On the Detection of Digital Face Manipulation

Hao Dang* Feng Liu* Joel Stehouwer* Xiaoming Liu Anil Jain
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824

Abstract

Detecting manipulated facial images and videos is an increasingly important topic in digital media forensics. As advanced face synthesis and manipulation methods are made available, new types of fake face representations are being created which have raised significant concerns for their use in social media. Hence, it is crucial to detect manipulated face images and localize manipulated regions. Instead of simply using multi-task learning to simultaneously detect manipulated images and predict the manipulated mask (regions), we propose to utilize an attention mechanism to process and improve the feature maps for the classification task. The learned attention maps highlight the informative regions to further improve the binary classification (genuine face v. fake face), and also visualize the manipulated regions. To enable our study of manipulated face detection and localization, we collect a large-scale database that contains numerous types of facial forgeries. With this dataset, we perform a thorough analysis of data-driven fake face detection. We show that the use of an attention mechanism improves facial forgery detection and manipulated region localization. The code and database are available at `cvlab.cse.msu.edu/project-ffd.html`.

1. Introduction

Human faces play an important role in human-human communication and association of side information, *e.g.*, gender and age with identity. For instance, face recognition is increasingly utilized in our daily life for applications such as access control and payment [50]. However, these advances also entice malicious actors to manipulate face images to launch attacks, aiming to be authenticated as the genuine user. Moreover, manipulation of facial content has become ubiquitous, and raises new concerns especially in social media content [41–43]. Recent advances in deep learning have led to a dramatic increase in the realism of face synthesis and enabled a rapid dissemination of "fake news" [7]. Therefore, to mitigate the adverse impact and benefit both *public security and privacy*, it is crucial to develop effective solutions against these facial forgery attacks.

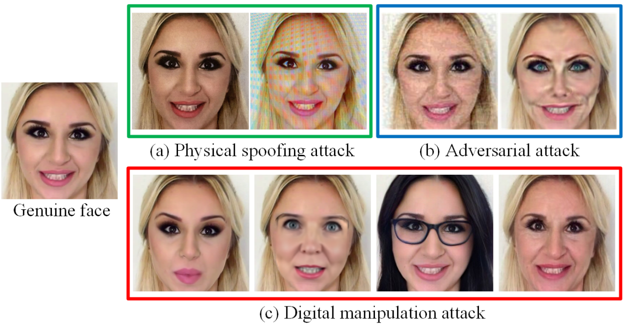
Genuine face
(a) Physical spoofing attack
(b) Adversarial attack
(c) Digital manipulation attack

Figure 1. Given one genuine face image, there are three types of facial forgery attacks: physical spoofing attack (print and replay attack), adversarial attack [18], and digital manipulation attack.

As shown in Fig. 1, there are *three* main types of facial forgery attacks. i) Physical spoofing attacks can be as simple as face printed on a paper, replaying image/video on a phone, or as complicated as a 3D mask [8, 24, 34, 35]. ii) Adversarial face attacks generate high-quality and perceptually imperceptible adversarial images that can evade automated face matchers [18, 20, 37, 57]. iii) Digital manipulation attacks, made feasible by Variational AutoEncoders (VAEs) [28, 40] and Generative Adversarial Networks (GANs) [19], can generate entirely or partially modified photorealistic face images. Among these three types, this work addresses only *digital manipulation attacks*, with the objectives of automatically detecting manipulated faces, as well as localizing modified facial regions. We use the term "face manipulation detection" or "face forgery detection" to describe our objective.

Digital facial manipulation methods fall into four categories: expression swap, identity swap, attribute manipulation and entire face synthesis (Fig. 2). 3D face reconstruction and animation methods [17, 32, 48, 64] are widely used for expression swap, such as *Face2Face* [47]. These methods can transfer expressions from one person to another in real time with only RGB cameras. Identity swap methods replace the face of one person with the face of an-

*denotes equal contribution by the authors.

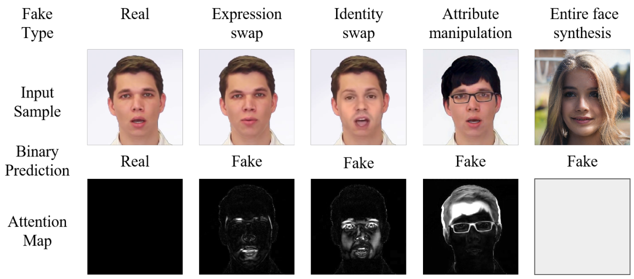

Figure 2. Our facial forgery detection method tackles faces generated by the four types of face manipulation methods. Given a face image, our approach outputs a binary decision (genuine v. manipulated face) and localizes the manipulated regions via an estimated attention map. For real or entirely synthetic faces, our estimated maps are assumed to be uniformly distributed in [0, 1].

other. Examples include *FaceSwap* [47, 53], which inserts famous actors into movie clips in which they never appeared and *DeepFakes* [3], which performs face swapping via deep learning algorithms.

