An Information Theoretic Approach for Attention-Driven Face Forgery Detection

Ke Sun[1,3], Hong Liu[2], Taiping Yao[3], Xiaoshuai Sun[1,4(✉)], Shen Chen[3], Shouhong Ding[3(✉)], and Rongrong Ji[1,4]

[1] Media Analytics and Computing Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China
xssun@xmu.edu.cn
[2] National Institute of Informatics, Tokyo, Japan
[3] Youtu Lab, Tencent, Shanghai, China
skjack@stu.xmu.edu.cn
[4] Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China

Abstract. Recently, Deepfake arises as a powerful tool to fool the existing real-world face detection systems, which has received wide attention in both academia and society. Most existing forgery face detection methods use heuristic clues to build a binary forgery detector, which mainly takes advantage of the empirical observation based on abnormal texture, blending clues, or high-frequency noise, etc.. However, heuristic clues only reflect certain aspects of the forgery, which might lead to model bias or suboptimization. Our recent observations indicate that most of the forgery clues are hidden in the informative region, which can be measured quantitatively by the classic information maximization theory. Motivated by this, we make the first attempt to introduce the self-information metric to enhance the feature representation for forgery detection. The proposed metric can be formulated as a plug-and-play block, termed self-information attention (SIA) module, which can be integrated with most of the top-performance deep models to boost their detection performance. The SIA module can explicitly help the model locate the informative regions and recalibrate channel-wise feature responses, which improves both model's performance and generalization with few additional parameters. Extensive experiments on several large-scale benchmarks demonstrate the superiority of the proposed method against the state-of-the-art competitors.

Keywords: Face forgery detection · Information maximization · Attention mechanism

1 Introduction

Recently, face forgery generation methods have received lots of attention in the computer vision community [12,21,40,46,49,53], which may cause severe trust

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19781-9_7.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13674, pp. 111–127, 2022.
https://doi.org/10.1007/978-3-031-19781-9_7

issues and seriously disturb the social order. For example, the producer can forge the video of world leaders to influence or manipulate politics and social sentiment. Even worse, these fake videos are of high quality and can be easily generated by open-source codes like DeepFaceLab. Therefore, it is urgent to develop effective face forgery detection methods to mitigate malicious abuse of face forgery.

A simple way is to model face forgery detection as a binary classification problem [1,12,16,44]. Basically, a pretrained convolutional neural network (CNN) is used to distinguish the authenticity of the input face, which is a golden standard in Deepfake-Detection-Challenge [12]. However, the generated fake faces become more and more authentic, which means the differences between real and fake faces are more subtle. Although CNN models possess discriminative features, it is still hard to directly use such models to capture those forgery clues in a unified framework, resulting in unsatisfactory performance.

To tackle this issue, many heuristic methods [7,11,17,23,32,45] usually use the prior knowledge or observed clues to learn more discriminative features, which are instrumental in distinguishing real and fake faces. For example, F3-Net [39] learns forgery patterns with the awareness of frequency, Gram-Net [32] leverages global image texture representations for robust fake image detection, and Face X-ray [27] takes advantage of blending boundary for a forged image to enhance the performance. Though these methods can help to improve the performance, these heuristic methods lack unified theoretical support and only reflect certain aspects of the face forgery, leading to model bias or sub-optimization.

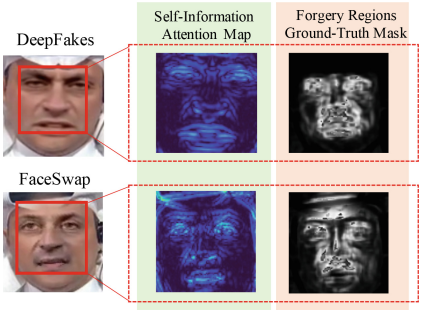

Fig. 1. Visualization of self-information map and manipulation ground truth mask for forgery faces by different manipulations (Deepfakes and FaceSwap). The self-information map is calculated by Eq. 1. The ground truth mask is generated by subtracting the forged faces and the corresponding real faces with some morphological transformations

To address this issue, we revisit face forgery detection from a new perspective, *i.e.*, face forgery is highly associated with high-information content. Inspired by [5,6], we make the first attempt to introduce self-information as a theoretic guidance to improve the discriminativeness of the model. Specially, self-information can be easily defined by the current or surrounding regions [43], where a high-information region is significantly different from their neighborhoods that can reflect the amount of information of the image content. Moreover, we find that most existing clues are always in high self-information regions. For example, due to the instability of the generative model, some abnormal textures always appear in forgery faces. These high-frequency artifacts are often very different from the surrounding facial features or skin, where the self-information can highlight these clues. Another example is blending artifacts. Face x-ray [27] demonstrate that the forged boundary is widely existed in forgery faces because

of blending operation. The skin color or texture difference between real and forgery part enlarge the self-information in blending artifacts regions. Motivated by this observation, we design a novel self-information attention module called Self-Information Attention (SIA), which calculates pixel-wise self-information and uses it as a spatial attention map to capture more subtle abnormal clues. Additionally, the SIA module calculates the average self-information of each channel's feature map and uses it as attention weights to select the most informative feature map. As shown in Fig. 1, the self-information map of the original image highlight the same region as the ground truth mask, which indicates its effectiveness in face forgery detection task.

