Face X-ray for More General Face Forgery Detection

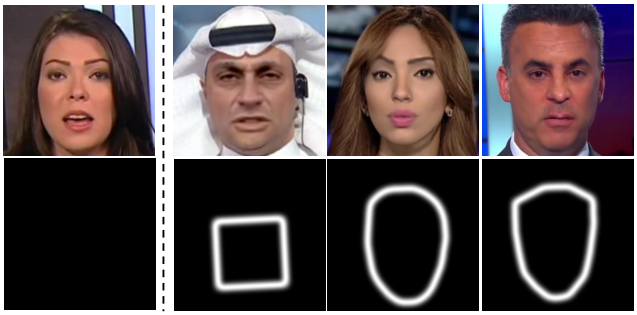
Lingzhi Li[1*†] Jianmin Bao[2*] Ting Zhang[2] Hao Yang[2] Dong Chen[2] Fang Wen[2] Baining Guo[2]
[1]Peking University [2]Microsoft Research Asia
lilingzhi@pku.edu.cn {jianbao,Ting.Zhang,haya,doch,fangwen,bainguo}@microsoft.com

Abstract

In this paper we propose a novel image representation called face X-ray for detecting forgery in face images. The face X-ray of an input face image is a greyscale image that reveals whether the input image can be decomposed into the blending of two images from different sources. It does so by showing the blending boundary for a forged image and the absence of blending for a real image. We observe that most existing face manipulation methods share a common step: blending the altered face into an existing background image. For this reason, face X-ray provides an effective way for detecting forgery generated by most existing face manipulation algorithms. Face X-ray is general in the sense that it only assumes the existence of a blending step and does not rely on any knowledge of the artifacts associated with a specific face manipulation technique. Indeed, the algorithm for computing face X-ray can be trained without fake images generated by any of the state-of-the-art face manipulation methods. Extensive experiments show that face X-ray remains effective when applied to forgery generated by unseen face manipulation techniques, while most existing face forgery detection algorithms experience a significant performance drop.

1. Introduction

Recent studies have shown rapid progress in facial manipulation, which enables an attacker to manipulate the facial area of an image and generate a new image, *e.g.*, changing the identities or modifying the face attributes. With the remarkable success in synthesizing realistic faces, it becomes infeasible even for humans to distinguish whether an image has been manipulated. At the same time, these forged images may be abused for malicious purpose, causing severe trust issues and security concerns in our society. Therefore, it is of paramount importance to develop effective methods for detecting facial forgery.



(a) (b)

Figure 1. Face X-ray reveals the blending boundaries in forged face images and returns a blank image for real images. (a) a real image and its face X-ray, (b) fake images and their face X-rays.

Our focus in this work is the problem of detecting face forgeries, such as those produced by current state-of-the-art face manipulation algorithms including DeepFakes [1], Face2Face [46], FaceSwap [2], and NeuralTextures [45]. Face forgery detection is a challenging problem because in real-world scenarios, we often need to detect forgery without knowing the underlying face manipulation methods. Most existing works [12, 44, 25, 35, 36] detect face manipulation in a supervised fashion and their methods are trained for known face manipulation techniques. For such face manipulation, these detection methods work quite well and reach around 98% detection accuracy. However, these detection methods tend to suffer from overfitting and thus their effectiveness is limited to the manipulation methods they are specifically trained for. When applied to forgery generated by unseen face manipulation methods, these detection methods experience a significant performance drop.

Some recent works [49, 13] have noticed this problem and attempted to capture more intrinsic forgery evidence to improve the generalizability. However, their proposed methods still rely on the generated face forgeries for supervision, resulting in limited generalization capability.

In this paper we propose a novel image representation, *face X-ray*, for detecting fake face images. The key observation behind face X-ray is that most existing face manipulation methods share the common step of blending an altered face into an existing background image, and there ex-

*Equal contribution
†Work done during an internship at Microsoft Research Asia

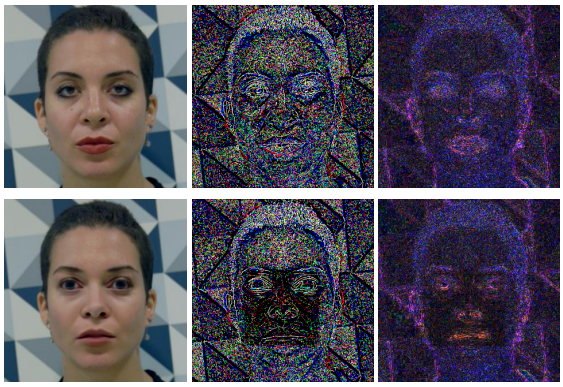


Figure 2. Noise analysis (middle column) and error level analysis (right column) of (a) a real image and (b) a fake image.

ist intrinsic image discrepancies across the blending boundaries. These discrepancies make the boundaries fundamentally detectable. Indeed, due to the acquisition process, each image has its own distinctive marks introduced either from hardware (*e.g.*, sensor, lens) or software components (*e.g.*, compression, synthesis algorithm) and those marks tend to present similarly throughout the image [39]. We illustrate noise analysis[1] and error level analysis [21] as two representative types of distinctive marks in Figure 2.

Face X-ray capitalizes on the above key observation and provides an effective way for detecting forgery produced by most existing face manipulation algorithms. For an input face image, its face X-ray is a greyscale image that can be reliably computed from the input. This greyscale image not only determines whether a face image is forged or real, but also identifies the location of the blending boundary when it exists, as shown in Figure 1.

