Exposing DeepFakes via Localizing the Manipulated Artifacts

Wenxin Li[1], Qi Wang[1], Run Wang[1,2], Lei Zhao[1,2(✉)], and Lina Wang[1,2]

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
leizhao@whu.edu.cn

[2] Key Laboratory of Aerospace Information Security and Trusted Computing,
Ministry of Education, Wuhan, China

Abstract. In recent years, *DeepFake* has become a public concern due to the abuse of advanced generative adversarial networks (GANs). Researchers have proposed various approaches to fight against DeepFakes by identifying whether an image or video is synthesized by GANs. Due to the imperfect design of GANs, the introduced artifacts serve as a promising clue for detection, which is captured by many proposed methods. However, these methods failed in presenting the artifacts in an interpretable manner. In this paper, we propose a novel approach by focusing on the artifact regions with dual attention (channel attention and spatial attention) to localize the observable and invisible artifacts for assisting *DeepFake* detection. Specifically, our proposed approach is agnostic to the specific backbone, which could be easily plugged into any DNN models to improve their performance. Experimental results show that our proposed dual attention could be deployed in any DNN based classifiers to improve their performance in detecting various DeepFakes. The detection accuracy on six current open-source *DeepFake* datasets is improved by 3.50%, 2.56%, 1.64%, 1.36%, and 0.89% in average on MesoNet, Meso-Inception, VGG-19, Xception, and EfficientNet, respectively. Besides, experimental results also show that our attention mechanism can serve as an asset for pixel-wise manipulation localization.

Keywords: DeepFake forensics · Localization · Dual attention

1 Introduction

Artificial Intelligence (AI) is developing rapidly at present. Among types of AI techniques, Generative Adversarial Networks (GANs) (e.g., ProGAN [12], StyleGAN [13] and FaceSwap-GAN [14]) have been widely used in multimedia synthesis and show great power. However, these advanced techniques can also be abused to generate so-called *DeepFake* videos or images. For example, multiple applications, such as *FaceApp* [2], *FaceSwap* [3], or *DeepFaceLab* [1] can generate faked videos where one's face can be swapped with someone else's. With these tools, users can generate *DeepFake* videos without any expert knowledge.

D. Gao et al. (Eds.): ICICS 2021, LNCS 12919, pp. 3–20, 2021.
https://doi.org/10.1007/978-3-030-88052-1_1

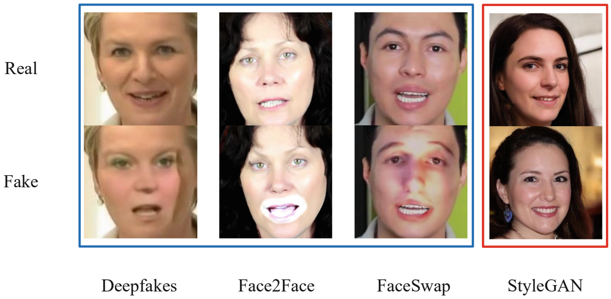


Fig. 1. Artifacts in DeepFake images.

The widespread of such *DeepFake* videos on the Internet seriously endangers the legitimate rights and interests of individuals and may result in a much worse impact if such videos are used for political and commercial purposes. Consequently, *DeepFake* brings a new yet significant threat to Cyber Security.

To fight against DeepFakes, previous studies have proposed multiple detection techniques. Based on the observation that synthesized images will inevitably introduce various artifacts due to the imperfect design of GANs, capturing artifacts is a promising approach for detection. Early techniques define features related to observable artifacts and then leverage these features for detection [5,14,27]. However, artifacts may not always be obvious because advanced GANs can often remove or reduce observable artifacts, which brings a challenge to early techniques. For example, we can observe from Fig. 1 that images synthesized with *Deepfakes* [20] have obvious blending boundaries, images created by *Face2Face* [20] show observable artifacts in the mouth region, and images created by *FaceSwap* [20] have unnatural light around the eyes and nose. By contrast, it is difficult to identify artifacts for images synthesized with *StyleGAN* [13].

To address the challenge caused by invisible artifacts, researchers proposed to design deep neural networks (DNNs) to learn the artifacts automatically [4,9,16]. A series of detection studies improve the generalization capability of DNNs using data augmentation [10] or detecting the blending boundary [24].

Despite the advances in *DeepFake* detection techniques, the interpretability of *DeepFake* has not been well investigated. For digital forensics, it is also critical to interpret how a detected DeepFake is manipulated by localizing the manipulated region. To achieve this purpose, a recent study [8] proposed an attention-mechanism-based technique for detecting and localizing artifacts in DeepFakes. This technique trains a neural network to focus on the ground truth mask regions and then leverage the attention map for localization. The main limitation of this technique is that it requires amounts of paired real/fake samples because it utilizes supervised learning that relies on ground truth manipulation masks for training. However, constructing the ground truth is challenging due to the scarce of paired samples. What's more, the resolution of the attention map is limited to a coarse-grained block level because it is designed to mask the high-dimensional features.

In this paper, we propose a novel approach for *DeepFake* detection and manipulated artifacts localization. The core of our approach is a dual-attention (channel attention and spatial attention) model to identify artifact regions. Compared with the previous technique [8] that leveraged supervised learning, our approach only relies on the labels and does not require the pair of synthesized images and the source images as the ground truth. To be more specific, we leverage neural networks to extract features and learn the connection between features and labels. Furthermore, dual attention enables the neural networks to adaptively refine critical features related to artifacts. Thus our attention is generated without the guidance of ground truth manipulation masks.

