Representative Forgery Mining for Fake Face Detection

Chengrui Wang, Weihong Deng*
Beijing University of Posts and Telecommunications
{crwang, whdeng}@bupt.edu.cn

Abstract

Although vanilla Convolutional Neural Network (CNN) based detectors can achieve satisfactory performance on fake face detection, we observe that the detectors tend to seek forgeries on a limited region of face, which reveals that the detectors is short of understanding of forgery. Therefore, we propose an attention-based data augmentation framework to guide detector refine and enlarge its attention. Specifically, our method tracks and occludes the Top-N sensitive facial regions, encouraging the detector to mine deeper into the regions ignored before for more representative forgery. Especially, our method is simple-to-use and can be easily integrated with various CNN models. Extensive experiments show that the detector trained with our method is capable to separately point out the representative forgery of fake faces generated by different manipulation techniques, and our method enables a vanilla CNN-based detector to achieve state-of-the-art performance without structure modification. Our code is available at https://github.com/crywang/RFM.

1. Introduction

The rapid development of face manipulation technology makes the manufacture of fake face more accessible than before, which further accelerates the spread of fake facial images on social media [2, 3, 33, 34]. Meanwhile, advanced techniques make it extremely difficult for human to distinguish between real and fake face [5, 11], raising constant concerns about the credibility of digital content [26, 27, 30]. To mitigate the toll that manipulation technology takes on society, Convolutional Neural Network (CNN) is widely used to construct detector for fake face detection [35].

Unfortunately, although vanilla CNN-based fake face detector can achieve satisfactory detection performance [28, 35], it may have a different understanding of forgery against that of humans. Concretely, the experiment in Section 4.5 shows that vanilla CNN-based detector tends to check forgeries from a limited region of face, while humans usually find representative forgery over the entire face. For example, the representative forgeries of fake face generated by Deepfakes [2] and Face2Face [34] usually appear on the facial boundary, while the representative forgeries of fake face generated by StyleGAN [16] and PGGAN [15] are located in the entire face flexibly (shown in Figure 2).

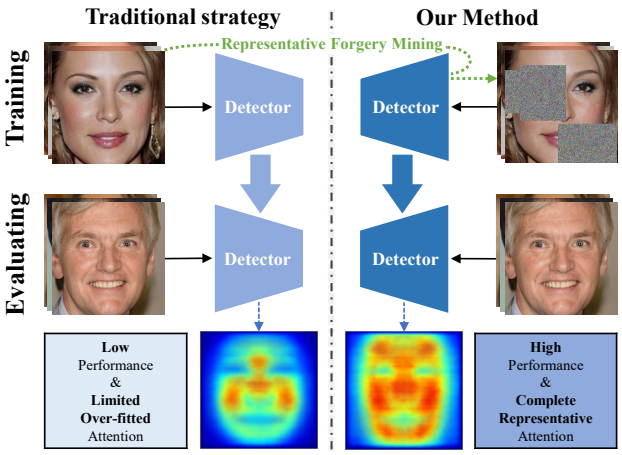

Figure 1. Comparison of detectors trained with traditional strategy and with our proposed RFM. Through refining training data, our method provides meaningful guidance to force a vanilla detector to allocate its attention to a larger and representative facial region, which greatly improves detection performance. Meanwhile, the region of interest raises a remarkable correlation to the corresponding manipulation technique.

For better detection, detectors should allocate more attention to the forgeries which can significantly represent the corresponding manipulation technique, rather than overfitting the forgeries which are mainly useful in minimizing the bi-classification loss function on training set. Recent remarkable breakthroughs [6, 8, 11, 13, 22, 40, 36, 38, 24] address this problem to some extent, which generally follow three directions: a) Through extracting the digital fingerprints produced by the defect of manipulation technique, some works [24, 36, 38, 40] achieve advanced generalization performance on CNN-generated facial images; b) Some works [6, 8] divide face into multiple patches and detect them independently, which compulsively optimizes the receptive field of detectors on fake face. c) With well-designed training dataset, some works [11, 13, 22] lever-

Figure 2. Examples of fake faces generated by different manipulation techniques. The faces generated by Deepfakes and Face2Face have forgeries mainly on the facial boundary, while the forgeries of faces generated by StyleGAN and PGGAN could appear anywhere on the face.

age the difference between real and fake faces of the same source to guide detector learn the forgeries on fake faces, which can further achieve forgery visualization.

In this paper, we propose an attention-based data augmentation method Representative Forgery Mining (RFM) to address the limited-attention problem by refining training data during training process. A brief comparison between the traditional strategy and our method is sketched in Figure 1. Concretely, our method consists of two steps including 1) using the gradient of detector to generate image-level Forgery Attention Map (FAM), which can precisely locate the sensitive facial region, and 2) utilizing Suspicious Forgeries Erasing (SFE) to intentionally occlude the Top-N sensitive regions of face, allowing detector to explore representative forgery from the previously ignored facial region. Specially, our method can be easily integrated with various CNN models without extra structure modification and sophisticated training set.