Face X-ray is a significant step forward in the direction of developing a general face forgery detector

as it only assumes the existence of a blending step and does not rely on any knowledge of the artifacts associated with a specific face manipulation algorithm. This level of generality covers most existing face manipulation algorithms. Moreover, the algorithm for computing face X-ray can be trained by self-supervised learning with a large number of blended images composited from real ones, without fake images generated by any of the state-of-the-art face manipulation methods. As a result, face X-ray remains effective when applied to forgery generated by an unseen face manipulation method, while most existing face forgery detection algorithms experience a significant performance drop.

Our experiments demonstrate that face X-ray significantly improves the generalization ability through a thorough analysis. We show that our framework achieves a remarkably high detection accuracy on unseen face forgeries, as well as the ability to predict face X-rays reliably and faithfully on all kinds of recent popular face manipulations. In comparison with other face forgery detectors, our framework largely exceeds the competitive state-of-the-arts.

[1] https://29a.ch/photo-forensics/#noise-analysis

2. Related Work

Over the past several years, forgery creation, of particular interest in faces given its wide applications, has recently gained significant attention.

With the complementary property of forgery creation and detection, face forgery detection also becomes an increasingly emerging research area. In this section, we briefly review prior image forensic methods including face forensics to which our method belongs.

Image forgery classification. Image forgery detection is mostly regarded as merely a binary (real or forgery) classification problem. Early attempts [30, 16, 17] aim to detect forgeries, such as copy-move, removal and splicing that once were the most common manipulations, by utilizing intrinsic statistics (*e.g.*, frequency domain characteristics)

of images. However, it is difficult to handcraft the most suitable and meaningful features. With the tremendous success of deep learning, some works [10, 33, 7] adopt neural networks to automatically extract discriminative features for forgery detection.

Recent advanced manipulation techniques, especially about faces, are capable of manipulating the images in a way that leaves almost no visual clues and can easily elude above image tampering detection methods. This makes face forgery detection increasingly challenging, attracting a large number of research efforts [12, 44, 25, 32, 28, 18, 3, 22, 35, 36]. For instance, a face forensic approach exploiting facial expressions and head movements customized for specific individuals is proposed in [4]. FakeSpotter [47] uses layer-wise neuron behavior instead of only the last neuron output to train a binary classifier. To handle new generated images, an incremental learning strategy is introduced in [27]. Lately, FaceForensics++ [36] provides an extensive evaluation of forgery detectors in various scenarios.

Image forgery localization. Besides classification, there are methods focusing on localizing the manipulated region. Early works [37, 8, 15] reveal the tampered regions using manually designed low-level image statistics at a local level. Subsequently, deep neural network is introduced in image forgery localization, where most works [5, 40, 29, 38] use multi-task learning to simultaneously detect the manipulated images and locate the manipulated region. Instead of simply using a multi-task learning strategy, Stehouwer et al. [41] highlight the informative regions through an attention mechanism where the attention map is guided by the groundtruth manipulation mask. Bappy et al. [6] present a localization architecture that exploits both frequency domain and spatial context. However, early works are not well suited for detecting advanced manipulations, while deep

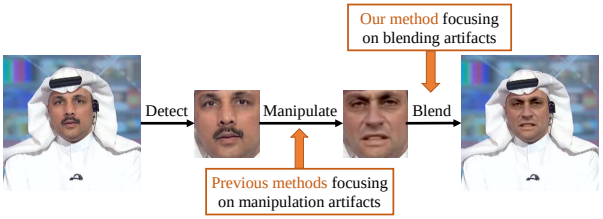


Figure 3. Overview of a typical face manipulation pipeline. Previous works detect artifacts produced by manipulation methods, while our approach focuses on detecting face X-ray.

learning based methods adopt supervised training, requiring a huge amount of corresponding groundtruth manipulation masks, which may be inaccessible in practice.

Generalization ability of Image forgery detection. With the evolution of new technologies, it has been noted in recent works [19, 11, 49, 13, 23] that the performance of current methods drop drastically on forgeries of new types. In particular, Xuan et al. [49] use an image preprocessing step to destroy low level unstable artifact, forcing the network to focus on more intrinsic forensic clues. ForensicTransfer [11] proposes an autorencoder-based neural network to transfer knowledge between different but related manipulations. LAE [13] also uses autorencoder to learn fine-grained representation regularized by forgery mask supervision. All above methods still need forged images to train a supervised binary classifier, resulting in limited generalization capability. Another related work is FWA [23], which targets the artifacts in affine face warping in a self-supervised way. However, FWA focuses on detecting DeepFake generated images such that the detection model is not applicable for other types of manipulations, *e.g.*, Face2Face.

3. Face X-Ray

We start by introducing the key observation behind face X-ray. Then we formally define the face X-ray of a given input image. Finally we provide details on obtaining labeled data (a set of pairs consisting of an image and its corresponding face X-ray) from real images to train our framework in a self-supervised manner.

As shown in Figure 3, a typical facial manipulation method consists of three stages: 1) detecting the face area; 2) synthesizing a desired target face; 3) blending the target face into the original image.