In detail, we leverage channel attention to figure out the channels related to the most critical features where artifacts may exist, and then leverage spatial attention to learn the distribution of features in the feature map. As channel attention and spatial attention are complementary, it is promising to combine such dual attention for better detection. Furthermore, the attention mechanism focuses on artifact regions. Thus, we can leverage the attention map to localize the manipulated pixels, and each pixel in the attention map corresponds to the pixel in the original image. In conclusion, our dual attention mechanism can focus on artifact regions that assist detection, as well as providing clues to locate the manipulated pixels with artifacts at a fine-grained pixel level.

To demonstrate the effectiveness of our approach, we construct experiments on five backbone models of different architectures and six datasets containing a diversity of *DeepFake* synthetic techniques. Experimental results show that our attention can be deployed to different models, improving the accuracy by 3.50% on MesoNet, 2.56% on Meso-Inception, 1.64% on VGG-19, 1.36% on Xception, and 0.89% on EfficientNet in average. Take MesoNet as an example, the model with dual attention reaches an accuracy of more than 92% on most datasets and an AUC of more than 0.92 on all datasets, outperforming the corresponding model without attention, with single attention, or with an extra convolution layer. Meanwhile, our approach performs fine-grained manipulation localization, and the PBCA of our attention maps reaches more than 0.7 while the IINC is less than 0.7 on all datasets, which outperforms the previous technique [8].

In summary, we make the following contributions:

- We propose to leverage the dual attention mechanism to detect artifacts that existed in the synthesized images. The attention mechanism focuses on channels and spatial distributions of critical features where artifacts may exist. And our approach can capture both observable and invisible artifacts.
- Beyond detection, our attention mechanism can further help to locate the manipulated pixels in an unsupervised manner. We learn the attention from the labels, instead of the ground truth manipulation masks, and leverage the attention map for localization.
- The dual attention mechanism proposed in our approach is scalable to convolutional neural networks of different architectures. Evaluation results show

that our attention mechanism improves the performance on different backbone models including MesoNet, Meso-Inception, VGG-19, Xception, and EfficientNet.

2 Related Work

2.1 DeepFake Creation

GANs have made great progress in image synthesis. Recently, a variety of GANs like ProGAN [12], StyleGAN [13], and FaceSwap-GAN [14] have been widely used for *DeepFake* generation. According to the tampered regions, there are two types of fake images synthesized with GANs: entire face synthesis and partial manipulation. As for the entire fake face synthesis, the whole facial image is synthesized with GANs. And partial manipulation includes expression swap, identity swap, and facial attributes editing. Specifically, expression swap transfers the facial expression from the source image to the target image, while identity swap replaces the whole target face with the source, and facial attributes editing modifies the facial attributes such as eyes or mouth. However, artifacts are inevitably introduced for both entire face synthesis and partial manipulation due to the imperfect design of GANs. Some artifacts are observable to humans while others are not, which could be captured by carefully designed models, and both of them provide critical clues for *DeepFake* detection.

2.2 DeepFake Detection

Existing *DeepFake* detection techniques can be roughly divided into two categories. One is detecting manually defined features with observable artifacts. Yang et al. [27] built the head pose vector using facial landmarks to capture the inconsistency between facial landmarks and head position, and trained an SVM classifier for detection. Agarwal et al. [5] designed an individual-specific monitor to capture the biological signals and trained an SVM classifier to detect them. But these techniques are rarely used at present, because existing artifacts may be removed with the development of GANs, and it cannot adapt to the changes. The other is based on DNNs by learning the artifacts automatically. Afchar et al. [4] built a convolutional neural network called MesoNet for *DeepFake* detection. They also added the Inception module and designed Meso-Inception to enable the model to extract additional features. Meanwhile, classical convolutional neural networks were also applied for *DeepFake* detection, such as VGGNet [21], InceptionNet [9], and Xception [20], etc. Among them, EfficientNet [22] has been proved to achieve the state-of-the-art performance in *DeepFake* detection. Besides, Wang et al. [25] improved the detection approaches by training a classifier based on ResNet and generalized it to unseen architectures using pre-/post-processing and data augmentation. Li et al. [15] trained their model to detect the blending boundary. Due to the strong self-learning ability of neural networks, deep learning is still the mainstream for *DeepFake* detection. However, there are still limitations because they provide no clues about the existence and location of artifacts in the fake images.

2.3 Attention Mechanism

Attention mechanism was widely adopted in NLP at first. Bahdanau et al. [6] applied the attention mechanism to solve the long-term dependence on contextual semantics problems in machine translation. Google's machine translation team also adopted the attention structure for the machine translation task [23]. Later, research about attention mechanism in the field of computer vision has been raised to simulate the way human views images. It can be divided into soft attention and hard attention. Soft attention is differentiable, which is trained by gradient descent and back propagation algorithm, while hard attention usually requires the prediction of the areas to be focused and applies reinforcement learning. On the whole, the soft attention mechanism has better performance in computer vision and is easier to be implemented in an end-to-end neural network. Woo et al. [26] proposed Convolutional Block Attention Module (CBAM), which is an effective soft attention module for feed-forward convolutional neural networks. Inspired by the excellent performance of the CBAM module, we apply the similar soft attention mechanism to help the neural networks capture observable and invisible artifacts in the synthesized images.

3 Our Approach

3.1 Insight

We observe that the images synthesized with GANs, including entire face synthesis and partial manipulation, will inevitably introduce various artifacts due to the imperfection of GANs. Some artifacts are observable while others are not, and both of them serve as a clue for *DeepFake* detection. However, a key challenge is that existing detection techniques cannot locate the pixels where artifacts exist even if they correctly detect fake images. In other words, they provide no evidence about the existence and location of artifacts, thus fail to present the artifacts in an interpretable manner.