We conduct our experiments on several widely-used benchmarks. And experimental results show that our proposed method significantly outperforms the state-of-the-art competitors. Particularly, the proposed SIA module can be flexibly plugged into most CNN architectures with little parameter increase. Our main contributions can be summarized as follows:

- We propose a new perspective for face forgery detection based on information theory, where self-information is introduced as a theoretic guidance for detection models to capture more critical forgery cues.
- We specially design a novel attention module based on self-information, which helps the model capture more informative regions and learn more discriminative features. Besides, the SIA attention can be plugged into most existing 2D CNNs with negligible parameter increase.
- Extensive experiments and visualizations demonstrate that our method can achieve consistent improvement over multiple competitors with a comparable amount of parameters.

2 Related Work

2.1 Forgery Face Manipulation

Face forgery generation methods have a security influence on scenarios related to identity authentication, which achieve more and more attention in computer vision communities. In particular, *deepfakes* is the first deep learning based face identity swap method [49], which uses two Autoencoders to simulate changes in facial expressions. The other stream of research is to design GAN based models [4,14,15,21] for generating entire fake faces. Recently, graphics-based approaches are widely used for identity transfer, which are more stable compared with deep learning based approaches. For instance, Face2Face [50] is can operate face swap using only an RGB camera in real-time. Averbuch-Elor *et al.* [3] proposed a reenactment method that deforms the target image to match the expressions of the source face. NeuralTextures [48] renders a fake face via computing reenactment result with neural texture. Kim *et al.* [25] combined image-to-image translation network with computer graphics renderings to convert face attributes. These forgery methods focus on manipulate high-information areas and may leave some high-frequency subtle clues, thus we introduce self-information learning to assist in identifying forged faces.

2.2 Face Forgery Detection

To detect the authenticity of input faces, early works usually extract low-level features such as RGB patterns [36], inconsistency of JPEG compression [2], visual artifacts [35]. More recently, binary convolution neural network has been widely used to this task [12] and achieve better performance. However, directly using vanilla CNN tend to extract semantic information while may ignore the subtle and local forgery patterns [54]. Thus, some heuristic methods are proposed, which leverage observation or prior knowledge to help model to mine the forgery pattern. For instance, Face X-ray [27] is supervised by the forged boundary. F3-Net [39] leverage frequency clues as to the auxiliary to RGB features. Local-Relation [9] measures the similarity between features of local regions based on the observation of inconsistency between forgery parts and real parts. However, these methods still cannot cover all the forgery clues, leading to sub-optimal performance. Thus, we introduce self-information to help the model capture informative region adaptively. In addition, our proposed method only contains a few parameters and can serve as a plug-and-play module upon several backbones.

3 Proposed Method

3.1 Preliminaries

Problem Formulation. Many works [24] have been proposed to identify a given face, real or fake, but most of them are based on experimental observations that show the remarkable difference between real and fake faces. Recent work [35] found that these observations belong to the discriminative artifacts clues that are subtle but abnormal compared with their neighborhoods, because of the generative model's instability and the imperfection of the blending methods. On the other hand, existing models just consider one or a small number of these different clues, which are integrated into the vanilla CNN, leading to bias or sub-optimization model. These raise a natural question that, *is there a metric that can adaptively capture differential information?* To answer this question, this paper focuses on the information theory and uses classical self-information to adaptively qualify the saliency clues.

Self-information Analysis. The self-information is a metric of the information content related to the outcome of a random variable [5], which is also called *surprisal*, *i.e.*, it can reflect the surprise of an event's outcome. Given a random variable X with probability mass function P_X, the self-information of X as outcome x is $I_X(x) = -log(P_X(x))$. As a result, we can derive that the smaller its probability, the higher the self-information it has. That is, the more different the region from its neighboring patches, the more self-information it contains. Inspired by [5,43], self-information can apply to the joint likelihood of statistics in a local neighborhood of the current patch, which provides a transformation between probability and the degree of information inherent in the

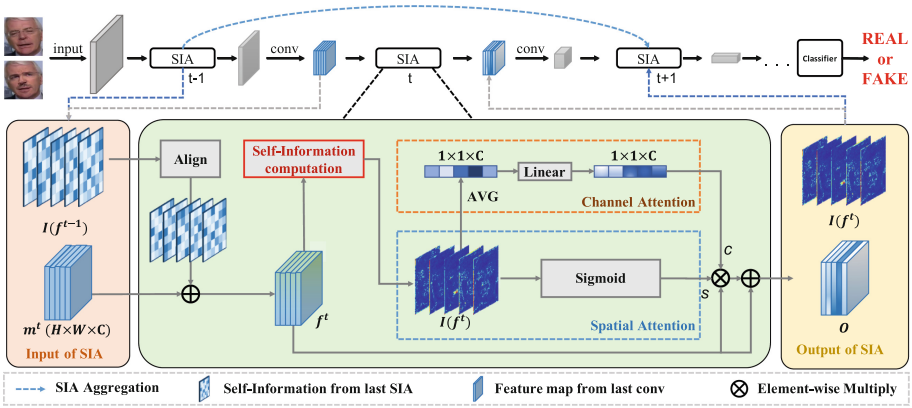

Fig. 2. The overview of our face forgery detection framework with Self-information attention (SIA) module. The SIA module is embedded in the middle layer of CNN. The orange dotted block means the channel attention part, while the blue dotted block represents the spatial attention part. The details of self-information is shown in Fig. 3 (Best viewed in color) (Color figure online)

local statistics. For face forgery detection, the heuristic unusual forgery clues (such as high-frequency noise, blending boundary, abnormal textures, *etc.*) are hidden in the high-information context. Therefore, it is intuitive to introduce the self-information metric into face forgery detection to help model additively learn high-information features.

