenbo Zhou[1,†] Dongdong Chen[2]
Tianyi Wei[1] Weiming Zhang[1,†] Nenghai Yu[1]
University of Science and Technology of China[1] Microsoft Cloud AI[2]
{zhq2015@mail, welbeckz@, bestwty@mail, zhangwm@, ynh@}.ustc.edu.cn
cddlyf@gmail.com

Abstract

Face forgery by deepfake is widely spread over the internet and has raised severe societal concerns. Recently, how to detect such forgery contents has become a hot research topic and many deepfake detection methods have been proposed. Most of them model deepfake detection as a vanilla binary classification problem, i.e, first use a backbone network to extract a global feature and then feed it into a binary classifier (real/fake). But since the difference between the real and fake images in this task is often subtle and local, we argue this vanilla solution is not optimal. In this paper, we instead formulate deepfake detection as a fine-grained classification problem and propose a new multi-attentional deepfake detection network. Specifically, it consists of three key components: 1) multiple spatial attention heads to make the network attend to different local parts; 2) textural feature enhancement block to zoom in the subtle artifacts in shallow features; 3) aggregate the low-level textural feature and high-level semantic features guided by the attention maps. Moreover, to address the learning difficulty of this network, we further introduce a new regional independence loss and an attention guided data augmentation strategy. Through extensive experiments on different datasets, we demonstrate the superiority of our method over the vanilla binary classifier counterparts, and achieve state-of-the-art performance. The models will be released recently at `https://github.com/yoctta/multiple-attention`.

1. Introduction

Benefiting from the great progress in generative models, deepfake techniques have achieved significant success recently and various face forgery methods [19, 41, 21, 31, 32, 44, 28, 38] have been proposed. As such techniques can generate high-quality fake videos that are even indistinguishable for human eyes, they can easily be abused by malicious users to cause severe societal problems or political threats. To mitigate such risks,many deepfake detection approaches [27, 34, 22, 33, 26, 45] have been proposed. Most of them model deepfake detection as a vanilla binary classification problem (real/fake). Basically, they often first use a backbone network to extract global features of the suspect image and then feed them into a binary classifier to discriminate the real and fake ones.

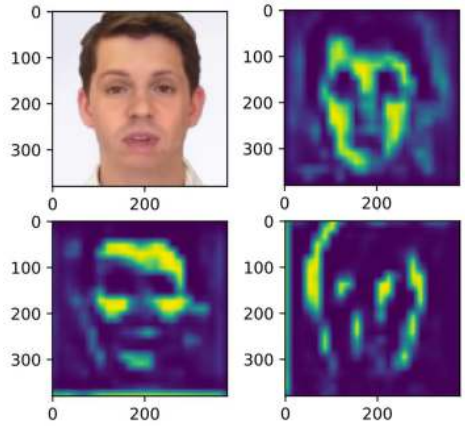

Figure 1: Example of the multiple attentional regions obtained by our method. The attention regions are separated and respond to different discriminative features.

However, as the counterfeits become more and more realistic, the differences between real and fake ones will become more subtle and local, thus making such global feature based vanilla solutions work not well. But actually, such subtle and local property shares a similar spirit as the fine-grained classification problem. For example, in the fine-grained bird classification task, some species look very similar and only differentiate from each other by some small and local differences, such as the shape and color of the beak. Based on this observation, we propose to model deepfake detection as a special fine-grained classification problem with two categories.

Inspired by the success of parts based model in the

† Corresponding Author.

fine-grained classification field, this paper presents a novel multi-attention network for deepfake detection. First, in order to make the network attend to different potential artifacts regions, we design multi-attention heads to predict multiple spatial attention maps by using the deep semantic features. Second, to prevent the subtle difference from disappearing in the deep layers, we enhance the textural feature obtained from shallow layers and then aggregate both low-level texture features and high-level semantic features as the representation for each local part. Finally, the feature representations of each local part will be independently pooled by a bilinear attention pooling layer and fused as the representation for the whole image. Figure 1 gives an example of the discriminative features obtained by our method.

However, training such a multi-attentional network is not a trivial problem. This is mainly because that, unlike single-attentional network [6] which can use the video-level labels as explicit guidance and be trained in a supervised way, the multi-attentional structure can only be trained in an unsupervised or weakly-supervised way. By using a common learning strategy, we find the multi-attention heads will degrade to a single-attention counterpart, i.e., only one attention region produces a strong response while all remaining attention regions are suppressed and can not capture useful information. To address this problem, we further propose a new attention guided data augmentation mechanism. In detail, during training, we will deliberately blur some high-response attention region (**soft attention dropping**) and force the network to learn from other attention regions. Simultaneously, we introduce a new regional independence loss to encourage different attention heads to attend to different local parts.