To address this challenge, we propose a novel approach for *DeepFake* detection and manipulated artifacts localization by focusing on the artifact regions with dual attention. Specifically, we leverage channel and spatial attention to figure out the channels and spatial distributions of the most critical features related to the artifacts. Thus, our attention mechanism helps the neural networks automatically capture and focus on the artifacts of different GANs as well as locating the manipulated pixels. On the whole, our detection approach has three main advantages. The first is that the detection performance of different DeepFakes can be further improved since the attention mechanism helps neural networks concentrate on the specific artifacts to better cope with the changes and differences of GANs. The second is locating manipulated regions at pixel-level through unsupervised learning. The attention is obtained according to the labels instead of the ground truth manipulation masks so that we don't need to collect paired real/fake samples. Thirdly, our attention has a good generalization capability, which is agnostic to the specific backbone and can be easily plugged into any convolutional neural network to improve their performance significantly.

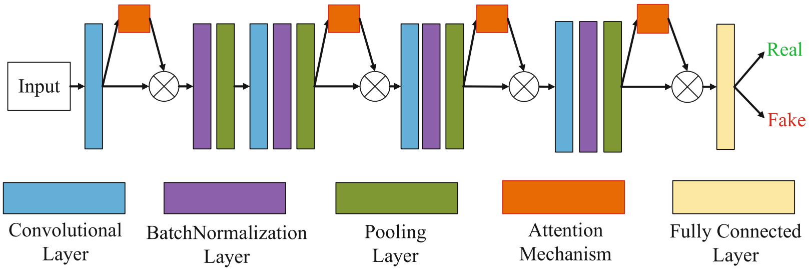


Fig. 2. The framework of our model. It consists of four basic blocks, each block contains a convolution layer, a batch normalization layer, and a pooling layer, and then connect to a fully connected layer. Attention is added to the extracted feature map of convolution layer for block one and pooling layers for the rest.

3.2 Framework

Here, to better illustrate the basic idea of our proposed approach, we choose MesoNet as the backbone model case and add our attention mechanism to the extracted feature maps. First, we focus on the artifacts of the *DeepFake* image rather than the whole image like prior studies. So we choose a shallow CNN that pays more attention to local features where artifacts exist. Additionally, we expect the network to pay more attention to the channels and regions concerning about the main artifacts in *DeepFake* images, so we add the attention mechanism to build our CNN model. Our attention mechanism can also be applied to arbitrary convolutional neural networks and more explanation refers to Sect. 4.

The framework is shown in Fig. 2. It consists of four basic blocks, each block includes a convolution layer, a batch normalization layer, a pooling layer, and further with a connection to a fully connected layer. For each basic block, the convolution layer extracts the latent semantic features of the image, and the feature maps are passed to the pooling layer for feature selection. The batch normalization layer is added in the middle to enhance the fitting ability of the model. For the first block, we calculate the attention and add it to the convolution layer, since the first convolution layer extracts the most detailed features that provide fine-grained localization. And for the rest blocks, attention is added to the pooling layers which have already filtered out excess information and we hope that our attention can mask the high-dimensional features to assist detection.

3.3 Dual Attention for Detection and Localization

The attention mechanism is used to simulate the way human views images. Humans usually only focus on important areas of images, and the attention mechanism enables important features in the feature maps to be further concerned and expressed, while other less important features are inhibited. We add two modules of attention mechanism: channel attention and spatial attention [26].

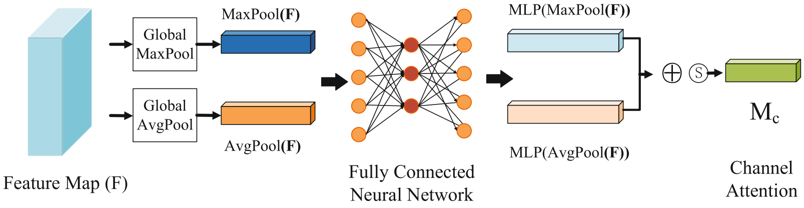


Fig. 3. The process of calculating channel attention.

Channel attention helps the model focus on important channels where artifacts may exist, and spatial attention focuses on the location distribution of artifacts in the feature map.

Channel Attention. For a convolutional neural network, different channels represent different features in the extracted feature map. Some channels represent important features for *DeepFake* detection such as eyes or mouth, while others are unnecessary like the background. Therefore, we add channel attention to help the neural network focus on channels concerned about the most important features where artifacts may exist. Suppose the output feature map of a convolution layer in CNN is $F \in R^{C*H*W}$, where C represents the channels of the feature map, H and W are height and width. Then we add channel attention M_c to the feature map. The new feature map with channel attention added can be represented as $F^{'} = M_c(F) \bigotimes F$, where $M_c \in R^{C*1*1}$, and $\bigotimes$ means element-wise multiplication.

The specific implementation process of calculating channel attention M_c is shown in Fig. 3. First, the input feature map is compressed using global max pooling and global average pooling. We use them at the same time, because average pooling has a global receptive field and captures overall characteristics, while max pooling focuses on unique features in the feature map. The combination of them helps to fully express channel attention. Then these vectors are fed into a fully connected neural network with shared dense layers. The network is used to explicitly model the correlation between channels, and the importance of each channel is expressed through nonlinear transformation. To reduce the parameters of the network, the hidden layer parameters are compressed, which also filters out some unimportant information. After obtaining these one-dimensional vectors representing the correlation and importance between channels, we calculate the sum of these two vectors and get the final channel attention M_c. This process can be expressed with Formula 1:

$$M_c(F) = \sigma(MLP(GAP(F)) + MLP(GMP(F))) \tag{1}$$

In formula (1), $M_c(F)$ is a one-dimensional matrix representing channel attention of the feature map. σ represents the activation function *Sigmoid*. *GAP(F)* means global average pooling and *GMP(F)* means global max pooling.