Existing face forgery detection methods usually focus on the second stage and train a supervised per-frame binary classifier based on datasets including both synthesized videos generated from manipulation methods and real ones. Although near perfect detection accuracy is achieved on the test dataset, we observe significantly degraded performance when applying the trained model to unseen fake images, which is empirically verified in Section 5.1.

We take a fundamentally different approach. Instead of

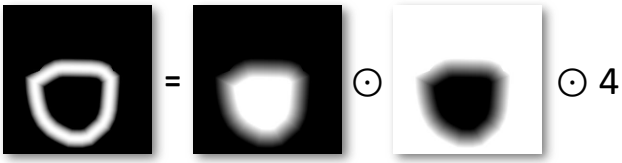


Figure 4. Illustrating the relationship between face X-ray and the mask. $\odot$ represents the element-wise multiplication.

capturing the synthesized artifacts of specific manipulations in the second stage, we try to locate the blending boundary that is universally introduced in the third stage of the face manipulation pipeline. Our approach is based on a key observation: *when an image is formed by blending two images, there exist intrinsic image discrepancies across the blending boundary.*

It is noted in the literature [14] that each image has its own distinctive marks or underlying statistics, which mainly come from two aspects: 1) hardware, *e.g.* color filter array (CFA) interpolation introducing periodic patterns, camera response function that should be similar for each of the color channels, sensor noise including a series of on-chip processings such as quantization and white balancing, introducing a distinct signature; 2) software, *e.g.*, lossy compression schemes that introduce consistent blocking artifacts, GAN based synthesis algorithms that may leave unique imprints [26, 50]. All above factors contribute to the image formation, leaving specific signatures that tend to be periodic or homogeneous, which may be disturbed in an altered image. As a result, we can detect a forged face image by discovering the blending boundary using the inconsistencies of the underlying image statistics across the boundary.

3.1. Face X-Ray Definition

Given an input face image I, we wish to decide whether the image is a manipulated image I_M that is obtained by combining two images I_F and I_B

$$I_M = M \odot I_F + (1 - M) \odot I_B, \tag{1}$$

where $\odot$ specifies the element-wise multiplication. I_F is the foreground manipulated face with desired facial attributes, whereas I_B is the image that provides the background. M is the mask delimiting the manipulated region, with each pixel of M having greyscale value between 0.0 and 1.0. When all the entries are restricted to 0 and 1, we have a binary mask, such as the mask used in Poisson blending [31]. Note that color correction techniques (*e.g.* color transfer [34]) are usually applied over the foreground image I_F before blending so that its color matches the background image color.

We would like to define the face X-ray as an image B such that if the input is a manipulated image, B will reveal the blending boundary and if the input is a real image, then B will have zero for all its pixels.

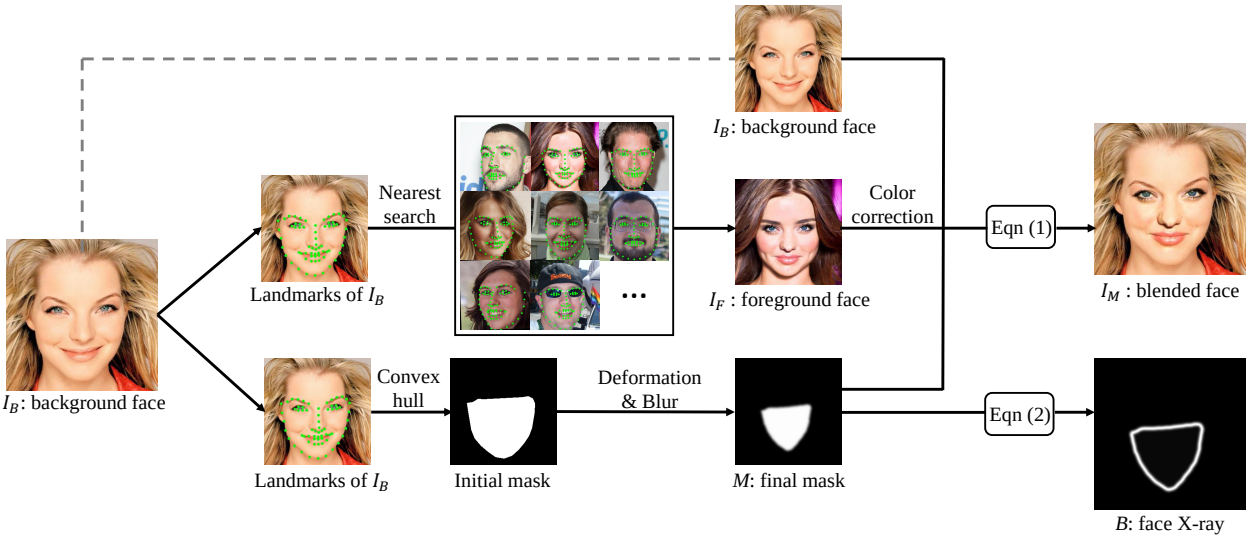


Figure 5. Overview of generating a training sample. Given a real face I_B, we seek another real face I_F to represent the manipulated variant of I_B and produce a mask to delimit the manipulated region. Then the blended face and its corresponding face X-ray can be obtained through Equation (1) and Equation (2). .