3.2 Overall Framework

In this paper, we design a new attention mechanism, that is based on the self-information metric, which could highlight the manipulated regions. We call this newly defined model as Self-Information Attention (SIA) module, whose overview framework is shown in the Fig. 2. In particular, the proposed SIA module mainly contains three key parts: 1) **Self-Information Computation**: To capture the high-information content region, we calculate the self-information from the input feature map and output a new discriminative attention map. 2) **Self-Information based Dual Attention**: To maximize the ability of using self-information by backbone model, the self-information from the input feature map would be used on both channel-wise attention and spatial-wise attention. 3) **Self-Information Aggregation**: Motivated by [19,54], we densely forward all previous self-information feature maps to the current SIA block, which is to preserve the detail area to the greatest extent.

3.3 Self-information Computation

Let $f^t \in R^{C\times H\times W}$ denotes the input of the t-th SIA module with C channels and spatial shape of $H \times W$, where $f_k^t(i,j)$ denotes the k-th channel's

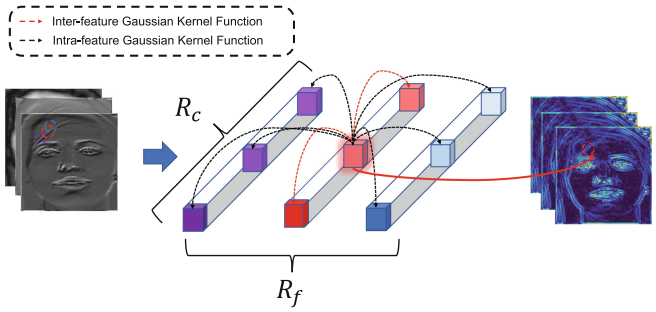

Fig. 3. Visualization of Self-Information Computation. R_f denotes the local receptive filed region, and R_c is the channel offset region (Best viewed in color)

pixel of f^t located by the coordinate (i, j). As mentioned before [5], the self-information can be approximated by the joint probability distribution of the current pixel together with its neighborhoods with Gaussian kernel function. Different with previous work [5], we consider the self-information through two orthogonal dimensions, one is to find the neighborhoods in the spatial dimension, and the other is to search the neighborhoods in channel dimension.

We define the spatial space intra-feature self-information as:

$$I_{\text{intra}}(f_k^t(i,j)) = -\log \sum_{m,n \in R_f} e^{-\frac{||f_k^t(i,j) - f_k^t(i+m,j+n)||^2}{2h^2}}, \tag{1}$$

where R_f are the local receptive filed region near the pixel (i, j), m and n are the pixel indexes in the R_f, and h is the bandwidth.

When the neighborhoods are located in the channel dimension, we define the self-information in channel as inter-feature self-information I_{inter}, which is shown as:

$$I_{\text{inter}}(f_k^t(i,j)) = -\log \sum_{s \in R_c} e^{-\frac{||f_k^t(i,j) - f_{k+s}^t(i,j)||^2}{2h^2}}, \tag{2}$$

where s is the index of the channel offset region R_c. The inter-feature self-information could help us avoid some observation noise that exists in the channels.

As a result, the whole self-information $I(f_k)$ can be formulated as:

$$I(f_k^t(i,j)) = I_{\text{intra}}(f_k^t(i,j)) + \lambda I_{\text{inter}}(f_k^t(i,j)), \tag{3}$$

where λ is the weight parameter that balance the importance of the inter-feature self-information. The Fig. 3 illustrates the computation of self-information.

3.4 Self-information Based Dual Attention

We propose a new dual attention model, where the saliency is qualified by the self-information measure [6]. Inspired by [20], we consider the saliency features through spatial dimension and channel dimension.

Spatial-Wise Attention Module. We introduce a spatial attention module based on self-information, as the flowchart shown in Fig. 2. In detail, we calculate each pixel's self-information features $I(f_k^t)$ via Eq. 3. Then, we use the Sigmoid function to normalize such features and output the self-information based spatial attention map. Finally, we perform an element-wise multiply operation with the input feature f_k^t. The whole formulation of the Spatial-wise Attention Module is shown as follows:

$$s_k = \text{Sigmoid}(I(f_k^t)) * f_k^t. \tag{4}$$

This attention map focuses on the high-information region with little parameter improvement, which can adaptively enhance many artifact subtle clues, such as blending boundary and high-frequency noise. For more details please refer to the Sect. 4.

Channel-Wise Attention Module. Apart from the spatial attention module, we further introduce the channel attention module, which pipeline is illustrated in Fig. 2. Similar to [20], we calculate the average self-information of channel feature maps and generate channel-wise statistical feature c_k for the k-th element of f^t as follows:

$$c_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} I(f_k^t). \tag{5}$$

To mitigate the problem of training stability, we opt to employ a simple linear transform with sigmoid activation on vector $c = \{c_1, c_2 ... c_C\}$ as channel attention c':

$$c' = \text{Sigmoid}(Wc), \tag{6}$$

where $W \in R^{C \times C}$ is the linear function. This module could improve the self-information of the feature map that contains high-information, which helps locate the saliency in explicit contents.