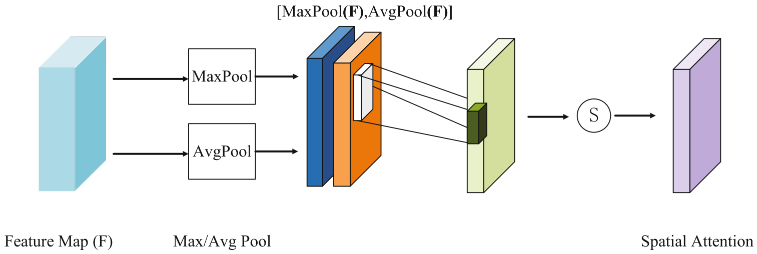


Fig. 4. The process of calculating spatial attention.

Spatial Attention. Different from channel attention, the spatial attention mechanism mainly focuses on the location information of the feature map, because the important features concentrate in a certain region and the location is uncertain. The idea is similar to the channel attention mechanism. A neural network can pay more attention to the areas with important features concerning artifacts by assigning different weights to different areas in the feature map. Suppose M_s represents spatial attention of the feature map, then we have $M_s(F) \in R^{H*W}$, where F is the input feature map, H and W represent height and width.

Figure 4 shows the specific implementation process of calculating M_s. First, max pooling and average pooling are used on the channel dimensions of the feature map. Similarly, max pooling focuses on the location of unique features, while average pooling focuses on the overall information of target features. And an effective feature descriptor of the feature map is formed by these two pooling operations. Then the convolution operation is performed on the feature descriptors to encode and model the locations that need to be emphasized or suppressed in the feature map. After that, we get the spatial attention $M_s \in R^{H*W}$. This process can be expressed with Formula 2:

$$M_s(F) = \sigma(f([F^S_{avg}, F^S_{max}])) \tag{2}$$

In formula (2), $M_s(F)$ is a two-dimensional matrix representing spatial attention of the feature map. σ represents the activation function *Sigmoid* and f represents convolution operation. F^S_{avg} is the descriptor of the average pooling on the feature map and F^S_{max} is the descriptor of the max pooling.

Dual Attention. The channel attention mechanism and spatial attention mechanism of the feature map are complementary. Channel attention focuses on information about the feature and its importance, while spatial attention focuses on the distribution of features in the feature map. Therefore, it is effective to combine them for better detection. It can not only help the neural network to pay more attention to key features, but also improve the ability to take control of the key feature distribution in space. Furthermore, the regions that our attention mechanism focuses on provide clues for localization.

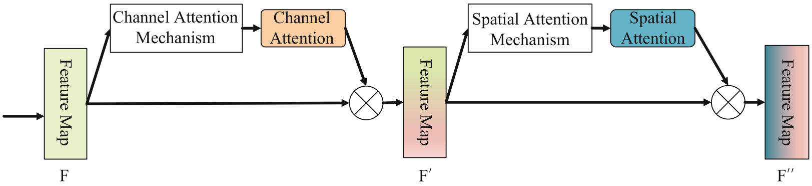


Fig. 5. The process of combining channel attention with spatial attention. For the input feature map, channel attention is calculated at first, because it will not interfere with the calculation and addition of spatial attention. Spatial attention is calculated later and added to the feature map that channel attention has been added to.

By combining channel attention with spatial attention and perform a linear transformation on the input feature maps, attention can be added to channel and space at the same time. The process of combining channel attention with spatial attention is shown in Fig. 5. For the input feature map, channel attention is calculated at first, because it will not interfere with the calculation and addition of spatial attention. After getting the feature map $F^{'}$ with channel attention added, spatial attention of $F^{'}$ is calculated. Then we add spatial attention to $F^{'}$ and get feature map $F^{''}$ with both attention added.

4 Experiment

4.1 Experimental Setup

Datasets. In this paper, we select six available open-source datasets for evaluation, namely Deepfake-Timit (DT) [14], UADFV [27], FaceForenscics (FF) [19], Celeb-DF (v2) [17], DeeperForensics (DeeperF) [11], and StyleGAN [13], including both images and videos. These datasets cover a wide range of GANs for *DeepFake* generation and their quality is relatively high. Among them, Deepfake-Timit uses FaceSwap-GAN for training, and the corresponding real videos come from VidTIMIT. UADFV is consist of both real and fake videos generated with *FakeApp*. FaceForenscics contains real YouTube videos and fake videos generated with *FaceSwap*. They are divided into three categories by the compression rate: Raw, HQ (High Quality), and LQ (Low Quality). Celeb-DF (V2) is also based on public YouTube videos, using an improved *DeepFake* synthesis technique, and has better visual quality. DeeperForensics is a video dataset generated by a newly proposed end-to-end face-swapping framework. These videos are first split into several frames, then the facial area of each frame is extracted to produce training and test images. We divide each dataset into disjoint training and test sets on a scale of approximately 9 to 1 except Celeb-DF (v2) according to its official documents. More details of the datasets are shown in Appendix A.

Implementation Details. We choose MesoNet [4], Meso-Inception [4], VGG-19 [21], Xception [7] and EfficientNet-B0 [22] as backbone to evaluate the

Table 1. Effectiveness performance of our approach on each test set.