Through decoupling detector's attention from the oversensitive facial region, our method achieves competitive detection performance with state of the art, and significantly maintains the detection performance on fake faces which only contain few technical forgeries. Moreover, the region of interest visualized by average FAM shows that our method contributes to mining the representative forgery of different manipulation techniques. The main contributions of this work are as follows:

– We propose a tracer method called FAM to precisely locate the facial region to which detector is sensitive, and further use it as the guidance for data augmentation.

– We propose an attention-based data augmentation method called SFE to help detector allocate more attention to representative forgery under the guidance of FAM.

– We finally provide a framework called RFM, which visualizes representative forgery without well-designed supervision and enables a vanilla CNN-based detector to achieve SOTA performance on DFFD and Celeb-DF.

2. Related Work

Face Manipulation Techniques. According to technical procedure, well-known face manipulation techniques [2, 3, 4, 9, 15, 16, 17, 33, 34] can be roughly divided into one-stage and two-stage technique:

The main procedure of **two-stage technique** [2, 4, 21, 33, 34] can be briefly described as: 1) generating target face or extracting face from target individual, 2) blending target identity into source face by utilizing mask, graphics-based technique, etc. In practice, two-stage techniques are widely used for identity swap and expression manipulation. Concretely, Thies *et al*. [34] propose a method to transfer facial expressions from target to source face while maintaining the identity of source person. "Synthesizing Obama" [33] composites synthesized mouth texture with proper 3D pose to help source face match the mouth in target video. *FaceSwap* [4] can swap the face of a person seen by camera with the face in the provided image. *Deepfake* [2] is the symbol of CNN-based face identity swap, which uses an autoencoder to swap the identity of face. Li *et al*. [23] generate a large-scale fake face dataset by using improved synthesis process, solving the low resolution, color mismatch, inaccurate face masks and temporal flickering problems. Li *et al*. [21] propose a two-stage algorithm, achieving high fidelity and occlusion aware face swapping.

Most of **one-stage techniques** [3, 9, 15, 16, 17] are implemented based on GANs, which can achieve entire face synthesis or expression and attributes manipulation without constructing complex physical models. *FaceApp* [3] is a consumer-level mobile application, providing multiple filters to selectively modify facial attributes. Choi *et al*. [9] achieve facial attribute transfer and facial expression synthesis for multiple domains by using only a single model. ProGAN [15] is a popular GAN structure that can synthesize high-resolution facial images by progressively growing both generator and discriminator. StyleGAN [16] achieves the control of synthesis through a new structure, which can automatically separate the high-level attributes and stochastic variation in generated images. Karras *et al*. [17] further improve the perception quality of synthesized images by redesigning the structure of StyleGAN.

Fake Face Detection. Recent studies [6, 8, 11, 13, 22, 24, 25, 36, 38, 40] propose a variety of methods for fake face detection. Dang *et al*. [11] assemble an attention-based layer into detector to locate forgery region and improve detection performance. Huang *et al*. [13] locate forgery region by using a modified semantic segmentation network. Chen *et al*. [8] propose a detector which combines both spatial domain and frequency domain as inputs for detection. Chai *et al*. [6] modify the structure of Xception [10], regarding each receptive field as a patch and detecting them independently. Li *et al*. [22] achieve high generalization detection performance without using fake images generated

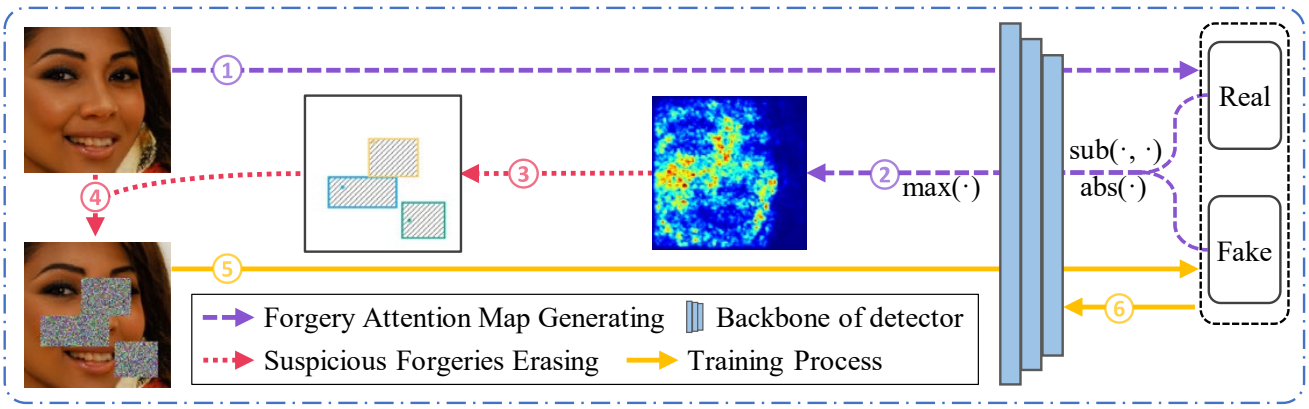

Figure 3. The procedure of RFM, which can be divided into three parts. Firstly (in steps 1, 2), we generate FAM for each original image of a single mini-batch. Then (in steps 3 and 4), we utilize SFE to erase the original images under the guidance of FAMs generated before. Finally (in steps 5 and 6), we use the erased images as inputs to train detector. Specially, in contrast to offline pre-processing, RFM refines training data dynamically during training.

by any existed manipulation methods. Specially, a series of methods [24, 25, 36, 38, 40] are proposed to detect GAN-generated fake face by leveraging the detectable digital fingerprints produced by generators.