Dual Attention Module Embedded in CNN. Finally, we combine the two attention modules mentioned above and perform an element-wise sum operation between the processed attention map and f_k to output a residual error feature $o_k \in R^{C \times H \times W}$, which formulated as:

$$O_k = c'_k * s_k + f_k^t. \tag{7}$$

The proposed SIA module is a flexible module, which can be easily inserted into any CNN-based architecture. Also, we can also flexibly choose the spatial attention module and the channel attention module. The SIA module does not increase many parameters yet can enhance the performance of the model.

3.5 Self-information Aggregation

General CNN such as EfficientNet [47] usually use down-sampling operations to reduce the parameters and expand the receptive field, which tend to eliminate subtle clues with high information content in face forgery detection task.

To overcome this problem, inspired by [22], we design a self-information aggregation operation, cascading different levels of SIA modules via self-information attention map. Thus the local and subtle forgery clues can be preserved. As shown in Fig. 2, we add the attention map of the previous stage with the current input feature map to preserve the shallow high informative texture. Due to the different sizes of attention maps at different levels, we use 1×1 convolution to align the number of channels and use the interpolation method to align the size of the feature map. This alignment operation could be presented as the function Align. As a result, the t-th input feature f^t can be defined as:

$$f^t = \sum_{i=1}^{t-1} \text{Align}_i(I(f^i)) + m^t, \tag{8}$$

where m^t is the feature map adjacent to the t-th SIA module.

3.6 Loss Function

We use the Cross-entropy as loss function, which is defined as:

$$L_{ce} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i}), \tag{9}$$

where n is the number of images, $\hat{y_i}$ is the prediction of the i-th fake image, and y_i is the label of the sample.

4 Experiment

In this section, we evaluate the proposed SIA module against some state-of-the-art face forgery detection methods [1,8,10,11,13,18,27,34,39,47,54] and some attention techniques [20,32,42]. We explore the robustness under unseen manipulation methods and conduct some ablation studies, and further give some visualization results.

4.1 Experimental Setup

Datasets. We conduct our experiments on several challenging dataset to evaluate the effectiveness of our proposed method. **FaceForensics++** [40] is a large-scale deepfake detection dataset containing 1,000 videos, in which 720 videos are used for training and the rest 280 videos are used for validation or testing. There are four different face synthesis approaches, including two graphics-based methods (*Face2Face* and *FaceSwap*) and two learning-based approaches (*DeepFakes* and *NeuralTextures*). The videos in FaceForensics++ have two kinds of video quality: high quality (C23) and low quality (C40). **Celeb-DF** [30] is another widely-used Deepfakes dataset, which contains 590 real videos and 5,639 fake videos. In Celeb-DF, the DeepFake videos are generated by swapping faces for each pair of the 59 subjects. Following the prior works [39,47,54], we use the

Table 1. Comparison on FaceForensics++ dataset in terms of ACC and AUC with different qualities (HQ and LQ). The highest results are highlighted in bold. The F3-Net use 0.5 as a threshold

Methods	ACC (LQ)	AUC (LQ)	ACC (HQ)	AUC (HQ)
MesoNet [1]	70.47	–	83.10	–
Face X-ray [27]	–	61.60	–	87.40
Xception [10]	86.86	89.30	95.73	96.30
Xception-ELA [18]	79.63	82.90	93.86	94.80
Xception-PAF [8]	87.16	90.20	–	–
Two Branch [34]	86.34	86.59	86.34	86.59
EfficientNet-B4 [47]	86.95	88.91	96.63	99.18
F3-Net [39]	86.89	93.30	97.31	98.10
MAT [54]	88.69	90.40	97.60	99.29
SPSL [31]	81.57	82.82	91.50	95.32
RFM [51]	87.06	89.83	95.69	98.79
Freq-SCL [26]	89.00	92.39	96.69	99.28
Ours	**90.23**	**93.45**	**97.64**	**99.35**

Table 2. Cross-dataset evaluation from FF++ (LQ) to deepfake class of FF++ and Celeb-DF in terms of AUC. The highest results are highlight in bold

Methods	FF++	Celeb-DF
Two-stream [55]	70.10	53.80
Meso4 [1]	83.00	53.60
FWA [29]	80.10	56.90
DSP-FWA [29]	93.00	64.60
Xception [40]	95.50	65.50
EN-b4 [12]	96.39	71.10
Multi-task [37]	76.30	54.30
Capeule [38]	96.60	57.50
SMIL [28]	96.80	56.30
Two Branch [34]	93.18	73.41
MAT [54]	96.41	72.50
GFF [33]	95.73	74.12
SPSL [31]	96.91	76.88
Ours	**96.94**	**77.35**

multi-task cascaded CNNs to extract faces, and we randomly select 50 frames from each video to construct the training set and test set. **WildDeepfake** [56] is a recently released forgery face dataset that contains 3805 real face sequences and 3509 fake face sequences, which is obtained from the internet. Therefore, wild deepfake has a variety of synthesis methods, backgrounds, and ids.

Evaluation Metrics. We apply accuracy score (ACC) and area under the receiver operating characteristic curve (AUC) as our basic evaluation metrics.