Datasets	TPR ↑	TNR ↑	FPR ↓	FNR ↓	Accuracy ↑	AUC ↑	PBCA ↑	IINC ↓
FF-RAW	0.9839	1.0	0.0	0.0164	0.9918	0.9997	0.7168	0.6006
FF-HQ	0.9493	0.9518	0.0482	0.0507	0.9505	0.9925	0.7082	0.6404
FF-LQ	0.8492	0.7307	0.1693	0.1508	0.8401	0.9217	0.7113	0.6319
Celeb-DF (v2)	0.9469	0.8658	0.1811	0.0378	0.8975	0.9673	–	–
UADFV	0.9528	0.9960	0.0042	0.0451	0.9744	0.9792	0.8404	0.6869
DT-HQ	1.0	1.0	0.0	0.0	1.0	1.0	–	–
DT-LQ	1.0	1.0	0.0	0.0	1.0	1.0	–	–
DeeperF	0.9875	0.9815	0.0184	0.0126	0.9845	0.9982	0.7094	0.6211
StyleGAN	0.9060	0.9385	0.0636	0.0910	0.9223	0.9748	–	–

effectiveness of our added attention in improving the detection performance. Specifically, we add channel and spatial attention to every block in MesoNet as described in Sect. 3.2 (referred to as **our approach** in the following sections). And we insert our attention module after both block 1 and 4 in Meso-Inception, between block 1 and 2 of Xception, and at the last block for VGG-19 and EfficientNet. For VGG-19, we also add batch normalization layer after each block to avoid over-fitting. The performance when our attention is added to different places of the backbone models is analyzed in Sect. 4.3. We used Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a fixed learning rate of 0.001. In the module of channel attention mechanism, the compression ratio in the shared fully connected layer is set to 8. And in the module of spatial attention mechanism, the kernel size used for modeling spatial attention is set to 7 * 7.

4.2 Detection Effectiveness

In evaluation, we adopt six different metrics, namely TPR (recall, true positive rate), TNR (true negative rate), FPR (false positive rate), FNR (false negative rate), accuracy, and AUC (area under curve). The detection effectiveness of our approach on each test set is shown in Table 1. Its detection accuracy in all datasets except FaceForensics LQ and Celeb-DF (v2) reaches over 92%, and the average FPR and FNR of all datasets are 0.0539 and 0.0449. The AUC of our approach on all datasets can reach more than 0.92, indicating the good detection performance for a variety of GANs. On the whole, results show that our model performs well in various datasets and achieves good performance in these evaluation metrics. AUC and accuracy are relatively low in the FaceForensics LQ dataset possibly because the high compression rate causes serious image distortion and it is even hard for humans to distinguish, which also causes a high FNR.

Baselines. To evaluate the performance of our approach comprehensively, we manually reproduce the existing facial forgery detection techniques to train them on the same training sets as our model and test on the corresponding

Table 2. Comparison of detection accuracy with baseline techniques on each dataset. The improvement relative to the corresponding baseline has been highlighted.

Approach	FF-RAW	FF-HQ	FF-LQ	Celeb-DF (v2)	UADFV	DT-HQ	DT-LQ	DeeperF	StyleGAN
Head pose	0.5253	0.5301	0.4822	–	0.9444	0.7099	0.7326	0.4400	–
VA-MLP	0.7439	0.7105	0.6688	0.6908	0.7170	0.8205	0.8707	0.7384	0.5833
VA-logistic	0.7730	0.7398	0.6426	0.6661	0.6604	0.8405	0.9138	0.7783	0.6401
FWA	0.8357	0.7680	0.5544	0.5969	0.7215	0.7520	0.9986	0.4028	0.4923
MesoNet	0.9616	0.9017	0.8048	0.8470	0.9439	1.0	1.0	0.9750	0.8818
MesoNet+Att.	**0.9918**	**0.9505**	**0.8401**	**0.8975**	**0.9744**	1.0	1.0	**0.9845**	**0.9223**
Meso-Inception	0.9877	0.9170	0.8411	0.8616	0.8869	1.0	1.0	0.9892	0.9153
Meso-Inception+Att.	**0.9918**	**0.9316**	**0.8694**	**0.9019**	**0.9661**	1.0	1.0	**0.9908**	**0.9267**
VGG-19	0.9918	0.9371	0.8647	0.9284	0.8734	1.0	1.0	0.9882	0.9798
VGG-19+Att.	0.9918	**0.9566**	**0.8713**	**0.9477**	**0.9331**	1.0	1.0	**0.9932**	**0.9848**
Xception	0.9915	0.9624	0.8502	0.9611	0.9606	1.0	1.0	0.9972	0.9892
Xception+Att.	**0.9918**	**0.9668**	**0.9022**	**0.9717**	**0.9801**	1.0	1.0	**0.9988**	**0.9962**
EfficientNet	0.9918	0.9719	0.8706	0.9683	0.9812	1.0	1.0	0.9968	0.9965
EfficientNet+Att.	0.9918	**0.9859**	**0.9048**	**0.9756**	**0.9845**	1.0	1.0	**0.9982**	**0.9985**

test sets for comparison. Table 2 reports the detection accuracy. According to the results, detection approaches based on manually defined features, including Head Pose [27], VA-MLP [18], VA-Logistic [18], and FWA [16], are only effective on some limited datasets. As for other learning-based approaches, including MesoNet [4], MesoNet-Inception [4], VGG-19 [21], Xception [20], and EfficientNet [22], accuracy is relatively high on all datasets, even for the simplest network like MesoNet. However, our approach still has an improvement compared with the corresponding backbone models. Specifically, the accuracy is improved by 3.50%, 2.56%, 1.64%, 1.36%, and 0.89% in average on MesoNet, Meso-Inception, VGG-19, Xception, and EfficientNet, respectively. Accuracy reaches 100% on Deepfake-Timit, suggesting that it is simple for learning-based approaches to detect, thus is not taken into account here and in the following sections. Results show that our attention mechanism has the capability to improve the detection performance for different *DeepFake* generation techniques and backbone models, and is especially effective for shallow networks such as MesoNet and Meso-Inception.

Cross-dataset Evaluation. We also evaluate the generalization capability of our approach, which is trained on FaceForensics RAW dataset and tested on other datasets. As shown in Table 3, generally, the detection accuracy of our approach is slightly lower than the backbone model when they are tested on datasets other than FF-RAW. The possible reason is that our attention mechanism focuses mainly on the artifacts of specific method, thus the trained model is less effective in dealing with unknown GANs.