Data Augmentation. Data augmentation is a useful approach in addressing the underfitting problem caused by insufficient data and preventing network from overfitting during training process [12, 20, 31, 32, 37, 39, 41]. The most commonly used random cropping [20] and random flip [31] extracts a random patch from the original image and randomly flips the original image, respectively. Dropout [32] can also be regarded as a data augmentation method, which randomly selects some hidden neurons and sets their outputs to zero. Mixup [39] creates new images to expand dataset by calculating the weighted average of two different images. Cutout [12] applies a fixed-size zero-mask to a random location in image, while Random Erasing [41] randomly selects a rectangle region in image and masks the region with random integers. Adversarial Erasing [37] selectively masks image based on the guidance of Class Activation Mapping (CAM) [42].

3. Proposed Method

In this section, we propose an attention-based data augmentation method called *Representative Forgery Mining* (RFM). As shown in Figure 3, our method is composed of two components. 1) *Forgery Attention Map* (FAM) is the foundation of RFM, which can reveal the sensitivity of detector on each facial region. 2) Based on FAM, *Suspicious Forgeries Erasing* (SFE) is applied to augment the original image for detector training. During training, each iteration with RFM only needs to propagate forward and backward twice. In the rest of this section, we explain the main difference between RFM and well-known erasing methods [37, 41]. Same as common settings, we take fake face detection as a binary classification problem.

3.1. Forgery Attention Map

In order to achieve guided erasing and forgeries visualization, FAM is proposed to precisely locate the region to which detector is sensitive. Concretely, the most sensitive region is defined as the region where perturbation has the most critical impact on detection result. In forward propagation, detector receives facial image I as input and outputs two logits O_{real} and O_{fake} to measure whether I is real or not. Because any perturbation would affect both two logits, the detection result should be determined by the relative magnitude of the two logits. By utilizing the $\nabla_I O_{real}$ and $\nabla_I O_{fake}$ to separately represent how perturbation in I impacts on the logits outputs, the maximum absolute difference between $\nabla_I O_{real}$ and $\nabla_I O_{fake}$ is regarded as FAM to simply represent the impact of perturbation on detection result. In other words, each value in FAM precisely indicates the sensitivity of detector to the corresponding pixel in image. Formally, FAM Map can be formulated as

$$\begin{aligned} Map_I &= max\big(abs(\nabla_I O_{fake} - \nabla_I O_{real})\big) \\ &= max\Big(\nabla_I\big(abs(O_{fake} - O_{real})\big)\Big)\,, \end{aligned} \tag{1}$$

where the function $max(\cdot)$ calculates the maximum value along channel axis and the function $abs(\cdot)$ obtains the absolute value of each pixel.

The difference between FAM and well-known Class Activation Mapping [7, 29, 42] can be demonstrated from two aspects. On the one hand, FAM locates the region where detector is sensitive, while Class Activation Mapping highlights the region which detector used for decision-making. On the other hand, FAM generates map at image level, while Class Activation Mapping calculates map based on the last convolutional layer of network.

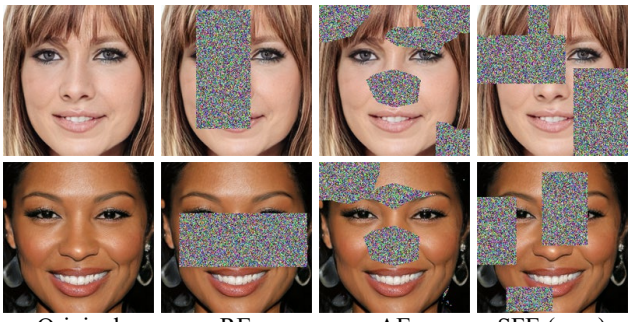

Figure 4. Examples of faces processed by RE, AE, and SFE. Obviously, our SFE is more flexible than other methods.

3.2. Suspicious Forgeries Erasing

Through occluding the Top-N sensitive facial regions calculated by FAM, our proposed erasing method SFE realizes dynamic refinement. In detail, we firstly generate FAM for each image in mini-batch. The sizes of both FAM and input image can be assumed as $H \times W$. Then, for each image, we sort coordinates in descending order according to the values in the corresponding FAM generated before. Next, each pixel is treated as an anchor according to the order calculated above. For each anchor, we use random integers to form a rectangle block whose size is smaller than $H_e \times W_e$ ($H_e \leq H, W_e \leq W$) to occlude the anchor if it has not been occluded before. We repeat the occlusion process until each image has been occluded by N blocks. The detail procedure of SFE is summarized in Algorithm 1.