To demonstrate the effectiveness of our multi-attentional network, we conduct extensive experiments on different existing datasets, including FaceForensics++[34], Celeb-DF[25] and DFDC[9]. It shows that our method is superior to the vanilla binary classifier baselines and achieves state-of-the-art performance. In summary, the contributions of this paper are threefold as below:

- We reformulate the deepfake detection as a fine-grained classification task, which brings a novel perspective for this field.
- We propose a new multi-attentional network architecture to capture local discriminative features from multiple face attentive regions. To train this network, we also introduce a regional independence loss and design an attention guided data augmentation mechanism to assist the network training in an adversarial learning way.
- Extensive experiments demonstrate that our method outperforms the vanilla binary classification baselines and achieves state-of-the-art detection performance.

2. Related Works

Face forgery detection is a classical problem in computer vision and graphics. Recently, the rapid progress in deep generative models makes the face forgery technique "deep" and can generate realistic results, which presents a new problem of deepfake detection and brings significant challenges. Most deepfake detection methods solve the problem as a vanilla binary classification, however, the subtle and local modifications of forgeried faces make it more similar to fine-grained visual classification problem.

2.1. Deepfake Detection

Since the face forgery causes great threat to societal security, it is of paramount importance to develop effective countermeasures against it. Many works [46, 23, 4, 53, 34, 22, 33, 26, 45, 43] have been proposed. Early works [46, 23] detect the forgery through visual biological artifacts, e.g., unnatural eye blinking or inconsistent head pose.

As the learning based methods become mainstream, some works [53, 34] have proposed frameworks which extract features from spatial domain and have achieved excellent performances on specific datasets. Recently, more data domains have been considered by emerging methods. [45] detects tampered faces through Spatial, Steganalysis and Temporal features. It adds a stream of simplified Xception with a constrained convolution layer and an LSTM. [26] uses a two-branch representation extractor to combine information from the color domain and the frequency domain using a multi-scale Laplacian of Gaussian (LoG) operator. [33] uses frequency-aware decomposition and local frequency statistic to expose deepfake artifacts in frequency domain and achieves state-of-the-art performance.

Most existing methods treat the deepfake detection as a universal binary classification problem. They focus on how to construct sophisticated feature extractors and then a dichotomy to distinguish the real and fake faces. However, the photo-realistic counterfeits bring significant challenge to this binary classification framework. In this paper, we redefine the deepfake detection problem as a fine-grained classification problem according to their similarity.

2.2. Fine-grained Classification

Fine-grained classification [50, 49, 13, 37, 12, 52, 47, 17, 10] is a challenging research task in computer vision, which captures the local discriminative features to distinguish different fine-grained categories. Studies in this field mainly focus on locating the discriminative regions and learning a diverse collection of complementary parts in weakly-supervised manners. Previous works [50, 49] build part models to localize objects and treat the objects and semantic parts equally. Recently, several works [52, 47, 10] have been proposed under a multiple attentional framework, the

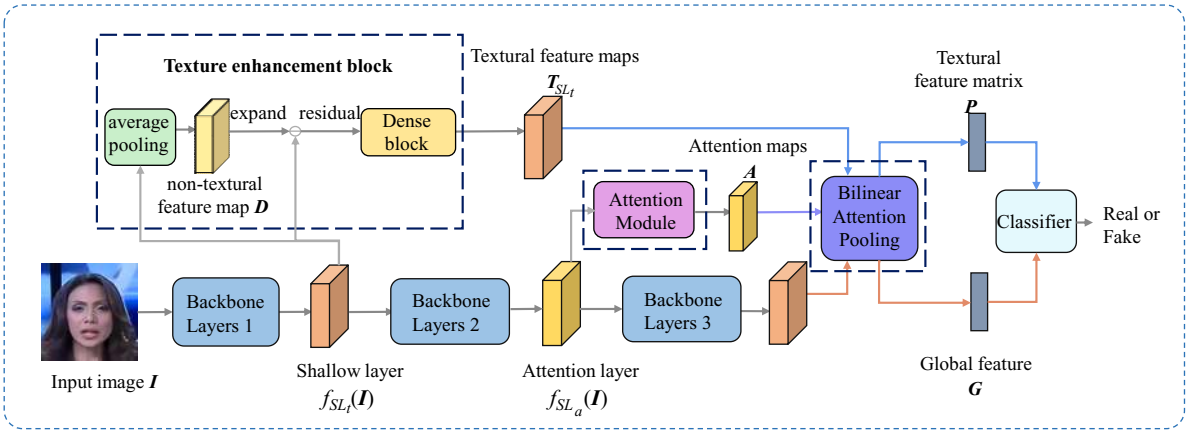

Figure 2: The framework of our method. Three components play an important role in our framework: an Attention Module for generating multiple attention maps, a texture enhancement block for extracting and enhancing the textural information and a bidirectionally used bilinear attention pooling for aggregating textural and semantic features.

core ideal of these method is that learning discriminative regions in multiple scales or image parts simultaneously and encouraging the fusion of these features from different regions. In addition, [17] designs attention cropping and attention dropping to obtain more balanced attention maps. In this paper, we model deepfake detection as a special fine-grained classification problem for the first time. It shares the same spirit in learning subtle and discriminative features, but only involves two categories, *i.e.*, real and fake.