Implementation Details. We use EfficientNet-b4 [47] pretrained on the ImageNet as our backbones, which are widely used in face forgery detection. The backbone contains seven layers, and we put our proposed SIA module in the output of layer1, layer2, and layer4. This is due to the shallow and middle layers contain low-level and middle-level features, which reflect the subtle artifact clues well. We resize each input face to 299×299. The hyperparameters λ in Eq. 3 is set to 0.5. We use Adam optimizer to train the network's parameters, where the weight decay is equal to $1e-5$ with betas of 0.9 and 0.999. The initial learning rate is set to 0.001, and we use StepLR scheduler with 5 step-size decay and gamma is set to 0.1. The batch size is set to 32.

4.2 Experimental Results

Intra-dataset Testing. We evaluate the performance under two quality settings on FaceForensics++. Note that the results of F3-Net use a threshold of

Table 3. Performance on Celeb-DF and WildDeepfake datasets in terms of ACC and AUC

Method	Celeb-DF		WildDeepfake	
	ACC	AUC	ACC	AUC
Xception [10]	97.90	99.73	77.25	86.76
EfficientNet-B4 [47]	97.63	99.20	81.63	90.36
RFM [51]	97.96	99.94	77.38	83.92
F3-Net [39]	95.95	98.93	80.66	87.53
MAT [54]	97.84	99.81	82.86	90.71
Ours	**98.48**	**99.96**	**83.95**	**91.34**

Table 4. ACC of different pretrained backbones on FaceForensics++ HQ, FaceForensics++ LQ and Celeb-DF datasets

Backbone	FF++ (HQ)	FF++ (LQ)	Celeb-DF
EffecinetNet-b0	96.32	85.66	97.81
EffecinetNet-b0+Ours	**96.95**	**87.03**	**98.37**
XceptionNet	95.73	86.86	97.90
XceptionNet+ours	**96.85**	**87.30**	**98.03**
MobileNet-v2	95.58	85.19	97.41
MobileNet-v2+Ours	**97.05**	**86.78**	**98.22**

0.5. The overall results in Table 1 show that the proposed method obtains state-of-the-art performance on the both high-quality and low-quality settings. Compared Freq-SCL [26] and SPSL [31] leverage frequency clues as to the auxiliary to RGB features, both of which convert RGB image into the frequency domain together with a dual-stream framework. Both two methods boost performance via input perspective, but our method takes consideration of promoting representation learning. Compared with the recent attention based Multi-attentional [54], our method achieves better performance. This is because that SIA gives more accurate guidance for the attention mechanism on both channel-wise and spatial-wise dimensions, which provide more adaptive information for the model.

To further demonstrate the effectiveness of our method, we evaluate the SIA module on two famous forgery datasets: Celeb-DF and WildDeepfake. The results are shown in Table 3. We can observe that our SIA outperforms all comparison methods. Specifically, compared with F3-Net which requires 80M parameters and 21G macs, our EN-b4+SIA only contains 35M parameters and 6.05G macs and achieves about 3% improvement on both two datasets. In addition, we evaluate the proposed SIA module on DFDC [12] dataset and achieve SOTA performance with 82.31% in terms of ACC and 90.96% in terms of AUC. Due to the page limit, we put the results in the supplementary material.

Cross-Dataset Testing. To further demonstrate the generalization of SIA, we conduct cross-dataset evaluations. Specifically, following the setting of [34], we train our model on FF++ (LQ) and test it on Deepfakes class and Celeb-DF. The quantitative results are shown in Table 2, we can observe that our method obtain state-of-the-art performance especially in cross-database setting. Our SIA outperforms by 4% and 2% in terms of AUC compared with the recent SPSL and GFF on cross-dataset setting and achieve slight improvement on the intra-dimain setting. The reason for the improvement is that our module guide the backbone focuses on the informative subtle details which are commonly present on all forgery face.

Dependency on Backbone. The proposed SIA module is a plug-and-play block, which can be embedded in any deep learning based model. Therefore, we verify the effectiveness of the SIA module by using different backbones. We select

Table 5. Quantitative results on Celeb-DF and FaceForensics++ dataset with different qualities (HQ and LQ). The compared methods are all plug-and-play attention modules. The last column represents the parameter increase after adding the corresponding module. The highest results are highlighted in bold

Method	FF++ (HQ)			FF++ (LQ)			Celeb-DF			Parameter increase
	ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER	
Baseline+Selayer	97.05	99.20	4.04	89.05	90.99	19.01	98.00	99.65	2.15	45k
Baseline+NL	96.79	99.16	4.20	89.55	91.34	19.50	97.89	99.61	2.20	764k
Baseline+Selayer+NL	97.21	99.25	3.95	89.78	91.56	18.98	97.80	99.68	2.04	811k
Baseline+GSA	94.68	97.92	7.03	88.92	89.03	22.32	97.88	99.63	2.63	1024k
Ours	**97.64**	**99.35**	**3.83**	**90.23**	**93.75**	**18.57**	**98.48**	**99.96**	**1.87**	**42k**

the EffecinetNet-b0, MobileNet-v2, and XceptionNet as other backbones, and we evaluate the results on FaceForensics++ and Celeb-DF. All the SIA module is embedded in the first and middle layer. For instance, for the EffiecientNet-b0, we put out the module after the 2th, 3th and 5th MBConvBlock. For the XceptionNet, our module is inserted between 3th block and 4 block. As for the MobileNet-v2, the module is embedded after the 3th and 7th InvertedResidual block. The result is shown in Table 4. We find that our methods do improve the network performance regardless of the types of backbones, which proves the flexibility and generality of our method.