Algorithm 1: Suspicious Forgeries Erasing

Input: Input facial image I;
Image size H and W;
Forgery Attention map Map;
Erasing Block count N;
Erasing probability p;
Max erase size H_{max} and W_{max}
Output: Erased image I^*.

```
if Rand(0,1) ≤ p then
    cnt = 0;
    while cnt < N do
        [i, j] = coordinate of the ind^th largest value in Map;
        if I[i, j] has not been occluded then
            H_t = Rand(1, H_max);
            W_l = Rand(1, W_max);
            H_b = H_max − H_t;
            W_r = W_max − W_l;
            Fill I[i − H_t : i + H_b, j − W_l : j + W_r]
              with a block composed of random integers;
            cnt = cnt + 1;
        end
    end
end
I* ← I;
return I*;
```

3.3. Comparison with well-known erasing methods.

To demonstrate why SFE is recommended for fake face detection, we take an analysis on the main difference between SFE and well-known erasing methods such as Random Erasing (RE) [41] and Adversarial Erasing (AE) [37].

Random Erasing partially occludes image at a random position with a single random-sized rectangle mask, making network robust to single occlusion. However, in term of fake face detection, the forgeries on which detector focuses may be far away from each other, causing it difficult for RE to erase all the forgeries while retaining as much effective facial information as possible. Meanwhile, due to the lack of effective guidance, RE can not selectively erase the forgeries which would lead to overfitting. Additionally, the inherent algorithm defect of RE makes RE more inclined to erase the central region of image.

Adversarial Erasing is a guided method that can progressively erase the discriminative object region. However, the Class Activation Mapping [42] which AE utilized to locate the erasing region is calculated on the last convolutional layer of network, which may cause the occlusion position to be different from the region that should be occluded. Moreover, the mask generating method which AE used is so fine-grained that may increase the risk of overfitting the shape and location of the mask.

In comparison with these methods, SFE can 1) precisely occlude sensitive facial regions under the guidance of FAM, 2) utilize multiple blocks to flexibly erase forgeries of different locations and preserve as much facial region as possible, 3) never leak extra information to detector and prevent detector from overfitting to the shape or location of erasing block. The example facial images produced by different erasing methods are illustrated in Figure 4.

4. Experiments

4.1. Dataset

We evaluate our method by performing experiments on two well-known datasets: DFFD [11] and Celeb-DF [23].

DFFD [11] contains 58,703 real facial images and 240,336 fake facial images. The manipulation techniques in DFFD are various in category, including face identity swap, face expression and attributes manipulation, and entire face synthesis. Moreover, both one-stage and two-stage manipulation techniques are used to generate fake faces in DFFD.

Specially, due to the inaccessibility of DFL [1], the DFFD we collected does not contain DFL database. According to the number of manipulation technical stages, we divided the fake faces in DFFD into Group A and Group B.

In detail, **Group A** contains fake faces generated by two-stage techniques such as FaceSwap [4], Deepfakes [2] and Face2Face [34], while the manipulation techniques of face images in **Group B** is composed of the one-stage techniques such as FaceAPP[3], StarGAN [9], PGGAN [15] and StyleGAN [16]. It must be pointed out that the images in Group A are collected from **FaceForensics++** [27].

Celeb-DF [23] is a symbol of the second generation manipulation technology, which generates fake face through a improved two-stage technique. Celeb-DF contains 590 real videos collected from YouTube video clips of 59 celebrities and 5,639 high-quality fake videos of celebrities generated using improved synthesis process. The fake faces in Celeb-DF are more difficult to distinguish than the fake faces in the previous datasets of the same category. For fake face detection, we extract facial images from the key frames of videos in Celeb-DF.

4.2. Experiment Settings

We firstly resize the aligned facial images into a fixed size of 256×256. Then, we apply random and center cropping into training and testing process to resize the images to 224×224, respectively. Moreover, we flip each image horizontally with a probability of 50% during training.

We adopt Xception [10] as the backbone of detector. All the detectors are trained by using Adam [19] optimizer with fixed learning rate of 0.0002. Following [11], the size of mini-batch is set to 16, and each mini-batch consists of 8 real and 8 fake facial images. On the basis of cross entropy loss function, we extra utilize the loss term in [14] with $b = 0.04$ to stabilize training. To fairly compare performance of the detectors trained with and without our method, all the detectors are trained from a same weight initialization. The hyper-parameters N, p, H_{max} and W_{max} are implicitly set as 3, 1.0, 120 and 120, respectively.

We report the detection performance by using the evaluation metrics such as Area Under Curve (AUC) of ROC, True Detect Rate (TDR) at False Detect Rate (FDR) of 0.01% (denoted as $\mathrm{TDR}_{0.01\%}$), and TDR at FDR of 0.1% (denoted as $\mathrm{TDR}_{0.1\%}$).

4.3. Ablation Study

In this section, we perform a number of ablation studies to better understand the contribution of each component and hyper-parameter in RFM.

Effect of Forgery Attention Map and Multiple Erasing Blocks. We conduct experiments on DFFD to investigate how Forgery Attention Map (FAM) and Multiple Erasing Blocks (MEB) boost detection performance. The functions of FAM and MEB are independent in RFM, where FAM plays the role of guidance and MEB emphasizes erasing with multiple blocks. The results are shown in Table 1, where "FAM&MEB" is the original setting, "w/ MEB" denotes placing the anchors of SFE randomly, "w/ FAM" denotes only occluding the Top-1 sensitive region under the guidance of FAM, and "w/o MEB|FAM" denotes using a single erasing block to occlude a random region of face.