3. Methods

3.1. Overview

In this section, we initially state the motivation of the designing and give a brief overview of our framework. As aforementioned, the discrepancy between real and fake faces is usually subtle and occurs in local regions, which is not easy to be captured by single-attentional structural networks. Thus we argue that decomposing the attention into multiple regions can be more efficient for collecting local feature for deepfake detection task. Meanwhile, the global average pooling which is commonly adopted by current deepfake detection approaches is replaced with local attention pooling in our framework. This is mainly because the textural patterns vary drastically among different regions, the extracted features from different regions may be averaged by the global pooling operation, resulting in a loss of distinguishability. On the other hand, we observe that the slight artifacts caused by forgery methods tend to be preserved in textural information of shallow features. Here, the textural information represents the high frequency component of the shallow features, just like the residual information of RGB images. Therefore, more shallow feature should be focused on and enhanced, which has not been considered by current state-of-the-art detection approaches.

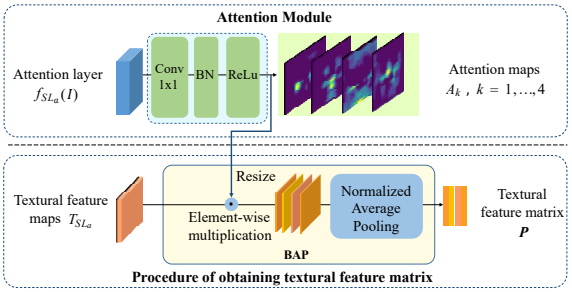

Figure 3: The structure of attention module and the procedure of obtaining textural feature matrix **P**. The proposed normalized average pooling is adopted instead of global average pooling.

Motivated by these observation, we propose a multi-attentional framework to solve the deepfake detection as a fine-grained classification problem. In our framework, three key components are integrated into the backbone network: 1) We employ an attention module to generate multiple attention maps. 2) We use densely connected convolutional layers [18] as a texture enhancement block, which can extract and enhance the textural information from shallow feature maps. 3) We replace the global average pooling with Bilinear Attention Pooling(**BAP**). And we use **BAP** to collect the textural feature matrix from the shallow layer and retain the semantic feature from the deep layer. The framework of our method is depicted in Figure 2.

Unlike single-attentional structure based network which can take the video-level labels as explicit guidance for training, the multi-attentional based network can only be trained in a unsupervised or weakly-supervised manner due to the lack of region-level labels. It could lead to a degradation of network that multiple attention maps focus on same region while ignoring other regions which may also provide discriminative information. To address the problem, we specifically design a Region Independence Loss, which

aims to ensure each attention map focusing on one specific region without overlapping and the focused region is consistent across different samples. Further, we employ the Attention Guided Data Augementation(**AGDA**) mechanism to decrease the salience of the most discriminative feature and force other attention maps to mine more useful information.

3.2. Multi-attentional Framework

Denote the input face image of network as I and the backbone network of our framework as f, the feature maps extracted from the intermediate stage of t-th layer is denoted as $f_t(I)$ with size of $C_t \times H_t \times W_t$. Here, C_t is the number of channels, H_t, W_t are the height and the width of feature maps, respectively.

Multiple Attention Maps Generation. As described above, given a real/fake face image I as input, our framework first uses an attention block to generate multiple attention maps for I. As shown in Figure 3, the attention block is a light weighted model which consists of a $1\times$ 1 convolutional layer, a batch normalization layer and non-linear activation layer ReLU. The feature map extracted from a specific layer SL_a will be fed into this attention block to obtain M attention maps A with size of $H_t \times W_t$, among which $A_k \in R^{\{H_t \times W_t\}}$ represents the k-th attention map and corresponds to one specific discriminative region, for example, eyes, mouth or even blending boundary defined in [22]. The determination of SL_a will be discussed in Section 4.

Textural Feature Enhancement. Most binary classification frameworks of deepfake detection do not pay attention to an important phenomenon, that is, the artifacts caused by forgery methods are usually salient in the textural information of shallow feature maps. The textural information here represents the high frequency component of the shallow features. Thus to preserve more textural information for capturing those artifacts, we design a textural feature enhancement block as shown in Figure 3. We first apply the local average pooling in patches to down-sample the feature maps from a specific layer SL_t and obtain the pooled feature map D. How to choose SL_t will be discussed in the following experiments part. Then similar to the texture representation of spatial image, we define the residual at the feature level to represent the texture information as below:

$$T_{SL_t} = f_{SL_t}(I) - D \tag{1}$$

Here T contains most textural information of $f_{SL_t}(I)$. We then use a densely connected convolution block with 3 layers to enhance T, the output is noted as $F \in R^{C_F \times H_s \times W_s}$, which is defined as "textual feature map".

Bilinear Attention Pooling. After getting the attention map A and textural feature map F, we use Bilinear Attention Pooling (BAP) to obtain feature maps. We bidirectionally use BAP for both shallow feature maps and deep feature maps. As shown in Figure 3, to extract shallow textural feature, we first use bilinear interpolation to resize the attention maps into the same scale with feature maps if they are not match. Then, we respectively element-wise multiply textural feature map F by each attention map A_k and obtain partial textural feature maps F_k.