Compared with Attention Methods. We compare the proposed method with several classical attention-based methods to show the effectiveness of self-information in this task: **(1) Baseline**: The EffecinetNet-b4 pretrained on the ImageNet. **(2) Baseline+SE-layer** [20]: The channel attention module. **(3) Baseline+Non-local** [52]: Non-local attention has been used in deepfake detection [32]. Here we use the Gaussian embedded version with both batchnorm layer and sub-sample strategy. **(4) Baseline+SE-layer+Non-local**: We use the SE-layer and Non-local to realize both channel attention and spatial attention module. **(5) Baseline+GSA** [42]: GSA is the state-of-the-art attention module that considers both pixel content and spatial information. Here the number heads are set to 8, and the dimensional key is set to 64.

The comparison results are reported in Table 5. The results show that our proposed SIA module outperforms all the reference methods on both two benchmarks. Specifically, after adding our SIA module on the baseline, the performance has about 1.5% ACC improvement with little parameters increase. This reflects that the self-information does fit for face forgery detection task.

4.3 Ablation Study

Impact of Different Components. To further explore the impact of different components of SIA module, we split each part separately for experimental verification. The ACC and AUC results on FF++ (LQ) are shown in Table 6. The results demonstrate that the three key components have a positive effect

Table 6. Abalation study on FaceForensics++(LQ) dataset

Spatial	Channel	Aggregation	ACC	AUC
✓			89.65	91.95
	✓		89.83	91.34
✓	✓		89.90	92.43
✓	✓	✓	90.23	93.75

Table 7. Comparative experiment of module insertion position

Layers	ACC on FF++	AUC on FF++
L1, L2, L3	89.56	92.73
L1, L2, L4	**90.23**	**93.75**
L3, L4, L5	89.14	91.15
L1, L4, L6	88.35	90.21

on the performance, all of which are necessary for the face forgery detection. Among them, spatial attention has a relatively large impact on performance, which demonstrate the importance of capturing high-information regions for the face forgery task.

Impact of Embedding Layer. We further conduct some ablation experiments to explore the effect of insertion place of our method. The attention module is embedded in different layers of EffiecientNet-b4 and tested on the FaceForensics++ LQ dataset. The results on the left of Table 7 show that the best performance is achieved when the attention module is embedded in layer1, layer 2 and layer4, which is in the shallow and middle of the backbone. The SIA module is derived from the theory of self-information (SI), which is usually built on the shallow structural and textural features. Therefore, it is intuitive to insert the SIA module in the shallow layers, which helps enhance SI. In the middle layers, SIA module helps reduce the global inconsistency bringing from long-range forgery patterns and pass the useful local and subtle forgery information via the self-information aggregation scheme. However, in deeper layers, down-sampling operation will neglect many local and subtle forgery information, in which SI can hardly find useful cues for forgery detection. In sum, it is natural and reasonable to plug SIA into either shallow or middle layers (L1, L2, L4), and our experiments indeed had verified this.

4.4 Visualization and Analysis

Analysis on SIA Module. To analysis our attention module, we visualize the feature maps from different channels sorted by channel-wise attention weight and the highest weight channel's SIA map. Figure 4 shows the result (all visualizations are colored according to the normalized feature map). We can observe that the channels with high self-information contain more local high-frequency clues and subtle details, while the lower ones have more semantic information and smoother clues which is less helpful for the face forgery detection task. In addition, the self-information based attention map enhances high-information areas such as mouth, eyes, high-frequency textures, and blending boundary, while weakening repetitive low-frequency areas. These visualizations demonstrate that our SIA module can effectively mine the informative channels and subtle clues, which are critical for performance improvement.

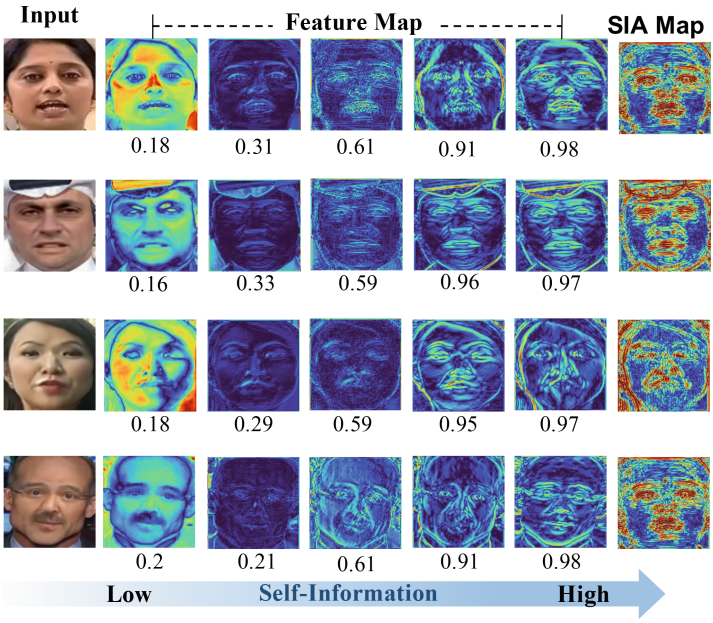

Fig. 4. Visualization of our channel-wise attention scores and their corresponding SIA maps. The feature maps are sorted according to the attention weights from low to high. The last column shows the SIA map calculated by the highest channel score feature map (Best viewed in color)