Method	AUC	$\mathrm{TDR}_{0.1\%}$	$\mathrm{TDR}_{0.01\%}$
Xception	99.94	94.47	87.17
+Ours, *w/o MEB|FAM*	99.95	97.21	92.62
+Ours, *w/ MEB*	99.95	97.40	93.13
+Ours, *w/ FAM*	99.96	98.06	94.83
+Ours, *w/ FAM&MEB*	**99.97**	**98.35**	**95.50**

Table 1. Ablation for the effect of different settings in RFM on DFFD. **MEB**: Multiple Erasing Blocks, **FAM**: Forgery Attention Map.

Iter.	Method	AUC	$\mathrm{TDR}_{0.1\%}$	$\mathrm{TDR}_{0.01\%}$
250k	+Ours, *w/ PSFE*	99.94	**96.91**	**92.84**
	+Ours, *w/ SFE*	**99.95**	96.17	92.78
350k	+Ours, *w/ PSFE*	99.97	98.25	95.34
	+Ours, *w/ SFE*	99.97	**98.35**	**95.50**

Table 2. Comparison of SFE with PSFE on DFFD. We separately compared their performance under 250k and 350k training iterations.

Although the ordinary erasing ("w/o MEB|FAM") can improve detection performance, we observe that MEB and FAM further improve $\mathrm{TDR}_{0.01\%}$ by 0.51% and 2.21% on DFFD, respectively, demonstrating that either MEB or FAM has contribution to our algorithm and FAM is more effective than MEB. Compared with baseline detector, the detector trained with combining both MEB and FAM leads to significant improvements of 8.33% $\mathrm{TDR}_{0.01\%}$ and 3.88% $\mathrm{TDR}_{0.1\%}$ on DFFD, which also outperforms all the other settings by a large margin.

Comparison with Progressive Suspicious Forgeries Erasing. We also try to design an iterative erasing procedure to replace SFE. The new method progressively occludes the most sensitive region of face rather than using multiple erasing blocks to occlude face at a time. Concretely, we erase the Top-1 sensitive region of face under the guidance of FAM, and then re-generate FAM of the erased face for next erasing. The two steps work in an alternative manner until the facial image has been erased N times. We call this method *Progressive Suspicious Forgeries Erasing* (PSFE). To evaluate the effectiveness of PSFE, we separately compare the performance of detectors trained under different iterations. Although PSFE consumes three times the time of SFE in processing, the results in Table 2 show that PSFE does not achieve significant performance gains.

The impact of hyper-parameters. To investigate the optimal hyper-parameters under different compositions of training data, we extract two sub-datasets from the training

Method	Celeb-DF			DFFD (Group A)			DFFD (Group B)		
	AUC	$TDR_{0.1\%}$	$TDR_{0.01\%}$	AUC	$TDR_{0.1\%}$	$TDR_{0.01\%}$	AUC	$TDR_{0.1\%}$	$TDR_{0.01\%}$
Xception	99.85	89.11	84.22	99.94	97.67	94.57	99.92	92.87	83.46
+AE [37]	99.84	84.05	76.63	99.94	97.98	93.64	99.92	92.97	81.73
+RE [41]	99.89	88.11	85.20	99.95	98.35	95.08	99.96	96.53	91.89
+Ours (RFM)	**99.94**	**93.88**	**87.08**	**99.97**	**99.53**	**98.91**	**99.96**	**97.76**	**93.80**
Patch [6]	99.96	91.83	86.16	99.94	99.85	99.23	99.96	**99.58**	98.75
+Ours (RFM)	**99.97**	**93.44**	**89.58**	**99.95**	**99.87**	**99.68**	**99.97**	99.56	**98.87**

Table 3. Comparison of RFM with well-known erasing methods and state of the art on DFFD and Celeb-DF. We use Xception as backbone in the first cell and use Patch as backbone in the second cell.

Method	$Size$	p	AUC	$TDR_{0.1\%}$	$TDR_{0.01\%}$
Xception	-	-	96.87	40.19	25.14
+Ours	30	0.5	**97.89**	**48.64**	**35.89**
+Ours	30	1.0	97.32	47.23	29.38
+Ours	120	0.5	96.79	38.04	29.23
+Ours	120	1.0	96.41	34.96	26.48
Xception	-	-	96.71	41.94	34.07
+Ours	30	0.5	96.99	44.02	34.42
+Ours	30	1.0	96.93	43.57	33.75
+Ours	120	0.5	**97.95**	**50.39**	40.06
+Ours	120	1.0	97.87	50.17	**40.86**
Xception	-	-	99.36	70.34	59.10
+Ours	30	1.0	99.38	70.95	59.40
+Ours	120	1.0	**99.53**	**76.26**	**62.12**

Table 4. Detection performance under different hyper-parameters on Celeb-DF. Results in the three cells are from the detectors trained on **Celeb-DF-subsetA**, **Celeb-DF-subsetB** and **Celeb-DF-subsetA&B**, respectively. p: Erasing probability, $Size$: Max erase size $H_{max}\&W_{max}$.