Attribute manipulation edits single or multiple attributes in a face, *e.g.*, gender, age, skin color, hair, and glasses. The adversarial framework of GANs is used for image translation [23, 62, 63] or manipulation in a given context [10, 45], which diversifies facial images synthesis. *FaceApp* [4] has popularized facial attribute manipulation as a consumer-level application, providing 28 filters to modify specific attributes [4]. The fourth category is entire face synthesis. Fueled by the large amounts of face data and the success of GANs, any user is capable of producing a completely synthetic facial image, whose realism is such that even humans have difficulty assessing if it is genuine or manipulated [15, 25, 26].

Research on face manipulation detection has been seriously hampered by the lack of large-scale datasets of manipulated faces. Existing approaches are often evaluated on small datasets with limited manipulation types, including Zhou *et al.* [61], Deepfake [29], and FaceForensics/FaceForensics++ [41, 42]. To remedy this issue, we collect a Diverse Fake Face Dataset (DFFD) of 2.6 million images from all four categories of digital face manipulations.

Due to the fact that the modification of a face image can be in whole or in part, we assume that a well-learned network would gather different information *spatially* in order to detect manipulated faces. We hypothesize that correctly estimating this spatial information can enable the network to focus on these important spatial regions to make its decision. Hence, we aim to not only detect manipulated faces, but also automatically locate the manipulated regions by estimating an image-specific attention map, as in Fig. 3. We present our approach to estimate the attention map in both supervised and weakly-supervised fashions. We also demonstrate that this attention map is beneficial to the final task of facial forgery detection. Finally, in order to quantify the attention map estimation, we propose a novel metric for attention map accuracy evaluation. In the future, we anticipate the predicted attention maps for manipulated face images and videos would reveal hints about the type, magnitude, and even intention of the manipulation.

In summary, the contributions of this work include:

⋄ A comprehensive fake face dataset including 0.8M real and 1.8M fake faces generated by a diverse set of face modification methods and an accompanying evaluation protocol.

⋄ A novel attention-based layer to improve classification performance and produce an attention map indicating the manipulated facial regions.

⋄ A novel metric, termed Inverse Intersection Non-Containment (IINC), for evaluating attention maps that produces a more coherent evaluation than existing metrics.

⋄ State-of-the-art performance of digital facial forgery detection for both seen and unseen manipulation methods.

2. Related Work

Digital Face Manipulation Methods. With the rapid progress in computer graphics and computer vision, it is becoming difficult for humans to tell the difference between genuine and manipulated faces [42]. Graphics-based approaches are widely used for identity or expression transfer by first reconstructing 3D models for both source and target faces, and then exploiting the corresponding 3D geometry to warp between them. In particular, Thies *et al.* [46] present expression swap for facial reenactment with an RGB-D camera. *Face2Face* [47] is a real-time face reenactment system using only an RGB camera. Instead of manipulating expression only, the extended work [27] transfers the full 3D head position, rotation, expression, and eye blinking from a source actor to a portrait video of a target actor. "Synthesizing Obama" [45] animates the face based on an input audio signal. *FaceSwap* replaces the identity of 3D models while preserving the expressions.

Deep learning techniques, not surprisingly, are popular in synthesizing or manipulating faces [48]. The term *Deepfakes* has become a synonym for deep learning based face identity replacement [42]. There are various public implementations of *Deepfakes*, most recently by *ZAO* [5] and *FaceAPP* [4]. *FaceAPP* can selectively modify facial attributes [4]. GAN-based methods can produce entire synthetic faces, including non-face background [25, 26, 49].

Fake Face Benchmarks. Unfortunately, large and diverse datasets for face manipulation detection are limited in the community. Zhou *et al.* [61] collected a dataset with face-swapped images generated by an iOS app and an open-source software. Video-based face manipulation became available with the release of FaceForensics [41], which contains 0.5M *Face2Face* manipulated frames from over 1,000 videos. An extended version, FaceForensics++ [42], further