However, we argue that as long as our attention mechanism has seen the images synthesized with the new GANs, even in the case of multiple mixed generation methods, it can also pay attention to the respective artifacts of different methods. To confirm this, we mix training images in FF-RAW, Celeb-DF (v2),

Table 3. Detection accuracy in the cases of cross-dataset and mixed training set.

Approach	Train	Test				
		FF-RAW	Celeb-DF (v2)	UADFV	DeeperF	StyleGAN
MesoNet	FF-RAW	0.9616	**0.6155**	0.4990	**0.4980**	**0.5002**
MesoNet+Att.	FF-RAW	**0.9918**	0.5897	**0.5165**	0.4810	0.4980
MesoNet	mixed	0.8562	0.7482	0.9175	**0.9740**	0.8383
MesoNet+Att.	mixed	**0.9597**	**0.8109**	**0.9330**	0.9645	**0.8658**

Table 4. Comparison of accuracy when added with single attention, dual attention, and a normal convolution layer.

Approach	FF-RAW	FF-HQ	FF-LQ	Celeb-DF (v2)	UADFV	DeeperF	StyleGAN	Parameters
MesoNet	0.9616	0.9017	0.8048	0.8470	0.9439	0.9750	0.8818	28073
MesoNet+Att-c	0.9906	0.8816	0.8059	0.8205	0.9321	0.9812	0.9062	28287
MesoNet+Att-s	0.9912	0.9189	0.7909	0.8641	0.9589	0.9832	0.9190	28465
MesoNet+Att.	**0.9918**	**0.9505**	**0.8401**	**0.8975**	**0.9744**	0.9845	**0.9223**	28679
MesoNet+Conv.	0.9836	0.9101	0.7972	0.8892	0.9404	**0.9848**	0.9137	29681

UADFV, DeeperForensics, and StyleGAN to create a mixed training set. Then we train MesoNet as well as our approach on this dataset, and they are tested on the original test sets. Results in Table 3 show that detection accuracy of our approach substantially outperforms the backbone model, especially in FF-RAW and Celeb-DF (v2), indicating the capability for our attention mechanism to focus on artifacts of different GANs.

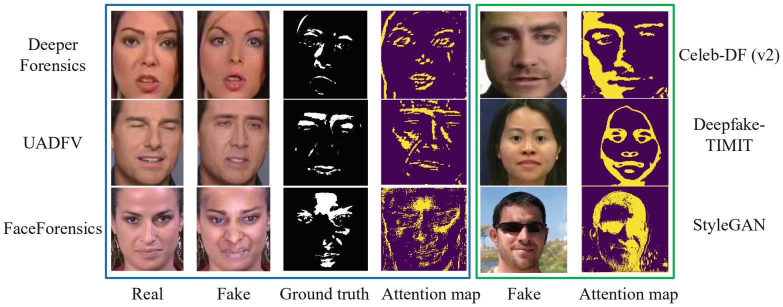
4.3 Ablation Study

Effectiveness of Dual Attention. To evaluate the effectiveness of dual attention, we compare our approach with several cases, including: (i) MesoNet, (ii) MesoNet+Att-c: MesoNet with only channel attention added to the corresponding places of each block, (iii) MesoNet+Att-s: MesoNet with only spatial attention added, (iv) MesoNet+Att.: MesoNet with dual attention added, (v) MesoNet+Conv.: an extra convolution layer is added to the first block of MesoNet to match the number of parameters with our approach. The detection accuracy and the number of trainable parameters are shown in Table 4. In general, the best results are obtained by using dual attention mechanism, while single attention mechanism improves detection accuracy slightly in some datasets and decreases in others. By adding a convolution layer, the accuracy is improved in most datasets. The average improvement is 1.47%, less than 3.50% when added with dual attention, even though it has more trainable parameters. This suggests the effectiveness of dual attention and shows that the improvement in accuracy does not entirely come from the increase in the number of parameters.

Table 5. Comparison of accuracy when our attention is added to different places of backbone models.

Approach	Position	FF-RAW	FF-HQ	FF-LQ	Celeb-DF (v2)	UADFV	DeeperF	StyleGAN	Improve(avg)
MesoNet	–	0.9616	0.9017	0.8048	0.8470	0.9439	0.9750	0.8818	–
	block1	0.9902	0.9366	0.8332	**0.8973**	**0.9671**	**0.9882**	**0.9203**	3.10%
	block2	0.9771	0.9384	0.8464	0.8913	0.9661	0.9872	0.9035	2.78%
	block3	**0.9912**	0.9294	0.8376	0.8756	0.9589	0.9820	0.9042	2.33%
	block4	0.9903	**0.9472**	**0.8596**	0.8770	0.9610	0.9882	0.9123	**3.14%**
Meso-Inception	–	0.9877	0.9170	0.8411	0.8616	0.8869	0.9892	0.9153	–
	block1	**0.9921**	0.9458	**0.8728**	0.8794	0.9293	**0.9930**	**0.9395**	**2.19%**
	block2	0.9918	0.9399	0.8568	0.8890	0.9311	0.9900	0.9255	1.79%
	block3	0.9918	0.9183	0.8662	0.8773	0.9066	0.9858	0.9093	0.81%
	block4	0.9918	**0.9534**	0.8348	**0.8936**	**0.9567**	0.9910	0.9187	2.02%
VGG-19	–	0.9918	0.9371	0.8647	0.9284	0.8734	0.9882	0.9798	-
	block1	0.9918	0.9530	0.8199	0.9465	0.9003	0.9918	**0.9930**	0.47%
	block2	0.9918	0.9274	0.8600	0.9348	0.8938	0.9910	0.9685	0.06%
	block3	0.9915	0.9427	0.8508	0.9141	0.9217	0.9895	0.9832	0.43%
	block4	0.9918	0.9453	0.8458	0.9373	0.9049	0.9840	0.9758	0.31%
	block5	0.9918	**0.9566**	**0.8713**	**0.9477**	**0.9331**	**0.9932**	0.9848	**1.64%**
Xception	–	0.9915	0.9624	0.8502	0.9611	0.9606	0.9972	0.9892	–
	block1	0.9918	0.9668	**0.9022**	**0.9717**	**0.9801**	**0.9988**	**0.9962**	**1.36%**
	block4	0.9918	0.9627	0.8650	0.9692	0.9681	0.9982	0.9962	0.56%
	block12	0.9918	0.9652	0.8584	0.9674	0.9795	0.9985	0.9762	0.35%
	block13	0.9918	**0.9718**	0.8863	0.9652	0.9634	0.9975	0.9918	0.79%
EfficientNet	–	0.9918	0.9719	0.8706	0.9683	0.9812	0.9968	0.9965	–
	stem	0.9918	0.9747	0.8715	0.9720	0.9819	0.9975	0.9960	0.12%
	middle	0.9918	0.9775	0.8742	0.9685	0.9842	0.9978	0.9965	0.19%
	top	0.9918	**0.9859**	**0.9048**	**0.9756**	**0.9845**	**0.9982**	**0.9985**	**0.89%**