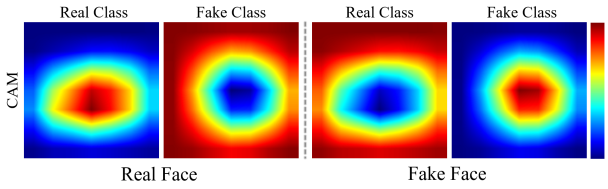

Figure 5. Average Class Activation Mapping (CAM) [42] on 512 facial images in Celeb-DF. The first and second colomns represent the average CAM on real and fake faces, respectively.

set in Celeb-DF. The fake faces in Celeb-DF can be represented as replacing the source face with target identity. By limiting the count of source faces and target identities in training set to 8, we can construct two sub-datasets **Celeb-DF-subsetA** and **Celeb-DF-subsetB**, respectively. Basing on these datasets, we separately train a series of detectors with different settings on Max erase size $H_{max}\&W_{max}$ and Erasing probability p. All the detectors are trained about 120k iterations, and the results are shown in Table 4. It is obvious that both the composition of training set and the setting of hyper-parameters have significant impact on

Method	AUC	$TDR_{0.1\%}$	$TDR_{0.01\%}$
Xception [11]	99.61	85.26	77.42
+Reg.[11]	99.64	90.78	83.83
Xception	99.87	85.55	77.92
+Ours (RFM)	**99.94**	**96.94**	**90.44**

Table 5. Comparison of RFM with state of the art on DFFD. Results shown in the first cell are from [11], and results in the second cell represent the detectors trained with the same iterations as [11] on DFFD without DFL.

RFM, and $H_{max}\&W_{max}$ is more decisive on detection performance than p. Moreover, large $H_{max}\&W_{max}$ and p help detector improve its detection performance when training on Celeb-DF-subsetA&B or Celeb-DF-subsetB, while the large hyper-parameters lead to performance degradation when training on Celeb-DF-subsetA. Unlike training with larger parameters, training with smaller parameters helps to improve performance in most cases, which means that the parameters can be set from small to large until the detector reaches the optimal performance. Anyway, in order to achieve the optimal effect, it is essential to set proper parameters during training.

4.4. Experiment on DFFD and Celeb-DF

Comparison with well-known erasing methods. We separately conduct experiments on DFFD and Celeb-DF to compare RFM with well-known erasing methods such as Adversarial Erasing (AE) [37] and Random Erasing (RE) [41]. Random integers are used to compose erasing blocks for all three methods. And the hyper-parameters in both AE and RE are set as the original settings.

The results in Table 3 shows that RFM outperforms baseline and other erasing methods by a large margin. Meanwhile, it is counter-intuitive that the usage of AE leads to performance degradation. To further figure out why AE does not work on fake face detection, we generate the average CAM on 512 real and fake faces separately. As shown in Figure 5, we find that the CAM which AE used only con-

Part	Method	Celeb-DF			DFFD (Group A)			DFFD (Group B)		
		AUC	$TDR_{0.1\%}$	$TDR_{0.01\%}$	AUC	$TDR_{0.1\%}$	$TDR_{0.01\%}$	AUC	$TDR_{0.1\%}$	$TDR_{0.01\%}$
Eyes Real	Xception	69.99	00.83	00.39	93.41	33.62	20.04	97.58	44.88	22.79
	+Ours (RFM)	**93.43**	**20.47**	**13.86**	**99.72**	**78.19**	**52.63**	**99.30**	**61.12**	**29.67**
Nose Real	Xception	82.77	02.92	01.33	95.50	59.99	45.59	98.51	76.01	59.50
	+Ours (RFM)	**98.88**	**48.33**	**35.19**	**99.95**	**98.20**	**95.90**	**99.92**	**93.45**	**85.58**
Mouth Real	Xception	98.30	46.18	32.36	98.76	67.93	49.94	99.25	66.61	48.44
	+Ours (RFM)	**98.50**	**46.30**	**34.83**	**99.95**	**97.77**	**94.39**	**99.84**	**86.37**	**75.73**
Skin Real	Xception	85.33	08.20	04.53	92.73	16.81	6.779	94.95	47.38	29.24
	+Ours (RFM)	**97.87**	**45.44**	**34.36**	**93.79**	**35.80**	**28.97**	**96.80**	**61.06**	**48.52**
Eyes Real	Patch	97.06	02.25	01.28	99.91	87.86	12.70	99.87	60.64	06.92
	+Ours (RFM)	**97.82**	**06.56**	**04.69**	**99.93**	**97.21**	**20.56**	**99.90**	**81.89**	**28.40**
Nose Real	Patch	98.76	21.75	13.98	99.97	99.64	89.83	99.96	97.64	68.10
	+Ours (RFM)	**99.52**	**56.17**	**47.21**	**99.97**	**99.84**	**97.91**	**99.97**	**97.68**	**84.17**
Mouth Real	Patch	**99.32**	12.68	06.46	99.96	99.81	87.86	99.96	99.16	65.08
	+Ours (RFM)	**99.32**	**29.84**	**23.03**	**99.97**	**99.86**	**95.28**	**99.97**	**99.22**	**83.65**
Skin Real	Patch	98.98	29.85	15.26	98.88	59.18	25.06	**99.87**	80.81	20.01
	+Ours (RFM)	**99.55**	**47.30**	**32.70**	**99.66**	**77.52**	**39.47**	99.73	**80.89**	**39.77**

Table 6. Comparison of RFM with baseline methods on the facial images which only have few technical forgeries.