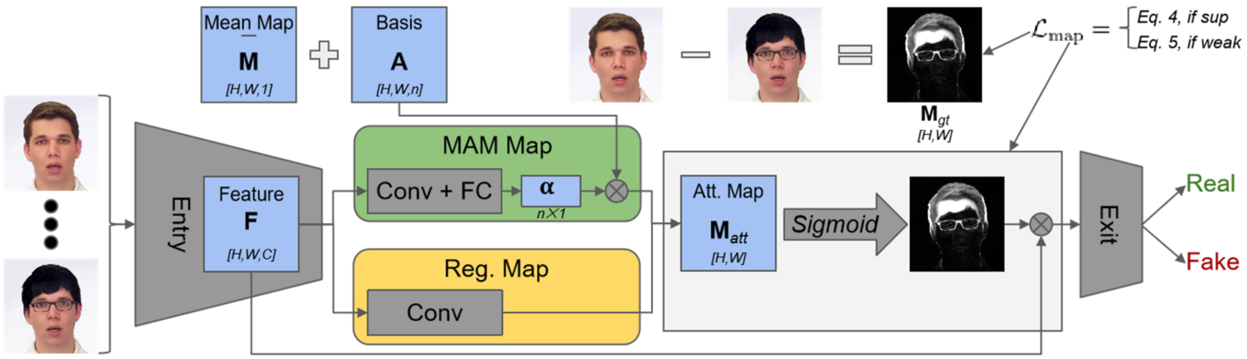

Figure 3. The architecture of our face manipulation detection. Given any backbone network, our proposed attention-based layer can be inserted into the network. It takes the high-dimensional feature $\mathbf{F}$ as input, estimates an attention map $\mathbf{M}_{att}$ using either *MAM*-based or *regression*-based methods, and channel-wise multiplies it with the high-dimensional features, which are fed back into the backbone. In addition to the binary classification supervision $\mathcal{L}_{\text{classifier}}$, either a supervised or weakly supervised loss, $\mathcal{L}_{\text{map}}$, can be applied to estimate the attention map, depending on whether the ground truth manipulation map $\mathbf{M}_{gt}$ is available.

augments the collection with *Deepfake* [3] and *FaceSwap* manipulations. However, these datasets are still limited to two fake types: identity and expression swap. To overcome this limitation, we collect the first fake face dataset with diverse fake types, including identity and expression swapped images from FaceForensics++, face attribute manipulated images using *FaceAPP*, and complete fake face images using StyleGAN [26] and PGGAN [25].

Manipulation Localization. There are two main approaches to localize manipulated image regions: segmenting the entire image [9, 39], and repeatedly performing binary classification via a sliding window [42]. These methods are often implemented via multi-task learning with additional supervision, yet they do not necessarily improve the final detection performance. In contrast, we propose an *attention mechanism* to automatically detect the manipulated region for face images, which requires very few additional trainable parameters. In computer vision, attention models have been widely used for image classification [12, 51, 56], image inpainting [33, 60] and object detection [11, 59]. Attention not only serves to select a focused location but also enhances object representations at that location, which is effective for learning generalizable features for a given task. A number of methods [22, 54, 55] utilize an attention mechanism to enhance the accuracy of CNN classification models. Residual Attention Network [51] improves the accuracy of the classification model using 3D self-attention maps. Choe *et al*. [14] propose an attention-based dropout layer to process the feature maps of the model, which improves the localization accuracy of CNN classifiers. To our knowledge, this is the *first* work to develop the attention mechanism to face manipulation detection and localization.

3. Proposed Method

We pose the manipulated face detection as a binary classification problem using a CNN-based network. We further propose to utilize the attention mechanism to process the feature maps of the classifier model. The learned attention maps can highlight the regions in an image which influence the CNN's decision, and further be used to guide the CNN to discover more discriminative features.

3.1. Motivation for the Attention Map

Assuming the attention map can highlight the manipulated image regions, and thereby guide the network to detect these regions, this alone should be useful for the face forgery detection. In fact, each pixel in the attention map would compute a probability that its receptive field corresponds to a manipulated region in the input image. Digital forensics has shown that camera model identification is possible due to "fingerprints" in the high-frequency information of a real image [13]. It is thus feasible to detect abnormalities in this high-frequency information due to algorithmic processing. Hence we insert the attention map into the backbone network where the receptive field corresponds to appropriately sized local patches. Then, the features before the attention map encode the high-frequency fingerprint of the corresponding patch, which may discriminate between real and manipulated regions at the local level.

Three major factors were considered during the construction and development of our attention map; *i*) explainability, *ii*) usefulness, and *iii*) modularity.

Explainability: Due to the fact that a face image can be modified entirely or in part, we produce an attention map that predicts where the *modified* pixels are. In this way, an auxiliary output is produced to explain which spatial regions the network based its decision on. This differs from prior works in that we use the attention map as a mask to remove any irrelevant information from the high-dimensional features *within the network*. During training, for a face image where the entire image is real, the attention map should ignore the entire image. For a modified or generated face, at least some parts of the image are manipulated, and therefore the ideal attention map should focus only on these parts.

Usefulness: One objective of our proposed attention map is that it enhances the final binary classification performance of the network. This is accomplished by feeding the attention map back into the network to ignore non-activated regions. This follows naturally from the fact that modified images may only be *partially modified*. Through the attention map, we can remove the real regions of a partial fake image so that the features used for final binary classification are purely from modified regions.

Modularity: To create a truly utilitarian solution, we take great care to maintain the modularity of the solution. Our proposed attention map can be implemented easily and plugged into existing backbone networks, through the inclusion of a single convolution layer, its associated loss functions, and masking the subsequent high-dimensional features. This can even be done while leveraging pre-trained networks by initializing only the weights that are used to produce the attention map.