Formally, for an input image I, we define its face X-ray as an image B with

$$B_{i,j} = 4 \cdot M_{i,j} \cdot (1 - M_{i,j}), \tag{2}$$

where the subscript (i, j) is the index denoting the pixel location and M is the mask that is determined by the input image I. If the input image is real, then the mask M is a trivial blank image with either all 0-pixels or all 1-pixels. Otherwise, the mask M will be a nontrivial image delimiting the foreground image region. Note that the maximum value of $M_{i,j} \cdot (1 - M_{i,j})$ is no greater than 0.25 and in fact only achieves the maximum value of 0.25 when $M_{i,j} = 0.5$. For this reason, the face X-ray pixel $B_{i,j}$ is always valued between 0 and 1. Figure 4 illustrates the relationship between the mask M and the face X-ray B with a toy example.

In our face X-ray definition, we always assume the mask M is soft and never use a binary mask. A binary mask M would pose a problem for our face X-ray definition because the corresponding face X-ray would be a blank image with all pixels valued as 0 even when M is not a trivial mask with all 0-pixels or all 1-pixels. This would defeat the purpose of detecting blending boundary in manipulated images. For this reason, we always adopt a 3×3 Gaussian kernel to turn a binary mask M into a soft mask before using it in Equation (2).

In essence, the face X-ray aims to discover a soft mask M with which the input image I can be decomposed into the blending of two images from different sources according to Equation (1). As mentioned earlier, images from different sources have undeniable differences that, despite their subtlety and invisibility to human eyes, are intrinsic due to the image acquisition process.

Face X-ray is a computational representation for discovering such differences in an input face image of unknown origin.

3.2. Self-Supervised Learning

Now that we have defined the concept of face X-ray, we shall explain one important thing in the rest of this section: how to get training data using only real images.

As is mentioned before, all real images naturally have their corresponding face X-rays with all 0-pixels. Yet those trivial face X-rays are not sufficient to guide the network learning, training data associated with nontrivial face X-rays is certainly crucial and indispensable. One intuitive solution is to access the manipulated images and accordingly the masks generated by facial manipulation methods. Nonetheless, we find that as face X-ray essentially cares about only the blending boundary, it is entirely possible to create nontrivial face X-rays by blending two real images.

To be specific, we describe the generation of nontrivial face X-rays in three stages. 1) First, given a real image I_B, we seek another real image I_F to take the place of the manipulated variant of I_B. We use the face landmarks (extracted by [9]) as the matching criteria to search from a random subset of the rest training videos according to the Euclidean distance between landmarks. To increase the randomness, we take 100 nearest neighbors and randomly choose one as the foreground image I_F. 2) In the second stage, we generate a mask to delimit the manipulated region. The initial mask is defined as the convex hull of the face landmarks in I_B. As face manipulation methods are not necessarily and always focused on the same area of the face, there exist various different shapes of manipulated regions in forged images, *e.g.*, some may be manipulated only

around the mouth region. In order to cover as many shapes of masks as possible, we first adopt random shape deformation using the 2-D piecewise affine transform estimated from the source 16 points (selected from a 4×4 grid) to the target 16 points (deformed from source using random offset), and then apply Gaussian blur with random kernel size, resulting in the final mask. 3) At last, the blended image is obtained through Equation (1), given the foreground image I_F, the background image I_B and the mask, and the blending boundary is attained by Equation (2) using the mask. Note that we apply the color correction technique (aligning the mean of the RGB channels respectively) to I_F, similar to existing facial manipulation methods, so as to match the color of I_B. A brief overview of generating a training sample is illustrated in Figure 5. In practice, we generate the labeled data dynamically along with the training process and train our framework in a self-supervised way.

4. Face Forgery Detection Using Face X-Ray

As described above, we are able to produce a huge number of training data by exploring only real images. Let the generated training dataset be $\mathcal{D} = \{I, B, c\}$ where I represents the image, B denotes the corresponding face X-ray and c is a binary scalar specifying whether the image I is real or blended. we adopt a convolutional neural network based framework due to the extremely powerful representation learning of deep learning. The proposed framework outputs the face X-ray given an input image I and then based on the predicted face X-ray, outputs the probabilities of the input image being real or blended.

Formally, let $\hat{B} = NN_b(I)$ be the predicted face X-ray where NN_b is a fully convolutional neural network, and $\hat{c} = NN_c(\hat{B})$ is the predicted probability with NN_c composed of a global average pooling layer, a fully connected layer, and a softmax activation layer in a sequential manner. During training, we adopt the widely used loss functions for the two predictions. For face X-ray, we use the cross entropy loss to measure the accuracy of the prediction,

$$L_b = -\sum_{\{I,B\}\in\mathcal{D}} \frac{1}{N}\sum_{i,j}(B_{i,j}\log\hat{B}_{i,j} + (1-B_{i,j})\log(1-\hat{B}_{i,j})), \quad (3)$$

where N is the total number of pixels in the feature map $\hat{B}$. For the classification, the loss is,

$$L_c = -\sum_{\{I,c\}\in\mathcal{D}} (c\log(\hat{c}) + (1-c)\log(1-\hat{c})). \quad (4)$$

Therefore, the overall loss function is $L = \lambda L_b + L_c$, where λ is the loss weight balancing L_b and L_c. In the experiments, we set $\lambda = 100$ to force the network focusing more on learning the face X-ray prediction. We train our framework in an end-to-end manner using the back propagation. More implementation details can be found in Section 5.

5. Experiments

In this section, we first introduce the overall experiment setups and then present extensive experimental results to demonstrate the superiority of our approach.