To the end of this step, the partial textural feature maps F_k should be fed into classifier after global pooling. However, considering the differences among the different region range, if using the traditional global average pooling, the pooled feature vector will be influenced by the intensity of attention map, which violates the purpose of focusing on textural information. To address the problem, we design a normalized average pooling:

$$v_k = \frac{\sum_{m=0}^{H_s-1}\sum_{n=0}^{W_s-1} F_{k,m,n}}{||\sum_{m=0}^{H_s-1}\sum_{n=0}^{W_s-1} F_{k,m,n}||_2} \tag{2}$$

The normalized attention features $v_k \in R^{1\times N}$ are then stacked together to obtain the textural feature matrix $\mathbf{P} \in R^{M\times C_F}$, which will be fed into the classifier.

As to deep features, we first splice each attention map to get a single channel attention map A_{sum}. Then we use BAP for A_{sum} and the feature map from the last layer of network to get the global deep feature $\mathbf{G}$, which will also be fed into the classifier.

3.3. Regional Independence Loss for Attention Maps Regularization

As aforementioned, training a multiple attention network may easily fall into a network degraded case due to the lack of fine-grained level labels. In details, different attention maps tend to focus on the same region as shown in Figure 4 which is not conducive to the network to capture rich information for a given input. In addition, for different input images, we hope that the each attention map locates in fixed semantic region, for example, attention map A_1 focuses on eyes in different image, A_2 focuses on mouth. Therefore, the randomness of captured information by each attention map will be reduced.

To achieve these goals, we propose a Region Independence Loss which helps to reduce the overlap among attention maps and keep the consistency for different inputs. We apply BAP on the pooled feature map D obtained in Section 3.2 to get a "semantic feature vector" : $V \in R^{M\times N}$, and the Regional Independence Loss is defined as below by modifying the center loss in [15]:

$$\begin{aligned}\mathcal{L}_{RIL} = &\sum_{i=1}^{B}\sum_{j=1}^{M}\max(\left\|V_j^i - c_j^t\right\|_2 - m_{in}(y_i), 0) + \\ &\sum_{k,l\in(M,M),k\neq l}\max\left(m_{out} - \left\|c_k^t - c_l^t\right\|_2, 0\right)\end{aligned} \tag{3}$$

where B is the batch size, M is number of attentions, m_{in} represents the margin between feature and the corresponding feature center and is set as different values when y_i is 0 and 1. m_{out} is the margin between each feature center. $c \in R^{M \times N}$ are feature centers of V, it is defined as below and updated in each iteration:

$$c^t = c^{t-1} - \alpha \left(c^{t-1} - \frac{1}{B} \sum_{i=1}^{B} V^i \right) \tag{4}$$

Here α is the updating rate of feature centers, we decay α after each training epoch. The first part of L_{RIL} is an intra-class loss that pulls V close to feature center c, the second part is an inter-class loss that repels feature centers scattered. we optimize c by calculating the gradient for V in each batch. Considering that the patterns of texture in fake faces should be more diverse than real ones for fakes are generated by multiple methods, thus we restrict part features of fake faces in the neighborhood from the feature center of real ones but with larger margin. In this way, we give a larger margin in the intra-class for searching useful information in fake faces.

For the objective function of our framework, we combine this Regional Independence Loss with the traditional cross entropy loss:

$$\mathcal{L} = \lambda_1 * \mathcal{L}_{CE} + \lambda_2 * \mathcal{L}_{RIL} \tag{5}$$

$\mathcal{L}_{CE}$ is a cross entropy loss, λ_1 and λ_2 are the balancing weights for these two terms. By default, we set $\lambda_1 = \lambda_2 = 1$ in our experiments.

3.4. Attention Guided Data Augmentations

Under the restraining of Regional Independence Loss, we reduce the overlap of different attention regions. However, although different attention regions can be well separated, the attention maps may still respond to the same discriminative features. For example, in Figure 5, the attention regions are not overlapped but they all strongly respond to the landmarks of input faces. To force the different attention maps to focus on different information, we propose the Attention Guided Data Augmentation (**AGDA**) mechanism.

For each training sample, one of the attention maps A_k is randomly selected to guide the data augmentation process, and it is normalized as Augmentation Map $A_k^* \in R^{H \times W}$. Then we use Gaussian blur to generate a degraded image. Finally, we use A_k^* as the weight of original image and degraded image:

$$I' = I \times (1 - A_k^*) + I_d \times A_k^* \tag{6}$$

Attention guided data augmentation helps to train the models in two aspects. Firstly, it can add blurry to some regions which ensure the model to learn more robust features from other regions. Alternatively, AGDA can erase the most saliently discriminative region by chance, which forces different attention maps focusing their response on different targets. Moreover, the AGDA mechanism can prevent a single attention region from expanding too much and encourage the attention blocks to explore various attention region dividing forms.