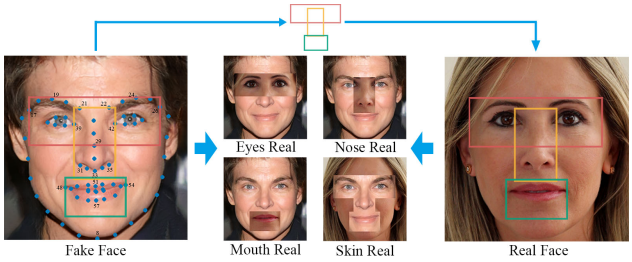

Figure 6. The generation procedure of less-forgery fake faces. The image with suffix "Real" means that the region in fake face have been replaced with the corresponding pixels from a random real face.

tains high-dimensional information and loses the representation ability for forgery region, providing insufficient guidance for erasing.

Comparison with state of the art. In order to further evaluate the effectiveness of RFM, we make comparisons with state of the art on DFFD and Celeb-DF separately. As shown in Table 3, a vanilla Xception with RFM can achieve competitive performance with Patch [6]. Moreover, by using RFM generated images for training, Patch achieves virtually the best performance than any other methods. Obviously, it is also a strong proof that RFM can be easily integrated with various models to improve fake face detection performance.

Additionally, we also compare RFM with [11]. Since the DFFD we collected lacks fake faces from Deep Face Lab (DFL) [1], we implement the same model selection strategy as that in [11] to conduct a roughly fair comparison. The whole process can be divided into two steps: Firstly, following the training iteration in [11], we trained an Xception-based detector on the incomplete DFFD until the detector achieves the same $TDR_{0.1\%}$ and $TDR_{0.01\%}$ as [11]. Then, we utilize RFM to train another detector under the same iteration. The results in Table 5 indicate that RFM yields improvements of 6.16% $TDR_{0.1\%}$ and 6.61% $TDR_{0.01\%}$ when compared with results in [11].

Robustness on less-forgery fake faces. To explore how RFM affects the detection performance on the fake faces with few technical forgeries, we propose to leverage semantic-based segmentation to generate less-forgery fake faces for testing. Concretely, we firstly obtain the locations of 68 facial landmarks by utilizing the facial landmarks extractor in dlib [18]. Then, the facial landmarks are used to divide fake face into four parts: eyes, nose, mouth and facial skin. Next, less-forgery fake faces are generated by separately replacing each region in fake face with the corresponding pixels of a real face. Finally, we construct four datasets that contain the fake faces with few technical forgeries on eyes, nose, mouth, and facial skin region respectively. The detailed process is shown in Figure 6.

We separately evaluate the detectors trained with or without RFM on these datasets. The results in Table 6 demonstrate that the detectors without RFM encounter severe performance degradation when facing less-forgery faces, while REF effectively helps detector maintain performance on less-forgery faces.

4.5. Representative Forgery Visualization

Visualization on several images. In order to exhibit the representative forgery region discovered by RFM, we generate average FAM on 512 fake facial images for each face manipulation technique. As illustarted in Figure 7, the de-

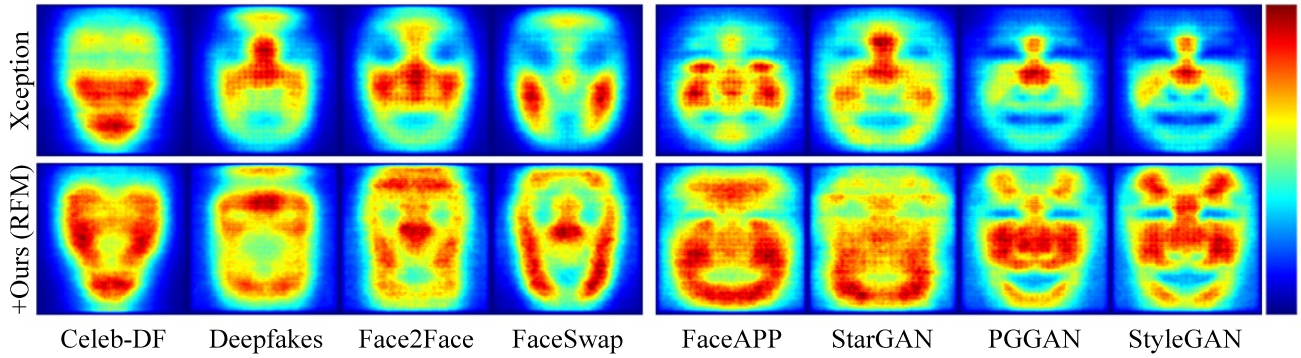

Figure 7. Average FAMs separately generated under the detectors trained with different methods. The manipulation techniques in the left and right colomn are consist of two-stage and one-stage techniques respectively.