3.2. Attention-based Layer

As shown in Fig. 3, the attention-based layer can be applied to any feature map of a classification model, and focus the network's attention on discriminative regions. Specifically, the input of the attention-based layer is a convolutional feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$, where H, W, C are height, width, and the number of channels, respectively. For simplicity, we omit the mini-batch dimension in this notation. Then we can generate an attention map $\mathbf{M}_{att} = \Phi(\mathbf{F}) \in \mathbb{R}^{H \times W}$ by processing $\mathbf{F}$, where $\Phi(\cdot)$ denotes the processing operator. The output of attention module is the refined feature map $\mathbf{F}'$, which is calculated as:

$$\mathbf{F}' = \mathbf{F} \odot \text{Sigmoid}(\mathbf{M}_{att}), \tag{1}$$

where $\odot$ denotes element-wise multiplication. The intensity of each pixel in the attention map is close to 0 for the real regions, and close to 1 for the fake regions. In other words, the pixel of the attention map indicates the probability of the original image patch being a fake region. This helps the subsequent backbone network to focus its processing to the non-zeros areas of the attention map, *i.e.*, the fake regions. Here, we propose two approaches to implement $\Phi(\cdot)$: manipulation appearance model and direct regression.

Manipulation Appearance Model (MAM). We assume that any manipulated map can be represented as a linear combination of a set of map prototypes:

$$\mathbf{M}_{att} = \bar{\mathbf{M}} + \mathbf{A} \cdot \alpha, \tag{2}$$

where $\bar{\mathbf{M}} \in \mathbb{R}^{(H \cdot W) \times 1}$ and $\mathbf{A} \in \mathbb{R}^{(H \cdot W) \times n}$ are the predefined average map and basis functions of maps. Thus the attention map generation can be translated to estimate the weight parameter $\alpha \in \mathbb{R}^{n \times 1}$, for each training image. We utilize one additional convolution and one fully connected layer to regress the weights from the feature map $\mathbf{F}$ (Fig. 3).

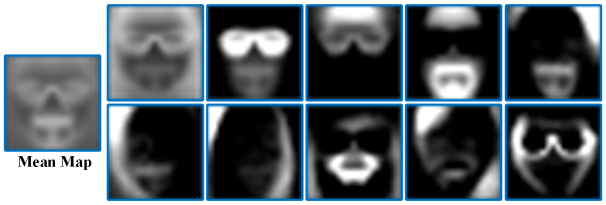

Figure 4. Mean map $\bar{\mathbf{M}}$ and 10 basis components $\mathbf{A}$.

The benefit of our proposed MAM is two fold. First, this constrains the solution space of map estimation. Second, the complexity of the attention estimation is decreased, which is helpful for generalization. To calculate the statistical bases $\mathbf{A}$, we apply Principal Component Analysis (PCA) to 100 ground-truth manipulation masks computed from *FaceAPP*. The first 10 principal components are used as bases, *i.e.*, $n = 10$. Fig. 4 shows the mean map and 10 bases (or templates).

Direct Regression. Another way to implement $\Phi(\cdot)$ is to estimate the attention map via a convolutional operation f: $\mathbf{F} \xrightarrow{f} \mathbf{M}_{att}$. f can consist of multiple convolutional layers or a single layer. This direct regression method is simple, yet effective, for adaptive feature refinement. Later we show that the benefits of our proposed attention-based layer are realized regardless of the choice of backbone networks. This further validates our claim that the proposed solution is modular and improves the usefulness and flexibility of the attention map.

3.3. Loss Functions

To train the binary classification network, we may begin with a pre-trained backbone network or to learn the backbone from scratch. Either way, the overall training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{classifier}} + \lambda * \mathcal{L}_{\text{map}}, \tag{3}$$

where $\mathcal{L}_{\text{classifier}}$ is the binary classification loss of Softmax and $\mathcal{L}_{\text{map}}$ is the attention map loss. λ is the loss weight.

For attention map learning, we consider three different cases: supervised, weakly supervised, and unsupervised.

Supervised learning. If the training samples are paired with ground truth attention masks, we can train the network in a supervised fashion, using Eqn. 4.

$$\mathcal{L}_{\text{map}} = ||\mathbf{M}_{att} - \mathbf{M}_{gt}||_1, \tag{4}$$

where $\mathbf{M}_{gt}$ is the ground truth manipulation mask. We use zero-maps as the $\mathbf{M}_{gt}$ for real faces, and one-maps as the $\mathbf{M}_{gt}$ for entirely synthesized fake faces. For partially manipulated faces, we pair fake images with their corresponding source images, compute the absolute pixel-wise difference in the RGB channels, convert into grayscale, and di-

Table 1. Comparison of fake face datasets along different aspects: number of still images, number of videos, number of fake types (identity swap (Id. swap), expression swap (Exp. swap), attributes manipulation, and entire image synthesis (Entire syn.)) and pose variation.

Dataset	Year	# Still images		# Video clips		# Fake types				Pose
		Real	Fake	Real	Fake	Id. swap	Exp. swap	Attr. mani.	Entire syn.	variation
Zhou *et al.* [61]	2018	$2,010$	$2,010$	-	-	2	-	-	-	Unknown
Yang *et al.* [58]	2018	241	252	49	49	1	-	-	-	Unknown
Deepfake [29]	2018	-	-	-	620	1	-	-	-	Unknown
FaceForensics++ [42]	2019	-	-	$1,000$	$3,000$	2	1	-	-	$[-30^\circ, 30^\circ]$
FakeSpotter [52]	2019	$6,000$	$5,000$	-	-	-	-	-	2	Unknown
DFFD (our)	2019	$58,703$	$240,336$	$1,000$	$3,000$	2	1	$28+40$	2	$[-90^\circ, 90^\circ]$

vide by 255 to produce a map in the range of $[0, 1]$. We empirically determine the threshold of 0.1 to obtain the binary modification map as $\mathbf{M}_{gt}$. We posit this strong supervision can help attention-based layer to learn the most discriminative regions and features for fake face detection.