Candidate of SL_t	Candidate of SL_a	ACC(%)
L2	L4	96.38
L2	L5	**97.26**
L3	L4	96.14
L3	L5	96.81

Table 1: Performance of our methods based on different combination of SL_t and SL_a.

4. Experiments

In this section, we first explore the optimal settings for our proposed multi-attentional framework and then present extensive experimental results to demonstrate the effectiveness of our method.

4.1. Implement Details

For all real/fake video frames, we use a state-of-the-art face extractor RetinaFace[8] to detect faces and save the aligned facial images as inputs with a size of 380×380. We set hyper-parameters $\alpha = 0.05$ in Equation 4, and decayed by 0.9 after each epoch. The inter-class margin m_{out} in Equation 3 is set to 0.2. The intra-class margin m_{in} are set as 0.05 and 0.1 respectively for real and fake images. We choose quantity of attention maps M, SL_a and SL_t by experiments. In AGDA we set the resize factor 0.3 and Gaussian blur $\sigma = 7$. Our models are trained with Adam optimizer[20] with learning rate 0.001 and weight decay 1e-6. We train our models on 4 RTX 2080Ti GPUs with batch size 48.

4.2. Determination of SL_a and SL_t

In this paper, we adopt EfficientNet-b4[39] as the backbone network of our multi-attentional framework. EfficientNet-b4 is able to achieve comparable performance to XceptionNet [3] with only half FLOPs. There are 7 main layers in total of EfficientNet, which are denoted from L1-L7, respectively.

As mentioned above, we observe that subtle artifacts tend to be preserved by textural features from shallow layers of the network, thus we choose L2 and L3 as the candidates of SL_t. Conversely, we want the attention maps to attend to different regions of the input, which needs the guidance of high-level semantic information to some extent. Therefore, we use deeper stage L4 and L5 as the candidates of SL_a.

Methods	LQ		HQ	
	ACC	AUC	ACC	AUC
Steg.Features[11]	55.98	-	70.97	
LD-CNN[5]	58.69	-	78.45	-
MesoNet[1]	70.47	-	83.10	-
Face X-ray[22]	-	61.60	-	87.40
Xception[3]	86.86	89.30	95.73	96.30
Xception-ELA[14]	79.63	82.90	93.86	94.80
Xception-PAFilters[2]	87.16	90.20	-	-
F^3-Net[33]	**90.43**	**93.30**	97.52	98.10
Two Branch[26]	-	86.59	-	98.70
EfficientNet-B4[39]	86.67	88.20	96.63	99.18
Ours(Xception)	86.95	87.26	96.37	98.97
Ours(Efficient-B4)	88.69	90.40	**97.60**	**99.29**

Table 2: Quantitative comparison on FaceForensics++ dataset with High-Quality and Low-Quality settings, respectively. The best performances are marked as bold.

By default setting $M = 1$, we train models with four combinations on FF++(HQ). From the results in Table 1, we find that the model reaches best performance when using L2 for SL_t and L5 for SL_a.

4.3. Comparison with Previous Methods

In this section, we compare our framework with current state-of-the-art deepfake detection methods. We evaluate the performance on FF++ [34] and DFDC [9], respectively. And we further evaluate the cross-dataset performance on Celeb-DF [25] in Section 4. We adopt ACC (accuracy) and AUC (area under Receiver Operating Characteristic Curve) as the evaluation metrics for extensive experiments.

4.3.1 Evaluation on FaceForensics++

FaceForensics++[34] is the most widely used dataset in many deepfake detection approaches, it contains 1000 original real videos from internet and each real video corresponds to 4 forgery ones, which are manipulated by Deepfakes, NeuralTextures[40], FaceSwap[48] and Face2Face[41], respectively. In the training process, we augment the original frames 4 times for real/fake label balance. We adopt EfficientNet-B4 as the backbone of our framework, and test the performances on HQ (c23) version and LQ (c40) version, respectively. Specially, when training our model on LQ, the parameters are initialized by those pretrained on HQ to accelerate the convergence. The comparison results are listed in Table 2.

The results in Table 2 demonstrate that our method achieves state-of-the-art performance on the HQ version of FF++. And the performances of different backbone verifies that our framework is not restricted by the backbone networks. However, the performance decreases 1.5% compared with F^3-Net [33] on the LQ version since F^3-Net is a specifically designed method for high-compressed deepfake videos detection. This is mainly because the videos in FF++(LQ) are highly compressed and cause a significant loss in textural information, which is a disaster to our texture enhancement designing. The results also reveal a limitation of our framework, that is, our framework is sensitive to high compression rate which blurs most of the useful information in spatial domain. We will make our framework more robust to compression in the future.

Method	Logloss
Selim Seferbekov[35]	0.1983
WM[51]	0.1787
NTechLab[7]	0.1703
Eighteen Years Old[36]	0.1882
The Medics[16]	0.2157
Ours	**0.1679**

Table 3: Comparison with DFDC winning teams' methods on the DFDC testing dataset. We participated in the competition as team WM.