Training datasets. In our experiments, we adopt recent released benchmark dataset FaceForensics++ [36] (FF++) for training. It is a large scale video dataset consisting of 1000 original videos that have been manipulated with four state-of-the-art face manipulation methods: DeepFakes (DF) [1], Face2Face (F2F) [46], FaceSwap (FS) [2], and NeuralTextures (NT) [45]. Another training dataset is the set of blended images that we constructed from real images. We denote such dataset with BI, meaning the blended data samples composited using real images in FF++.

Test datasets. To evaluate the generalization ability of the proposed model using face X-ray, we use the following datasets: 1) FaceForensics++ [36] (FF++) that contains four types of facial manipulations as described above; 2) DeepfakeDetection[2] (DFD) including thousands of visual deepfake videos released by Google in order to support developing deepfake detection methods; 3) Deepfake Detection Challenge[3] (DFDC) released an initial dataset of deepfakes accompanied with labels describing whether they are generated using facial manipulation methods; 4) Celeb-DF [24], a new DeepFake dataset including 408 real videos and 795 synthesized video with reduced visual artifacts.

Implementation detail. For the fully convolutional neural network NN_b in our framework, we adopt the recent advanced neural network architecture, *i.e.*, HRNet [42, 43], and then concatenate representations from all four different resolutions to the same size 64×64, followed by a 1×1 convolutional layer with one output channel, a bilinear upsampling layer with 256×256 output size and a sigmoid function. In the training process, the batch size is set to 32 and the total number of iterations is set to $200,000$. To ease the training process of our framework, we warm start the remaining layers with fixed ImageNet pre-trained HRNet for the first $50,000$ iterations and then finetune all layers together for the rest $150,000$ iterations. The learning rate is set as 0.0002 using Adam [20] optimizer at first and then is linearly decayed to 0 for the last $50,000$ iterations.

5.1. Generalization Ability Evaluation

We first verify that supervised binary classifiers experience a significant performance drop over unseen fake images. To show this,

we adopt the state-of-the-art detector, *i.e.* Xception [36].

Table 1 summarizes the results in terms of AUC (area under the Receiver Operating Characteristic curve) with

[2]https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[3]https://deepfakedetectionchallenge.ai/dataset

Model	Training set		Test set AUC				
	DF	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	99.38	75.05	49.13	80.39	76.34
HRNet	✓	–	99.26	68.25	39.15	71.39	69.51
Face X-ray	✓	–	99.17	94.14	75.34	93.85	90.62
	✓	✓	99.12	**97.64**	**98.00**	**97.77**	**97.97**
	F2F	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	87.56	99.53	65.23	65.90	79.55
HRNet	✓	–	83.64	99.50	56.60	61.26	74.71
Face X-ray	✓	–	98.52	99.06	72.69	91.49	93.41
	✓	✓	**99.03**	99.31	**98.64**	**98.14**	**98.78**
	FS	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	70.12	61.70	99.36	68.71	74.91
HRNet	✓	–	63.59	64.12	99.24	68.89	73.96
Face X-ray	✓	–	93.77	92.29	99.20	86.63	93.13
	✓	✓	**99.10**	**98.16**	99.09	**96.66**	**98.25**
	NT	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	93.09	84.82	47.98	99.50	83.42
HRNet	✓	–	94.05	87.26	64.10	98.61	86.01
Face X-ray	✓	–	99.14	98.43	70.56	98.93	91.76
	✓	✓	**99.27**	**98.43**	**97.85**	99.27	**98.71**
	FF++	BI	DF	F2F	FS	NT	FF++
Xception [36]	–	✓	98.95	97.86	89.29	97.29	95.85
HRNet	–	✓	99.11	97.42	83.15	**98.17**	94.46
Face X-ray	–	✓	**99.17**	**98.57**	**98.21**	98.13	**98.52**

Table 1. Generalization ability evaluation. Using only classifier suffers performance drop on other unseen facial manipulations. Our approach improves the generalization ability by detecting face X-ray and further gets significant improvement using the self-supervised data. It is worth noting that our framework only using the self-supervised data still obtains promising results.

respect to each type of manipulated videos. We observe that excellent performance (above 99%) is obtained on the known specific manipulation, while the performance drops drastically for unseen manipulations. The reason may be that the model quickly overfits to the manipulation-specific artifacts, achieving high performance for the given data but suffering from poor generalization ability.

Our approach tackles the forgery detection by using a more general evidence: face X-ray. We show that the improved generalization ability comes from two factors: 1) we propose detecting the face X-ray instead of paying attention to the manipulation-specific artifacts; 2) we construct a large number of training sample automatically and effortlessly composited from real images so that the model is adapted to focus more on the face X-rays. Finally, we show that our method, even only using the self-supervised data, is capable of achieving a high detection accuracy.

The effect of detection using face X-ray. We first evaluate our model detecting face X-rays using the same training set and training strategy as Xception [36]. In order to obtain accurate Face X-rays for the manipulated images, we again adopt the generation process in Section 3.2 by considering the real image as background and the fake image as foreground, given a pair of a real image and a fake one. For fair comparison, we also show the results of the binary classifier using the same network architecture with ours, which is denoted as HRNet in the table. The comparison results are shown in Table 1. It can be clearly seen that our approach gets significantly improvement on the unseen facial manipulations, verifying our hypothesis that explicitly detecting face X-ray is more generalizable.