Visualization of Grad-CAM. We apply Grad-CAM [41] and Guided Grad-CAM tools to the baseline model and our model, which are widely-used methods to explain the attention of deep neural networks. The Grad-CAM can identify the regions that the network considers import, while Guided Grad-CAM can reflect more details of activation. Through Fig. 5, we can observe that our module helps the network to capture more subtle artifacts compared with the baseline backbone. The red circle indicates the obvious high-information forgery details. We also find that the baseline model ignores these artifacts (white circle) while our SIA module helps networks pay more attention to these clues. For example, the forgery face in the fourth line has an obvious blending boundary, but the baseline CAM does not pay attention to this area. After going through our SIA module, the network clearly focuses on this high-information area. Furthermore, the activation area of the guided grad-cam is larger than the baseline, because our module help the network enhances the most informative channel.

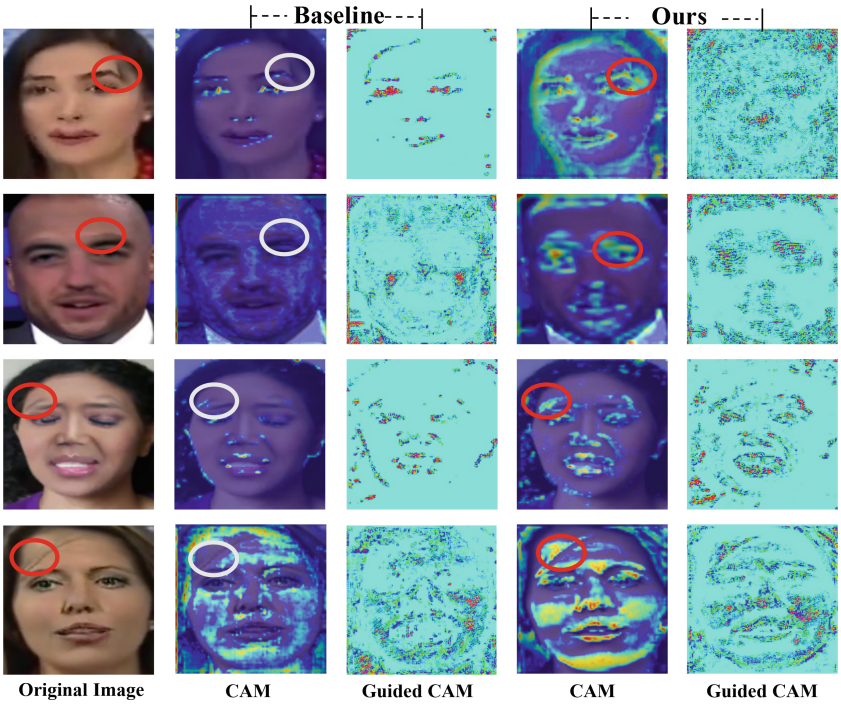

Fig. 5. Grad CAM and Guided Grad CAM on baseline model and our proposed model (layer 1 of EffiecientNet-B4). The red circles indicate obvious clues that are ignored by previous approach but well captured by our method (Best viewed in color) (Color figure online)

5 Conclusion

In this work, we propose an information theoretic framework with self-Information Attention (SIA) for effective face forgery detection. The proposed SIA module has a strong theoretic basis, which leads to an effective and interpretable method that can achieve superior face forgery detection performance with negligible parameter increase. Specially, self-information of each feature map is extracted as the bases of dual attention to help model capture the informative regions which contains critical forgery clues. Experiments on several datasets demonstrate the effectiveness of our method.

Future Work. Currently, we only evaluate our SIA module on the RGB domain. In future work, we will evaluate it in the frequency domain to further demonstrate its effectiveness and generality.

Acknowledgments. This work was supported by the National Science Fund for Distinguished Young Scholars (No. 62025603), the National Natural Science Foundation of China (No. U21B2037, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, and No. 62002305), Guangdong Basic and Applied Basic Research Foundation (No. 2019B1515120049, and the Natural Science Foundation of Fujian Province of China (No. 2021J01002).

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: MesoNet: a compact facial video forgery detection network. In: WIFS, pp. 1–7. IEEE (2018)
2. Agarwal, S., Farid, H.: Photo forensics from JPEG dimples. In: WIFS, pp. 1–6. IEEE (2017)
3. Averbuch-Elor, H., Cohen-Or, D., Kopf, J., Cohen, M.F.: Bringing portraits to life. ACM Trans. Graph. (TOG) **36**(6), 1–13 (2017)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: Advances in Neural Information Processing Systems, pp. 155–162 (2005)
6. Bruce, N., Tsotsos, J.: Attention based on information maximization. J. Vis. **7**(9), 950–950 (2007)
7. Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X.: End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4113–4122 (2022)
8. Chen, M., Sedighi, V., Boroumand, M., Fridrich, J.: JPEG-phase-aware convolutional neural network for steganalysis of JPEG images. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 75–84 (2017)
9. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: AAAI (2021)
10. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: CVPR, pp. 1251–1258 (2017)
11. Cozzolino, D., Poggi, G., Verdoliva, L.: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 159–164 (2017)
12. Dolhansky, B., et al.: The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397 (2020)
13. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. **7**(3), 868–882 (2012)
14. Gonzalez-Sosa, E., Fierrez, J., Vera-Rodriguez, R., Alonso-Fernandez, F.: Facial soft biometrics for recognition in the wild: recent works, annotation, and cots evaluation. IEEE Trans. Inf. Forensics Secur. **13**(8), 2001–2014 (2018)
15. Goodfellow, I., et al.: Generative adversarial nets. In: NeurIPS, pp. 2672–2680 (2014)
16. Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., Yi, R.: Exploiting fine-grained face forgery clues via progressive enhancement learning. In: AAAI, vol. 36, pp. 735–743 (2022)
17. Gu, Z., et al.: Spatiotemporal inconsistency learning for deepfake video detection. In: ACM MM, pp. 3473–3481 (2021)