4.3.2 Evaluating on DFDC Dataset

DeepFake Detection Challenge (DFDC) is the most recently released largest scale deepfake detection dataset, this dataset is public on the Deepfake Detection Challenge organized by Facebook in 2020. Currently, it is the most challenging dataset for deepfake detection task due to the excellent forgery quality of fake videos in this dataset. Seldom previous methods have been conducted on this dataset thus we train our model on the training set of this dataset and only compare the logloss score with the winning teams' methods of the DFDC contest. Here the provided logloss scores are calculated on the DFDC testing set(Ref. to Table 2 of [9]), which is one part of DFDC private set. Smaller logloss represent a better performance. The results in Table 3 demonstrate that our framework achieves state-of-the-art performance on DFDC dataset.

4.3.3 Cross-dataset Evaluation on Celeb-DF

In this part, we evaluate the transferability of our framework, that is trained on FF++(HQ) with multiple forgery methods but tested on Celeb-DF [25]. We sample 30 frames for each video to calculate the frame-level AUC scores. The results are shown in Table 4. Our method shows better transferability than most existing methods. Two-branch [26] achieves the state-of-the-art performance in transferability, however, its in-dataset AUC is far behind ours.

Method	FF++	Celeb-DF
Two-stream[53]	70.10	53.80
Meso4[1]	84.70	54.80
MesoInception4[1]	83.00	53.60
FWA[24]	80.10	56.90
Xception-raw[25]	99.70	48.20
Xception-c23[25]	99.70	65.30
Xception-c40[25]	95.50	65.50
Multi-task[29]	76.30	54.30
Capsule[30]	96.60	57.50
DSP-FWA[24]	93.00	64.60
Two Branch[26]	93.18	73.41
F^3-Net[33]	98.10	65.17
EfficientNet-B4[39]	99.70	64.29
Ours	99.80	67.44

Table 4: Cross-dataset evaluation on Celeb-DF (AUC(%)) by training on FF++. Results of some other methods are cited directly from [26]. Our method outperforms most deepfake detection approaches.

M	FF++(HQ)	Celeb-DF
1	97.26	67.30
2	97.51	65.74
3	97.35	66.86
4	**97.60**	**67.44**
5	97.39	66.82

Table 5: Ablation results on FF++(HQ) (Acc %) and Celeb-DF (AUC %) with different number of attention maps.

4.4. Ablation Study

4.4.1 Effectiveness of Multiple Attentions

To confirm the effectiveness of using multiple attentions, we evaluate how the quantity of attention maps affect the accuracy and transferability of our model. We train models in our framework with different attention quantities M on FF++(HQ), the other hyper-parameters are kept same as settings in Table 2. For the single attentional model, we do not use the regional independence loss and AGDA.

The Acc results on FF++(HQ) and AUC results on Celeb-DF are reported in table Table 5. In some cases, multi-attention based models perform better than the single attentional model, and we found that $M = 4$ provides the best performance.

4.4.2 Ablation Study on Regional Independence Loss and AGDA

As mentioned above, the regional independence loss and AGDA play an important role in regularized multiple attention maps training. In this part, we execute quantitative experiments and give some visualizations to demonstrate that

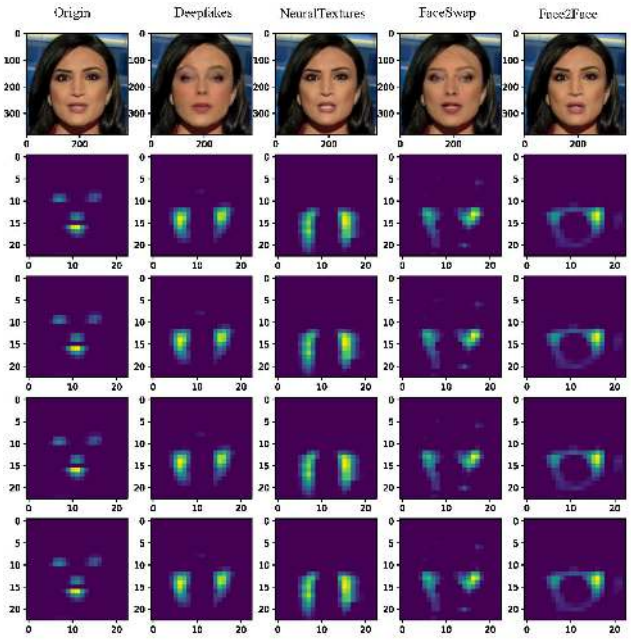

Figure 4: Attention maps trained without regional independence loss (RIL) and AGDA. Without RIL and AGDA, the network is easily degraded and the multiple attention maps locates the same regions of input.