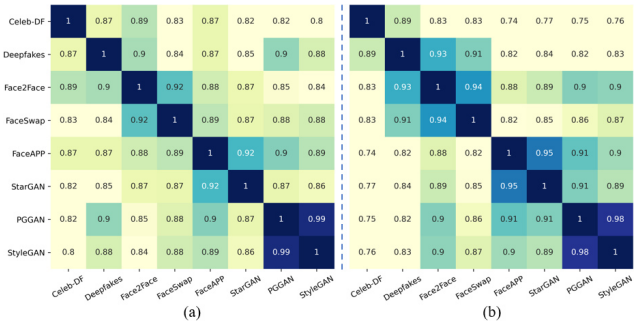

Figure 8. The correlation matrixes calculate the normalized cosine similarity of average FAMs between each pair of manipulation techniques. The average FAMs are generated under the detectors trained (a) with or (b) without RFM respectively.

tector trained with RFM pays attention to a more comprehensive and representative region of interest, where the facial boundary of faces generated by two-stage techniques and the entire skin of faces generated by one-stage techniques.

Furthermore, under the help of RFM, the average FAMs of fake faces generated by similar techniques tend to be similar to each other (as the correlation matrixes shown in Figure 8), which reflects that RFM produces a clustering effect on the fake faces of similar techniques. Therefore, our method can be further utilized to explore the technical procedure of a black-box manipulation technique. To achieve this goal, we firstly train a detector with RFM on the fake faces generated by both known and unknown techniques. Then, we generate the average FAM for each technique and calculate the normalized cosine similarity of FAMs between each pair of known and unknown techniques. After that, the technical procedure of a black-box technique can be determined according to its correlation with other known methods. As shown in the matrix, it can be inferred that FaceAPP is belonging to one-stage manipulation technique.

Visualization on a single video. Actually, fake faces generated by two-stage techniques mainly appear in the form of video. To further investigate how the number of

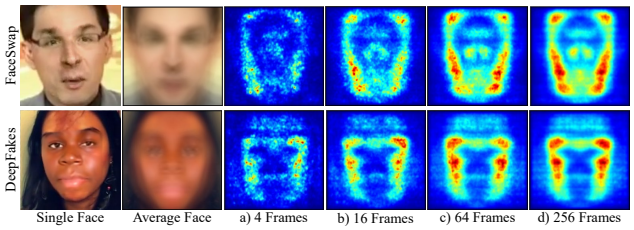

Figure 9. Comparison of average FAMs calculated on a single fake video with different frames. 'Single face' shows the fake face of single frame and 'Average face' shows the average fake face of multiple frames. The columns a) to d) shows the average FAMs calculated on 4 to 256 frames, respectively.

video frames of single fake video influences on the visualization effectiveness of our method, we exhibit the average FAMs generated based on different frames of a single video. As shown in Figure 9, average FAM generated on 4 frames is enough to show the representative forgery with discriminative contour, and the complete contour and inner of representative forgery appear gradually when the number of frames raises from 4 to 256. In conclusion, RFM performs well on representative forgery mining.

5. Conclusion

In this work, we provide insight into fake face detection that detection performance can be effectively improved by refining training data. Concretely, we propose a novel attention-based data augmentation method to guide detector explore representative forgery from the previously ignored facial region. In addition, the visualization result shows that our method can separately discover the corresponding representative forgery of different manipulation techniques without the need of well-designed supervision. With our method, a vanilla CNN-based detector can achieve state-of-the-art performance on the well-known fake face datasets DFFD and Celeb-DF.

Acknowledgment. This work was supported by National Key R&D Program of China (2019YFB1406504).

References

[1] Deepfacelab. https://github.com/iperov/DeepFaceLab. Accessed: 2019-09-04.

[2] Deepfakes github. https://github.com/deepfakes/faceswap. Accessed: 2018-10-29.

[3] Faceapp. https://faceapp.com/app. Accessed: 2019-09-04.

[4] Faceswap. https://github.com/MarekKowalski/FaceSwap. Accessed: 2018-10-29.

[5] This person does not exist. https://thispersondoesnotexist.com. Accessed: 2019-2-15.

[6] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, 2020.

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[8] Zehao Chen and Hua Yang. Manipulated face detector: Joint spatial and frequency domain attention network. *arXiv preprint arXiv:2005.02958*, 2020.

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

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

[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[13] Yihao Huang, Felix Juefei-Xu, Run Wang, Xiaofei Xie, Lei Ma, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. Fakelocator: Robust localization of gan-based face manipulations via semantic segmentation networks with bells and whistles. *arXiv preprint arXiv:2001.09598*, 2020.

[14] Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Do we need zero training loss after achieving zero training error? In *International Conference on Machine Learning*, pages 4604–4614. PMLR, 2020.

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[18] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[21] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020.

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

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

[24] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8060–8069, 2020.

[25] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019(5):532–1, 2019.

[26] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018.

[27] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

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

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via

gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[30] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[33] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

[34] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[35] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020.

[36] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 7, 2020.

[37] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[38] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7556–7566, 2019.

[39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

[40] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.

[41] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020.

[42] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.