Weakly supervised learning. For partially manipulated faces, sometimes the source images are not available. Hence, we can not obtain the ground truth manipulation mask as described above. However, we would still like to include these faces in learning the attention maps. To this end, we propose a weak supervision map loss as in Eqn. 5:

$$\mathcal{L}_{\text{map}} = \begin{cases} |\text{Sigmoid}(\mathbf{M}_{att}) - 0|, & \text{if real} \\ |\max(\text{Sigmoid}(\mathbf{M}_{att})) - 0.75|. & \text{if fake} \end{cases} \quad (5)$$

This loss drives the attention map to remain un-activated for real images, *i.e.*, all 0. For fake images, regardless of entire or partial manipulation, the maximum map value across the entire map should be sufficiently large, 0.75 in our experiments. Hence, for partial manipulation, an arbitrary number of the map values can be zeros, as long as at least one modified local region has a large response.

Unsupervised learning. The proposed attention module can also allow us to train the network without any map supervision when λ_m is set to 0. With only image-level classification supervision, the attention map learns informative regions automatically. More analysis of these losses is available in the experiments section.

4. Diverse Fake Face Dataset

One of our contributions is the construction of a dataset with diverse types of fake faces, termed Diverse Fake Face Dataset (DFFD). Compared with previous datasets in Tab. 1, DFFD contains greater diversity, which is crucial for detection and localization of face manipulations.

Data Collection. In Sec. 1, we introduced four main facial manipulation types: identity swap, expression swap, attribute manipulation, and entire synthesized faces. We thus collect data from these four categories by adopting respective state-of-the-art (SOTA) methods to generate fake images. Among all images and video frames, 47.7% are from male subjects, 52.3% are from females, and the majority of samples are from subjects in the range 21-50 years of age. For the face size, both real and fake samples have both low quality and high quality images. This ensures that the distributions of gender, age, and face size are less biased.

Real face images. We utilize FFHQ [26] and CelebA [36] datasets as our real face samples since the faces contained therein cover comprehensive variations in race, age, gender, pose, illumination, expression, resolution, and camera capture quality. We further utilize the source frames from FaceForensics++ [42] as additional real faces.

Identity and expression swap. For facial identity and expression swap, we use all the video clips from FaceForensics++ [42]. The FaceForensics++ contains $1,000$ real videos collected from YouTube and their corresponding $3,000$ manipulated versions which are divided into two groups: identity swap using *FaceSwap* and *Deepfake* [3], and expression swap using *Face2Face* [47]. From a public website [1], we collect additional identity swap data, which are videos generated by *Deep Face Lab* (DFL) [2].

Attributes manipulation. We adopt two methods *FaceAPP* [4] and StarGAN [15] to generate attribute manipulated images, where $4,000$ faces of FFHQ and $2,000$ faces of CelebA are the respective input real images. *FaceAPP*, as a consumer-level smartphone app, provides 28 filters to modify specified facial attributes, *e.g.*, gender, age, hair, beard, and glasses. The images are randomly modified with an automated script running on Android devices. For each face in FFHQ, we generate three corresponding fake images: two with a single random manipulation filter, and one with multiple manipulation filters. For each face in CelebA, we generate 40 fake images by StarGAN, a GAN-based image-to-image translation method. In total, we collect 92K attribute manipulated images.

Entire face synthesis. Recent works such as PGGAN [25] and StyleGAN [26] achieve remarkable success in realistic face image synthesis. PGGAN proposes a progressive training scheme both for generator and discriminator, which can produce high-quality images. StyleGAN redesigns the generator by borrowing from style transfer literature. Thus, we use the pre-trained model of PGGAN and StyleGAN to create 200k and 100k high-quality entire fake images, respectively. Figure 5 shows examples of DFFD.

Pre-processing. InsightFace [21] is utilized to estimate the bounding box and 5 landmarks for each image. We discard images whose detection or alignment fails. We further gen-