The effect of additional self-supervised data. Further, we train our framework with additional blended images that capture various types of face X-rays. The results are given in Table 1, showing that large improvement has been obtained again. We think that there are two advantages of using the additional blended images. One is the benefit of extra training data as it is known that increasing the training data always leads to better model and thus improved performance. Another important thing is that the region inside the boundary of blended images is actually real instead of synthesized, making the model less over-fitting to manipulation-specific artifacts.

Results of only using self-supervised data. Finally we present the results of our framework using only the self-supervised data, *i.e*. BI, in Table 1. The performance in terms of AUC on the four representative facial manipulations DF, F2F, FS, NT is $99.17\%, 98.57\%, 98.21\%, 98.13\%$ respectively. This shows that our model, even without fake images generated by any of the state-of-the-art face manipulation methods, still achieves a competitively high detection accuracy. We also show the results of classifer which is trained over BI by considering BI as fake images. The performance is much better than the previous above supervisely trained classifiers. This is probably because BI forced the classifier to learn the face X-rays, leading to better generalization. Nevertheless, our approach using face X-ray still gets overall better results and thus again validates the conclusion that using face X-ray is more generalizable.

5.2. Benchmark Results on Unseen Datasets

In order to facilitate the industry as well as the academia to develop advanced face forgery detectors, a growing number of datasets containing a large amount of high-quality deepfake videos have been released recently. Here we show the benchmark results of our framework on the detection of those unseen popular datasets.

Results in terms of forgery classification. We first show the forgery classification results in terms of AUC, AP (Average Precision) and EER (Equal Error Rate). The results of our framework on recent released large scale datasets, *i.e*. DFD, DFDC and Celeb-DF, are shown in Table 2. We also show the results of the state-of-the-art detector Xception [36] as a baseline. We can see that our framework, without using any images generated from facial manipulation methods, already performs better than the baseline. Moreover, if we exploit additional fake images even not from the same distribution as the test set, the performance

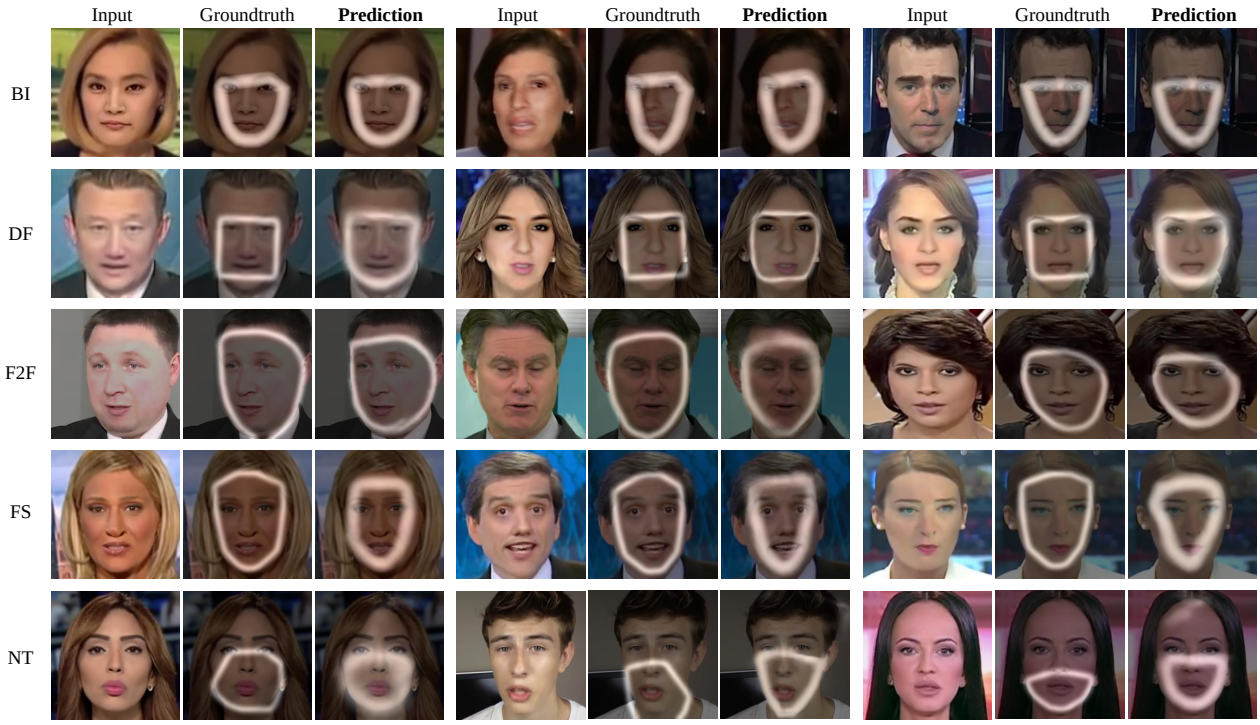


Figure 6. Visual results on various facial manipulation methods including our self-supervised generated blended images. For the facial manipulations, the groundtruth is obtained by computing the absolute element-wise difference between the manipulated image and the corresponding real image and then converting to grayscale followed by normalization. It can be clearly seen from the figure that the predicted face X-ray well captures the shape of the corresponding groundtruth. More visual results can be found in the supplementary.