Ablation Study on the Depth of Attention Module. We further investigate the effects of our attention mechanism when added to different depths of different backbone models. For MesoNet, Meso-Inception, and VGG-19, we add the attention module to the end of each block respectively. Since Xception and EfficientNet are much deeper, we select only four typical blocks of Xception, representing the beginning of entry flow (block1), middle flow (block4), exit flow (block12), and the ending of exit flow (block13), and three positions of EfficientNet that are before the first block (stem), between block3 and 4 (middle), and after the last block (top). Attention is added to these places respectively. Then we train and test each model with different placements of attention on each dataset. Results in Table 5 shows that accuracy of models with attention added is relatively higher than the original ones regardless of the placements, indicating that the attention mechanism can indeed improve the detection performance of different backbone models. Generally, it is more effective when attention is added to the beginning and exit flow, especially for deeper neural networks such as Xception. Moreover, since the feature extraction capability is relatively weak for shallow networks, attention can be added in multiple locations to enhance the detection effectiveness. For example, the accuracy is improved by 3.50% with attention added to all blocks of MesoNet (refer to Table 2), better than adding to any single block.



(a) Datasets with ground truth. (b) Datasets without ground truth.

Fig. 6. Visual effects of localization. For DeeperForensics, UADFV and FaceForensics, the attention maps are visually consistent with the ground truth. As for Celeb-DF (v2), Deepfake-TIMIT, and StyleGAN, attention is concentrated on facial areas.

Table 6. Comparison with FFD.

Approach	Accuracy ↑	PBCA ↑	IINC ↓
Xception	0.9729	–	–
FFD	0.9752	0.8289	0.6574
MesoNet	0.9664	–	–
Our approach	**0.9753**	**0.8315**	**0.6454**

4.4 Manipulation Localization

Our attention mechanism can focus on manipulated pixels of the synthesized images, hence we utilize the spatial attention from the first convolution layer for localization. Each pixel in the attention map corresponds to the pixel in the original image. To quantify the localization effectiveness, we need to compare our attention map with the ground truth, thus we only evaluate *DeepFake* images in datasets where ground truth can be obtained. We choose PBCA (Pixel-wise Binary Classification Accuracy) and IINC (Inverse Intersection Non-Containment) [8] as the evaluation metrics. Higher PBCA and lower IINC means better performance. As shown in Table 1, we have a considerable PBCA of more than 70% and IINC less than 70% on all datasets. And the visual effects of our attention maps are shown in Fig. 6. For datasets with ground truth, our attention maps are visually consistent with the ground truth maps. As for datasets without ground truth, attention is concentrated on facial areas, indicating that artifacts mainly exist in facial areas for generated videos and images. More visualization results can be found in Appendix B.

We also compare our localization effectiveness with previous work by Dang et al. [8] (referred to as FFD). The detection accuracy, PBCA, and IINC on *faceapp*, a dataset provided in Dang's work, are reported in Table 6, and the comparison of visual effects is shown in Fig. 7. Overall, the detection accuracy,

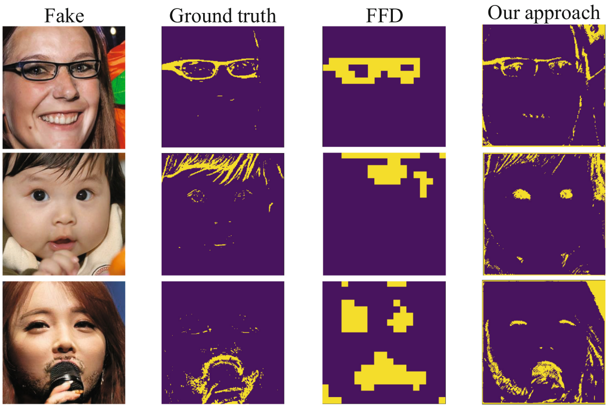


Fig. 7. Comparison of visual effects with FFD. By comparing the attention maps produced by both techniques with the ground truth, we can observe that our approach is obviously more fine-grained and precise.

PBCA, and IINC of our approach are slightly better than FFD. And though our approach is similar to FFD in terms of these metrics, the attention maps of our approach are obviously more fine-grained and precise.