18. Gunawan, T.S., Hanafiah, S.A.M., Kartiwi, M., Ismail, N., Za'bah, N.F., Nordin, A.N.: Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis. Indones. J. Electr. Eng. Comput. Sci. **7**(1), 131–137 (2017)
19. Guo, Z., Yang, G., Chen, J., Sun, X.: Fake face detection via adaptive manipulation traces extraction network. Comput. Vis. Image Underst. **204**, 103170 (2021)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
21. Huang, D., De La Torre, F.: Facial action transfer with personalized bilinear regression. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7573, pp. 144–158. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_11
22. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, pp. 4700–4708 (2017)
23. Huang, Y., et al.: FakePolisher: making deepfakes more detection-evasive by shallow reconstruction. arXiv preprint arXiv:2006.07533 (2020)
24. Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., Liu, Y.: Countering malicious deepfakes: survey, battleground, and horizon. arXiv preprint arXiv:2103.00218 (2021)
25. Kim, H., et al.: Deep video portraits. ACM Trans. Graph. (TOG) **37**(4), 1–14 (2018)
26. Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y.: Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In: CVPR, pp. 6458–6467 (2021)
27. Li, L., et al.: Face X-ray for more general face forgery detection. In: CVPR, pp. 5001–5010 (2020)
28. Li, X., et al.: Sharp multiple instance learning for deepfake video detection. In: ACM MM, pp. 1864–1872 (2020)
29. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
30. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: a new dataset for deepfake forensics. arXiv preprint arXiv:1909.12962 (2019)
31. Liu, H., et al.: Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: CVPR, pp. 772–781 (2021)
32. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: CVPR, pp. 8060–8069 (2020)
33. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: CVPR, pp. 16317–16326 (2021)
34. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12352, pp. 667–684. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58571-6_39
35. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: WACVW, pp. 83–92. IEEE (2019)
36. McCloskey, S., Albright, M.: Detecting GAN-generated imagery using color cues. arXiv preprint arXiv:1812.08247 (2018)
37. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876 (2019)
38. Nguyen, H.H., Yamagishi, J., Echizen, I.: Capsule-forensics: using capsule networks to detect forged images and videos. In: ICASSP, pp. 2307–2311. IEEE (2019)

39. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: face forgery detection by mining frequency-aware clues. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 86–103. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_6
40. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: learning to detect manipulated facial images. In: ICCV, pp. 1–11 (2019)
41. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)
42. Shen, Z., Bello, I., Vemulapalli, R., Jia, X., Chen, C.H.: Global self-attention networks for image recognition. arXiv preprint arXiv:2010.03019 (2020)
43. Shi, B., Zhang, D., Dai, Q., Wang, J., Zhu, Z., Mu, Y.: Informative dropout for robust representation learning: a shape-bias perspective. In: ICML, vol. 1 (2020)
44. Stehouwer, J., Dang, H., Liu, F., Liu, X., Jain, A.: On the detection of digital face manipulation. In: CVPR (2019)
45. Sun, K., et al.: Domain general face forgery detection by learning to weight. In: AAAI, vol. 35, pp. 2638–2646 (2021)
46. Sun, K., Yao, T., Chen, S., Ding, S., Li, J., Ji, R.: Dual contrastive learning for general face forgery detection. In: AAAI, vol. 36, pp. 2316–2324 (2022)
47. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: ICML (2019)
48. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: image synthesis using neural textures. ACM Trans. Graph. (TOG) **38**(4), 1–12 (2019)
49. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. **34**(6), 183-1 (2015)
50. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: real-time face capture and reenactment of RGB videos. In: CVPR, pp. 2387–2395 (2016)
51. Wang, C., Deng, W.: Representative forgery mining for fake face detection. In: CVPR, pp. 14923–14932 (2021)
52. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR, pp. 7794–7803 (2018)
53. Wang, X., Yao, T., Ding, S., Ma, L.: Face manipulation detection via auxiliary supervision. In: Yang, H., Pasupa, K., Leung, A.C.-S., Kwok, J.T., Chan, J.H., King, I. (eds.) ICONIP 2020. LNCS, vol. 12532, pp. 313–324. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63830-6_27
54. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: CVPR (2021)
55. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: CVPRW, pp. 1831–1839. IEEE (2017)
56. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: WildDeepfake: a challenging real-world dataset for deepfake detection. In: ACM MM, pp. 2382–2390 (2020)