Model	Training dataset	Test dataset								
		DFD			DFDC			Celeb-DF		
		AUC	AP	EER	AUC	AP	EER	AUC	AP	EER
Xception [36]	FF++	87.86	78.82	21.49	48.98	50.83	50.45	36.19	50.07	59.64
Face X-ray	BI	93.47	87.89	12.72	71.15	**73.52**	32.62	74.76	68.99	31.16
Face X-ray	FF++ and BI	**95.40**	**93.34**	**8.37**	**80.92**	72.65	**27.54**	**80.58**	**73.33**	**26.70**

Table 2. Benchmark results in terms of AUC, AP and EER for our framework and the state-of-the-art detector Xception [36] on unseen datasets. Our framework, trained in a self-supervised way, already performs better than the baseline. The performance is further greatly improved in most cases if we exploit additional fake images even not from the same distribution as the test set.

is further greatly improved in most cases.

Results in terms of face X-ray prediction. Our framework makes predictions about the forgery based on the existence of nontrivial face X-rays. We show that our method can reliably predict the face X-rays for unseen facial manipulations and thus provide explainable decisions for the inference. The visual examples on various types of fake images including the blended images generated in the proposed self-supervised learning are shown in Figure 6. For the facial manipulations, the groundtruth is obtained by computing the absolute element-wise difference between the manipulated image and the corresponding real image and then converting to grayscale followed by normalization. It can be clearly seen from the figure that the predicted face X-ray well captures the shape of the corresponding groundtruth.

5.3. Comparison with Recent Works

Some recent related works [23, 13, 11, 29] have also noticed the generalization issue and tried to solve the problem to a certain degree. FWA [23] also adopts a self-supervised way that creates negative samples from real images. Yet its goal is to characterize face warping artifacts only widely existed in DeepFake generated videos. The comparison is given in Table 3.

Three other related works are LAE [13], FT [11], both attempting to learn intrinsic representation instead of capturing artifacts in the training set, and MTDS [29] learning detection and localization simultaneously.

We present the comparison, which is evaluated over a new type of manipulated data when the model is trained on another type, in Table 4. Note that we directly cite the num-

Model	AUC	
	FF++/DF	Celeb-DF
FWA [23]	79.20	53.80
Face X-ray	**99.17**	**74.76**

Table 3. AUC comparison with FWA.

Model	Training set		Detection accuracy	
	F2F	FS	F2F	FS
LAE [13]	✓	–	90.93	63.15
FT-res [11]	✓	4 images	94.47	72.57
MTDS [29]	✓	–	92.77	54.07
Face X-ray	✓	–	**97.73**	**85.69**

Table 4. Detection accuracy comparison with recent methods. Note that here we use the HQ version (a light compression) of FF++ dataset for fair comparison.

bers from their original papers for fair comparison. From the two tables, we can see that our framework largely exceeds recent state-of-the-arts.

5.4. Analysis of the proposed framework

The effect of data augmentation. The overall goal of data augmentation in the self-supervised data generation is to offer a large amount of different types of blended images to give the model the ability to detect various manipulated images. Here, we study two important augmentation strategies: a) mask deformation which intends to bring larger variety to the shape of face X-ray; b) color correction in order to produce a more realistic blended image. We think that both strategies are crucial for generating diverse and high-quality data samples that are definitely helpful for network training.

To show this,

we present the comparison on FF++ and DFD in Table 5. It can be seen that both strategies are important and without either one of the two strategies will degrade the performance.

Generalization to other types of blending. We adopt alpha blending in the self-supervised data generation. Though we have demonstrated the performance of the proposed framework on unseen manipulations that might not use alpha blending, we here precisely present a study on the results of our approach with respect to Poisson blending, another widely used blending technique in existing face manipulation methods, and deep blending (GP-GAN [48]) which utilizes neural network rather than Equation (1). We construct the test data by using different types of blending and evaluate the model when the training data is constructed with alpha blending. The results are given in Table 6. We can see that our framework still gets satisfactory results on unseen blending types though with visible performance drops on Possion blending.

	AUC	
	FF++	DFD
w/o mask deformation	93.92	85.89
w/o color correction	96.21	89.91
Face X-ray	**98.52**	**93.47**

Table 5. Ablation study for the effect of (a) mask deformation and (b) color correction in the self-supervised data generation pipeline.

Blending type	AUC	AP	EER
Alpha blending	99.46	98.50	1.50
Possion blending	94.62	88.85	11.41
Deep blending [48]	99.90	98.77	1.36

Table 6. Results over test data using Possion blending and deep blending when the training data is constructed with alpha blending. Our framework still gets satisfactory results on unseen blending types though with visible performance drops on Poisson blending.

6. Limitations

While we have demonstrated satisfactory performance on general detection in the experiments, we are aware that there exist some limitations of our framework.

First, we realize that detecting face X-ray may fail in two aspects. 1) Our method relies on the existence of a blending step. Therefore, when an image is entirely synthetic, it is possible that our method may not work correctly. However the fake news

such as videos of someone saying and doing things they didn't, usually require blending as a post-processing step. This is because so far without blending, it is unlikely to completely generate a realistic image with desired target background. We indeed provide a promising way to detect those numerous blended forgeries. 2) We notice that

one can develop adversarial samples to against our detector.

This is inevitable since it is an arms race between image forgery creation and detection, which would inspire both fields to develop new and exciting techniques.