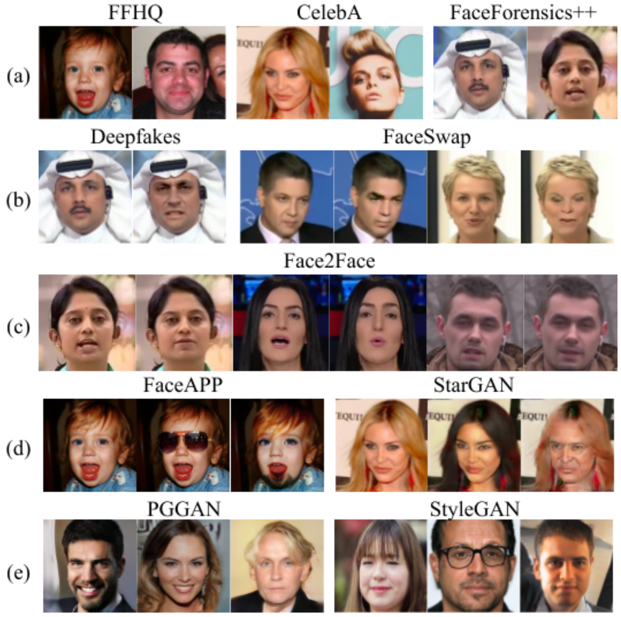

Figure 5. Example faces in our DFFD. (a) Real images/frames from FFHQ, CelebA and FaceForensics++ datasets; (b) Paired face identity swap images from FaceForensics++ dataset; (c) Paired face expression swap images from FaceForensics++ dataset; (d) Attributes manipulated examples by *FaceAPP* and StarGAN; (e) Entire synthesized faces by PGGAN and StyleGAN.

erate ground truth manipulation masks for fake images as described in Sec. 3.3. To enforce consistency, if a fake face image is derived from a source real face image, we use the same landmarks of the real face image for face cropping.

Protocols. We collect 781,727 samples for real image, and 1,872,007 samples for fake ones. Within these samples, we randomly select a subset of 58,703 real images and 240,336 fake ones to make the size of our dataset manageable and to balance the size of each sub-category. For video samples, we extract one frame per second in order to reduce the size without sacrificing the diversity of DFFD. We randomly split the data into 50% for training, 5% for validation and 45% for testing. All fake images manipulated from the same real image are in the same set as the source image.

5. Experimental Results

5.1. Experimental Setup

Implementation Details: The loss weight λ is set to 1 and the batch size is 16, where each mini-batch consists of 8 real and 8 fake images. We use XceptionNet [16] and VGG16 [44] as backbone networks. Both two networks were pre-trained on ImageNet and fine-tuned on DFFD. Adam optimizer is used with a learning rate of 0.0002 in all experiments. Depending on the backbone architecture, we train for 75k-150k iterations, which requires less than 8 hours on an NVidia GTX 1080Ti.

Table 2. Ablation for the benefit of the attention map, with various combinations of map generation methods and supervisions. [Key: top performance for **supervised** and weakly supervised methods]

Map Supervision	AUC	EER	$\text{TDR}_{0.01\%}$	$\text{TDR}_{0.1\%}$	PBCA
Xception	99.61	2.88	77.42	85.26	–
+ Reg., *unsup.*	99.76	2.16	77.07	89.70	12.89
+ Reg., *weak sup.*	99.66	2.57	46.57	75.20	30.99
+ Reg., *sup.*	99.64	**2.23**	**83.83**	**90.78**	**88.44**
+ Reg., *sup.* - map	**99.69**	2.73	48.54	72.94	**88.44**
+ MAM, *unsup.*	99.55	3.01	58.55	77.95	36.66
+ MAM, *weak sup.*	99.68	2.64	72.47	82.74	69.49
+ MAM, *sup.*	99.26	3.80	77.72	86.43	85.93
+ MAM, *sup.* - map	98.75	6.24	58.25	70.34	85.93

Metrics: For all experiments, we utilize the protocols defined in Sec. 4. For detection, we report Equal Error Rate (EER), Area Under Curve (AUC) of ROC, True Detect Rate (TDR) at False Detect Rate (FDR) of 0.01% (denoted as $\text{TDR}_{0.01\%}$), and TDR at FDR of 0.1% (denoted as $\text{TDR}_{0.1\%}$). For localization, with known ground-truth masks, we report Pixel-wise Binary Classification Accuracy (PBCA), which treats each pixel as an independent sample to measure classification accuracy, Intersection over Union (IoU), and Cosine similarity between two vectorized maps. We also propose a novel metric, termed Inverse Intersection Non-Containment (IINC) for evaluating face manipulation localization performance, as described in Sec. 5.4.

5.2. Ablation Study

Benefit of Attention map: We utilize the SOTA XceptionNet [16] as our backbone network. It is based on depthwise separable convolution layers with residual connections. We convert XceptionNet into our model by inserting the attention-based layer between Block 4 and Block 5 of the middle flow, and then fine-tune on DFFD training set.

In Tab. 2, we show a comparison of the direct regression (Reg.) and manipulation appearance model (MAM) with different supervision strategies, *i.e.*, unsupervised (*unsup.*), weakly supervised (*weak sup.*) and supervised (*sup.*) learning. While four detection metrics are listed for completeness, considering the overall strong performance of some metrics and the preferred operational point of low FDR in practice, TDR at low FDR (*i.e.,* $\text{TDR}_{0.01\%}$) should be the primary metric for comparing various methods.

Unsurprisingly, the supervised learning outperforms the weakly supervised and unsupervised, in both the detection and localization accuracies. Additionally, comparing two map estimation approaches, regression-based method performs better with supervision. In contrast, MAM-based method is superior for weakly supervised or unsupervised cases as MAM offers strong constraint for map estimation.

Finally, instead of using the softmax output, an alternative is to use the average of the estimated attention map for detection, since loss functions encourage low attention values for real faces while higher values for fake ones. The

Table 3. Our attention layer in two backbone networks.