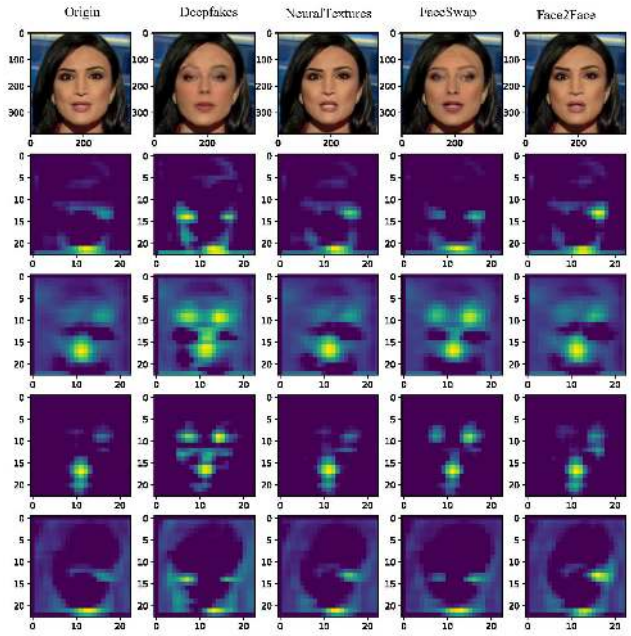

Figure 5: Attention maps trained without AGDA. Although the regional independence loss forces different attention maps to separate, they tend to respond to the same salient feature without the help of AGDA.

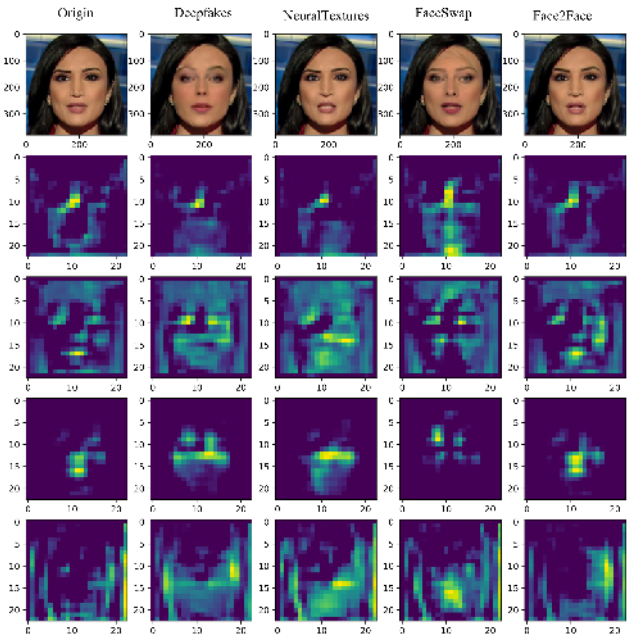

Figure 6: Attention maps trained with both regional independence loss and AGDA. The location and response of attention maps are correctly distributed.

these two components are necessary.

First, to demonstrate the effectiveness of our regional independence loss, we compare the performances of the models trained with different auxiliary losses. We keep all the settings identical with previous except for the loss function. With the same motivation in designing auxiliary loss, we substitute the regional independence loss with Additive Angular Margin softmax(AMS)[42] that can also force feature vectors close to their center.

Then we verify the effectiveness of our design for AGDA. As mentioned, we blur the original image to degrade the selected region of input. Thus the strategy of AGDA can be regarded as a "soft attention dropping". In this part, we alternatively adopt a "hard attention dropping", which directly erases pixels of selected region by binary attention mask BM:

$$BM_k(i,j) = \begin{cases} 0, & \text{if } A_k^*(i,j) > \theta_d \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

We set the attention dropping threshold $\theta_d = 0.5$ in this experiment. The comparison results of this ablation study are depicted in Table 6. The results verify that both regional independence loss (RIL) and attention guided data augmentations (soft attention dropping) have remarkable contribution to improve the performance of our framework.

To further help understanding of the function of regional independence loss and the AGDA strategy, we visualize the attention maps of models trained with/without these two components. Figure 4 illustrate the attention maps without RIL, it shows a clear trend that all attention maps are focused on same region. Figure 5 demonstrate that, although the attention regions are separated under the retraining of RIL, the different regions still exhibit similar response to the most salient features such as landmarks. This is not conductive for multiple attention maps to capture divergent information from different regions. While Figure 6 verifies that when both RIL and soft AGDA are adopted, the attention maps show response in discriminative regions with diverse semantic representations.

Loss type	AGDA type	FF++(HQ)	Celeb-DF
None	None	96.74	64.86
AMS	None	96.49	64.23
RIL	None	97.38	65.85
AMS	Hard	96.53	63.73
RIL	Hard	97.24	64.40
AMS	Soft	96.78	66.42
RIL	Soft	**97.60**	**67.44**

Table 6: Ablation results of different loss function and AGDA strategy. The model achieves best performance when using regional independence loss and soft AGDA mechanism. The metric on FF++(HQ) dataset is ACC, and on Celeb-DF is AUC.

5. Conclusion

In this paper, we research the deepfake detection from a novel perspective that is formulating the deepfake detection task as a fine-grained classification problem. We propose a multi-attentional deepfake detection framework. The proposed framework explores discriminative local regions by multiple attention maps, and enhances texture features from shallow layers to capture more subtle artifacts. Then the low-level textural feature and high-level semantic features are aggregated guided by the attention maps. A regional independence loss function and attention guided data augmentation mechanism are introduced to help train disentangled multiple attentions. Our method achieves good improvements in extensive metrics.