Besides, similar to all the other forgery detectors, our method also suffers from performance drop when encounter low resolution images. This is because classifying low resolution images is more challenging as the forgery evidence is less significant. We test our framework on the HQ version (a light compression) and the LQ version (a heavy compression) of FF++ dataset and the overall AUC are 87.35% and 61.6% respectively. This is expected as the heavier the compression, the less significant the forgery evidence and thus the lower the performance.

7. Conclusion

In this work, we propose a novel face forgery evidence, face X-ray, based on the observation that most existing face

manipulation methods share a common blending step and there exist intrinsic image discrepancies across the blending boundary, which is neglected in advanced face manipulation detectors. We develop a more general face forgery detector using face X-ray and the detector can be trained in a self-supervised manner, without fake images generated by any of the state-of-the-art face manipulation methods. Extensive experiments have been performed to demonstrate the generalization ability of face X-ray, showing that our framework is capable of accurately distinguishing unseen forged images and reliably predicting the blending regions.

References

[1] DeepFakes. www.github.com/deepfakes/faceswap. Accessed: 2019-09-18. 1, 5

[2] FaceSwap. www.github.com/MarekKowalski/FaceSwap. Accessed: 2019-09-30. 1, 5

[3] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 2

[4] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–45, 2019. 2

[5] Jawadul H Bappy, Amit K Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and BS Manjunath. Exploiting spatial structure for localizing manipulated image regions. In *Proceedings of the IEEE international conference on computer vision*, pages 4970–4979, 2017. 2

[6] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, 2019. 2

[7] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10. ACM, 2016. 2

[8] Tiziano Bianchi and Alessandro Piva. Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3):1003–1017, 2012. 2

[9] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014. 4

[10] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 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*, pages 159–164. ACM, 2017. 2

[11] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 3, 7, 8

[12] Xinyi Ding, Zohreh Raziei, Eric C Larson, Eli V Olinick, Paul Krueger, and Michael Hahsler. Swapped face detection using deep learning and subjective assessment. *arXiv preprint arXiv:1909.04217*, 2019. 1, 2

[13] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards generalizable forgery detection with locality-aware autoencoder. *arXiv preprint arXiv:1909.05999*, 2019. 1, 3, 7, 8

[14] Hany Farid. Image forgery detection–a survey. 2009. 3

[15] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, 2012. 2

[16] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 2

[17] Miroslav Goljan and Jessica Fridrich. Cfa-aware features for steganalysis of color images. In *Media Watermarking, Security, and Forensics 2015*, volume 9409, page 94090V. International Society for Optics and Photonics, 2015. 2

[18] Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang. Learning to detect fake face images in the wild. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pages 388–391. IEEE, 2018. 2

[19] Ali Khodabakhsh, Raghavendra Ramachandra, Kiran Raja, Pankaj Wasnik, and Christoph Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2018. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Neal Krawetz and Hacker Factor Solutions. A pictures worth... *Hacker Factor Solutions*, 6:2, 2007. 2

[22] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. Detection of deep network generated images using disparities in color components. *arXiv preprint arXiv:1808.07276*, 2018. 2

[23] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2, 2018. 3, 7, 8

[24] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 2019. 5

[25] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389. IEEE, 2018. 1, 2

[26] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019. 3

[27] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. Incremental learning for the detection and classification of gan-generated images. *arXiv preprint arXiv:1910.01568*, 2019. 2

[28] Huaxiao Mo, Bolin Chen, and Weiqi Luo. Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, pages 43–47. ACM, 2018. 2

[29] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. 2, 7, 8

[30] Xunyu Pan, Xing Zhang, and Siwei Lyu. Exposing image splicing with inconsistent local noise variances. In *2012 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2012. 2

[31] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003. 3

[32] Weize Quan, Kai Wang, Dong-Ming Yan, and Xiaopeng Zhang. Distinguishing between natural and computer-generated images using convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 13(11):2772–2787, 2018. 2

[33] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017. 2

[34] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 3

[35] 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 preprint arXiv:1803.09179*, 2018. 1, 2

[36] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint arXiv:1901.08971*, 2019. 1, 2, 5, 6, 7

[37] Seung-Jin Ryu, Matthias Kirchner, Min-Jeong Lee, and Heung-Kyu Lee. Rotation invariant localization of duplicated image regions based on zernike moments. *IEEE Transactions on Information Forensics and Security*, 8(8):1355–1370, 2013. 2

[38] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, 51:201–209, 2018. 2

[39] Victor Schetinger, Massimo Iuliani, Alessandro Piva, and Manuel M Oliveira. Digital image forensics vs. image composition: An indirect arms race. *arXiv preprint arXiv:1601.03239*, 2016. 2

[40] Kritaphat Songsri-in and Stefanos Zafeiriou. Complement face forensic detection and localization with facialland-marks. *arXiv preprint arXiv:1910.05455*, 2019. 2

[41] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation. *arXiv preprint arXiv:1910.01717*, 2019. 2

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019. 5

[43] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 5

[44] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, pages 81–87. ACM, 2018. 1, 2

[45] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv preprint arXiv:1904.12356*, 2019. 1, 5

[46] 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. 1, 5

[47] Run Wang, Lei Ma, Felix Juefei-Xu, Xiaofei Xie, Jian Wang, and Yang Liu. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019. 2

[48] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2487–2495. ACM, 2019. 8

[49] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the generalization of gan image forensics. In *Chinese Conference on Biometric Recognition*, pages 134–141. Springer, 2019. 1, 3

[50] 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. 3