Network	AUC	EER	$\text{TDR}_{0.01\%}$	$\text{TDR}_{0.1\%}$	PBCA
Xception	99.61	2.88	77.42	85.26	-
Xception + Reg.	**99.64**	**2.23**	**83.83**	**90.78**	**88.44**
Xception + MAM	99.26	3.80	77.72	86.43	85.93
VGG16	96.95	8.43	0.00	51.14	-
VGG16 + Reg.	99.46	3.40	44.16	61.97	**91.29**
VGG16 + MAM	**99.67**	**2.66**	**75.89**	**87.25**	86.74

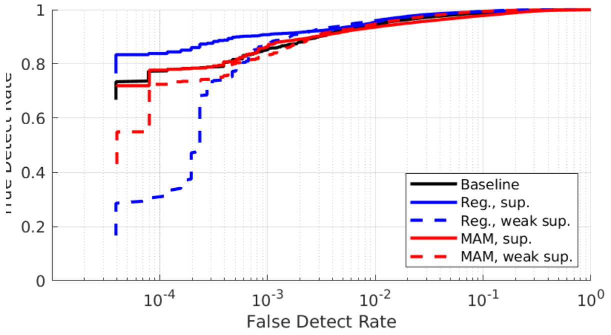

Figure 6. Forgery detection ROCs of the XceptionNet backbone with and without the attention mechanism.

performance of this alternative is shown in the rows of '*, *sup.* - map' in Tab. 2. Although this is not superior to the softmax output, it demonstrates that the attention map is itself useful for the facial forgery detection task.

Effect on Backbone Networks: We also report the results of a shallower backbone network VGG16 [44]. Tab. 3 compares XceptionNet and VGG16 with and without the attention layer. Both Reg. and MAM models are trained in the supervised case. We observe that using attention mechanism does improve the detection on both backbones.

Specifically, with a large and deep network (XceptionNet), the attention map can be directly produced by the network given the large parameter space. This directly produced attention map can better predict the manipulated regions than a map estimated from the MAM bases. Inversely, when using a smaller and shallower network (VGG16), we find that the direct production of the attention map causes contention in the parameter space. Hence including the prior of the MAM bases reduces this contention and allows for increased detection performance, though its estimation of the manipulated regions is constrained by the map bases.

5.3. Forgery Detection Results

We first show the ROCs on our DFFD in Fig. 6. Obviously, the direct regression approach for the attention map produces the best performing network at low FDR, which is not only the most challenging scenario, but also the most relevant to the practical applications. In addition, the proposed attention layer substantially outperforms the conventional XceptionNet, especially at lower FDR. Fig. 7 plots binary classification accuracy of our *Reg., sup* and baseline for different fake types of DFFD. The proposed approach

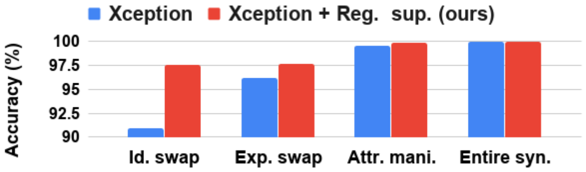

Figure 7. Binary classification accuracy for different fake types.

Table 4. AUC (%) on UADFV and Celeb-DF.

Methods	Training data	UADFV [58]	Celeb-DF [31]
Two-stream [61]	Private data	85.1	55.7
Meso4 [6] MesoInception4 [6]	Private data	84.3 82.1	53.6 49.6
HeadPose [58]	UADFV	89.0	54.8
FWA [30]	UADFV	97.4	53.8
VA-MLP [38] VA-LogReg [38]	Private data	70.2 54.0	48.8 46.9
Multi-task [39]	FF	65.8	36.5
Xception-FF++ [42]	FF++	80.4	38.7
Xception	DFFD	75.6	63.9
Xception	UADFV	96.8	52.2
Xception	UADFV, DFFD	97.5	67.6
Xception+Reg.	DFFD	84.2	64.4
Xception+Reg.	UADFV	**98.4**	57.1
Xception+Reg.	UADFV, DFFD	**98.4**	**71.2**

benefits forgery detection of all considered fake types, especially for the facial identity and expression swap.

We further validate our model on public datasets, where SOTA facial forgery detection methods have been tested. Table 4 summarizes the performance of all methods. Note that the performances shown here are not strictly comparable since not all methods are trained on the same dataset. First, we evaluate on the UADFV and Celeb-DF datasets with the models trained with DFFD. As shown in Tab. 4, our proposed approach significantly outperforms all the baselines on Celeb-DF and achieves competitive results on UADFV. FWA [30] and HeadPose [58] demonstrate superior performance on UADFV partially because they are trained on the same UADFV dataset, while this data source is not in our DFFD. Second, for a fair comparison, we train our method and baseline Xception on UADFV training set. In this case, our method outperforms all baselines on UADFV, and still shows superior generalization on Celeb-DF. Third, the results in Tab. 4 also help us to identify both the *source* and *amount* of improvements. For example, $75.6\% \rightarrow 84.2\%$ is an improvement due to attention mechanism while $52.2\% \rightarrow 63.9\%$ and $57.1\% \rightarrow 64.4\%$ are due to the greater diversity of DFFD dataset.