6. Acknowledgement

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, U1636201, 62002334, by the Anhui Science Foundation of China under Grant 2008085QF296, by the Exploration Fund Project of the University of Science and Technology of China under Grant YD3480002001 and the Fundamental Research Funds for the Central Universities under Grant WK2100000011.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security, WIFS 2018, Hong Kong, China, December 11-13, 2018*, pages 1–7. IEEE, 2018.

[2] M. Chen, V. Sedighi, M. Boroumand, and J. Fridrich. Jpeg-phase-aware convolutional neural network for steganalysis of jpeg images. *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, 2017.

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society, 2017.

[4] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[5] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In Matthew C. Stamm, Matthias Kirchner, and Sviatoslav Voloshynovskiy, editors, *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec 2017, Philadelphia, PA, USA, June 20-22, 2017*, pages 159–164. ACM, 2017.

[6] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020.

[7] Azat Davletshin. `https://github.com/NTech-Lab/deepfake-detection-challenge`.

[8] Jiankang Deng, J. Guo, Y. Zhou, Jinke Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *ArXiv*, abs/1905.00641, 2019.

[9] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[10] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Yi-Zhe Song, Zhanyu Ma, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, 2020.

[11] Jessica J. Fridrich and Jan Kodovský. Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.*, 7(3):868–882, 2012.

[12] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4476–4484. IEEE Computer Society, 2017.

[13] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society, 2014.

[14] T. S. Gunawan, Siti Amalina Mohammad Hanafiah, M. Kartiwi, Nanang Ismail, N. F. Za'bah, and A. N. Nordin. Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis. *Indonesian Journal of Electrical Engineering and Computer Science*, 7:131–137, 2017.

[15] Harald Hanselmann, Shen Yan, and Hermann Ney. Deep fisher faces. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017.

[16] James Howard and Ian Pan. `https://github.com/jphdotam/DFDC/`.

[17] Tao Hu and Honggang Qi. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *CoRR*, abs/1901.09891, 2019.

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[19] Ira Kemelmachershlizerman. Transfiguring portraits. *International Conference on Computer Graphics and Interactive Techniques*, 35(4):94, 2016.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[21] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. Head2head: Video-based neural head synthesis. *arXiv preprint arXiv:2005.10954*, 2020.

[22] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.

[23] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security, WIFS 2018, Hong Kong, China, December 11-13, 2018*, pages 1–7. IEEE, 2018.

[24] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *CVPR Workshops*, 2019.

[25] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[26] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. *arXiv preprint arXiv:2008.03412*, 2020.

[27] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[28] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. In *SIGGRAPH '18*, 2018.

[29] Huy H. Nguyen, Fuming Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, 2019.

[30] Huy H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019.

[31] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. pages 7184–7193, 2019.

[32] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Morenonoguer. Ganimation: Anatomically-aware facial animation from a single image. pages 835–851, 2018.

[33] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020.

[34] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019.

[35] Selim Seferbekov. `https://github.com/selimsef/dfdc_deepfake_challenge`.

[36] Jing Shao, Huafeng Shi, Zhenfei Yin, Zheng Fang, Guojun Yin, Siyu Chen, Ning Ning, and Yu Liu. `https://github.com/Siyu-C/RobustForensics`.

[37] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1143–1151. IEEE Computer Society, 2015.

[38] Ira Kemelmacher-Shlizerman Supasorn Suwajanakorn, Steven Seitz. Synthesizing obama: Learning lip sync from audio. *SIGGRAPH*, 36(4):95, 2017.

[39] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

[40] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: image synthesis using neural textures. *ACM Trans. Graph.*, 38(4):66:1–66:12, 2019.

[41] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niebner. Face2face: Real-time face capture and reenactment of rgb videos. pages 2387–2395, 2016.

[42] Feng Wang, Jian Cheng, Weiyang Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25:926–930, 2018.

[43] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *ICCV*, 2019.

[44] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. pages 622–638, 2018.

[45] Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2952–2956. IEEE, 2020.

[46] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 8261–8265. IEEE, 2019.

[47] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV*, 2018.

[48] Jiangning Zhang, Xianfang Zeng, Yusu Pan, Yong Liu, Yu Ding, and Changjie Fan. Faceswapnet: Landmark guided many-to-many face reenactment. *CoRR*, abs/1905.11805, 2019.

[49] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 834–849. Springer, 2014.

[50] Ning Zhang, Ryan Farrell, Forrest N. Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 729–736. IEEE Computer Society, 2013.

[51] Hanqing Zhao, Hao Cui, and Wenbo Zhou. `https://github.com/cuihaoleo/kaggle-dfdc`.

[52] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5219–5227. IEEE Computer Society, 2017.

[53] Peng Zhou, Xintong Han, V. Morariu, and L. Davis. Two-stream neural networks for tampered face detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839, 2017.