5 Conclusion

In this paper, we propose a novel *DeepFake* detection approach by focusing on the observable and invisible artifact regions with dual attention. We add channel attention and spatial attention to help the neural networks to better focus on essential features and their spatial distribution in the images where artifacts mainly exist. Our proposed approach can automatically learn artifacts for different GANs, which not only helps for better detection performance but also provides a clear clue for localization. We evaluate our approach with five backbone models of different architectures and six datasets covering a diversity of GANs. Experimental results show that the detection accuracy is improved by 3.50%, 2.56%, 1.64%, 1.36%, and 0.89% in average on MesoNet, Meso-Inception, VGG-19, Xception, and EfficientNet respectively, and our attention mechanism can perform pixel-wise manipulation localization.

Acknowledgement. This work was supported in part by National Natural Science Foundation of China under Grant U1836112 and 61876134, the Fundamental Research Funds for the Central Universities No. 2042021kf1030.

A Dataset Details

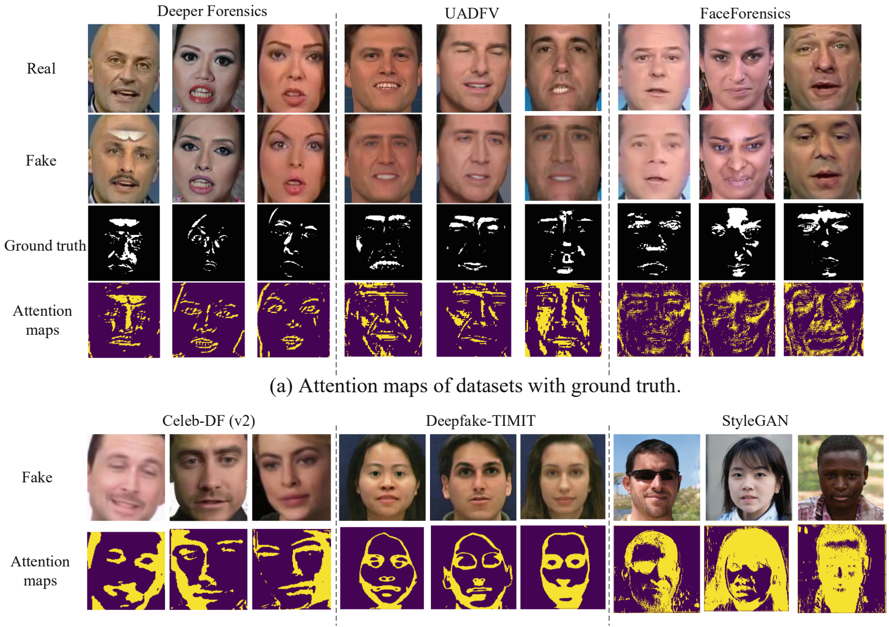
Table 7 details the number of training and testing videos and images of the six DeepFake datasets in our experiments.

Table 7. Datasets details.

Datasets	Label	Train videos	Test videos	Train images	Test images
UADFV	Fake	39	10	4,237	1,483
	Real	39	10	4,343	1,487
Deepfake-Timit HQ	Fake	288	32	6,728	703
	Real	507	52	6,591	676
Deepfake-Timit LQ	Fake	288	32	6,730	705
	Real	507	52	6,591	676
FaceForenscics RAW	Fake	135	15	1,2736	1,618
	Real	135	15	12,721	1,559
FaceForenscics HQ	Fake	135	15	12,736	1,618
	Real	135	15	12,720	1,576
FaceForenscics LQ	Fake	135	15	12,712	1,618
	Real	135	15	12,699	1,553
Celeb-DF (V2)	Fake	5,299	340	26,429	1,695
	Real	712	178	10,579	2,646
DeeperForensics	Fake	900	100	1,8000	2,000
	Real	900	100	18,000	2,000
StyleGAN	Fake	–	–	18,000	2,000
	Real	–	–	18,000	2,000

B Visualization

Figure 8 presents the visualization results of our attention maps. For datasets with ground truth, including DeeperForensics, UADFV, and FaceForensics, we pair the fake images with the real ones to calculate the manipulated pixels. And then we calculate the manipulated pixels from our spatial attention mechanism for localization. Experimental result shown that our attention maps are visually consistent with the ground truth. As for datasets without ground truth, we also calculate the attention maps obtained from the spatial attention mechanism to understand the artifact regions of these images. Results show that attention is concentrated on facial areas for Celeb-DF (v2), Deepfake-TIMIT, and StyleGAN, which are the main artifact regions.



(a) Attention maps of datasets with ground truth.

(b) Attention maps of datasets without ground truth.

Fig. 8. Visualization results of localizing the manipulated artifacts.

References

1. Deepfacelab. https://github.com/iperov/DeepFaceLab. Accessed 20 Apr 2021
2. Faceapp. https://faceappdownload.org/. Accessed 11 Apr 2021
3. Faceswap github. https://github.com/deepfakes/faceswap. Accessed 20 Apr 2021
4. 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)
5. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: CVPR Workshops, pp. 38–45 (2019)
6. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
7. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
8. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
9. Güera, 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. IEEE (2018)

10. Islam, A., Long, C., Basharat, A., Hoogs, A.: DOA-GAN: dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4676–4685 (2020)
11. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2886–2895. IEEE (2020)
12. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
14. Korshunov, P., Marcel, S.: DeepFakes: a new threat to face recognition? Assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
15. Li, L., Bao, J., Zhang, T., Yang, H., Guo, B.: Face x-ray for more general face forgery detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
16. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
17. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: a new dataset for deepfake forensics. arXiv preprint arXiv:1909.12962 (2019)
18. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 83–92. IEEE (2019)
19. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics: a large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 (2018)
20. 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)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
24. Wang, S.Y., Wang, O., Owens, A., Zhang, R., Efros, A.A.: Detecting photoshopped faces by scripting photoshop. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10072–10081 (2019)
25. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot... for now. arXiv preprint arXiv:1912.11035 (2019)
26. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
27. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265. IEEE (2019)