5.4. Manipulation Localization Results

We utilize three metrics for evaluating the attention maps: Intersection over Union (IoU), Cosine Similarity, and Pixel-wise Binary Classification Accuracy (PBCA). However, these three metrics are inadequate for robust evaluation of these diverse maps. Thus, we propose a novel metric defined in Eqn. 6, termed Inverse Intersection Non-

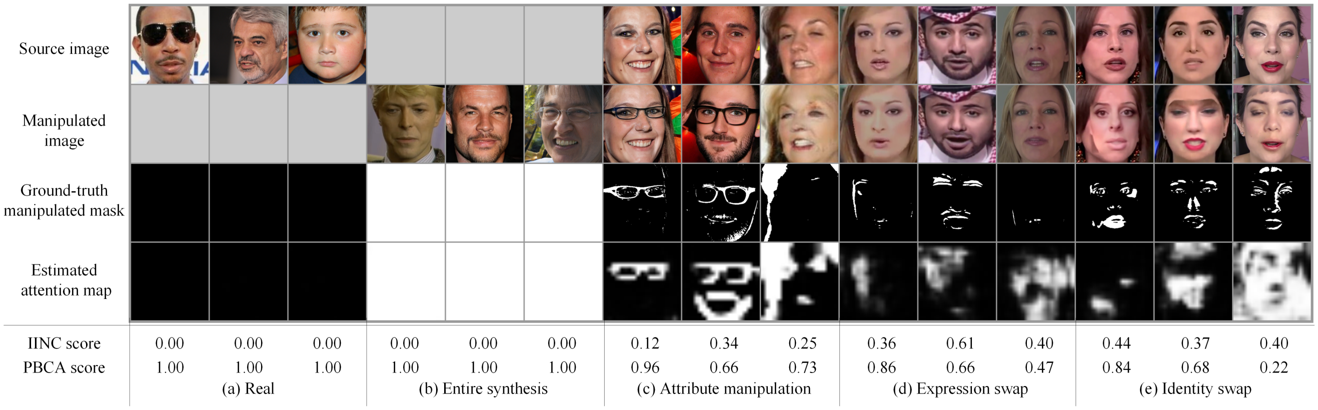

Figure 8. Estimated attention maps by applying Xception + Reg. *sup.* model to real and 4 types of manipulated images, with IINC and PBCA scores computed w.r.t. ground truth. While the overall areas of the attention maps are correct, their fidelity could be further improved.

Table 5. Evaluating manipulation localization with 4 metrics.

Data	IINC ↓	IoU ↑	Cosine Similarity ↓	PBCA ↑
All Real	0.015	–	–	0.998
All Fake	0.147	0.715	0.192	0.828
Partial	0.311	0.401	0.429	0.786
Complete	0.077	0.847	0.095	0.847
All	0.126	–	–	0.855

Containment (IINC), to evaluate the predicted maps:

$$\mathrm{IINC} = \frac{1}{3-|\mathbf{U}|} * \begin{cases} 0 & \text{if } \overline{\mathbf{M}}_{\text{gt}} = 0 \text{ and } \overline{\mathbf{M}}_{\text{att}} = 0 \\ 1 & \text{if } \overline{\mathbf{M}}_{\text{gt}} = 0 \text{ xor } \overline{\mathbf{M}}_{\text{att}} = 0 \\ (2 - \frac{|\mathbf{I}|}{|\mathbf{M}_{\text{att}}|} - \frac{|\mathbf{I}|}{|\mathbf{M}_{\text{gt}}|}) & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathbf{I}$ and $\mathbf{U}$ are the intersection and union between the ground truth map, $\mathbf{M}_{gt}$, and the predicted map, $\mathbf{M}_{att}$, respectively. $\overline{\mathbf{M}}$ and $|\mathbf{M}|$ are the mean and L_1 norm of $\mathbf{M}$, respectively. The two fractional terms measure the ratio of the area of the intersection w.r.t. the area of each map, respectively. IINC improves upon other metrics by measuring the non-overlap ratio of both maps, rather than their combined overlap, as in IoU. Additionally, the IoU and Cosine Similarity are undefined when either map is uniformly 0, which is the case for real face images.

The benefits of IINC as compared to other metrics are shown in Fig. 9. Note that IOU and Cosine similarity are not useful for cases (a-c), where the scores are the same, but the maps have vastly different properties. Similarly, PBCA is not useful for the cases (e-g), as the ratio of misclassification is not represented in PBCA. For example, case (g) over-estimates by a factor of 100% and case (e) over-estimates by 200%, while case (f) both over- and under-estimates by 150%. The IINC provides the optimal ordering by producing the same order as IOU when it is useful, cases (d-g), and similarly with PBCA when it is useful, cases (a-c). Thus, IINC is a more robust metric for comparing the attention maps than the previous metrics.

The localization ability of our Xception + Reg. *sup.* model to predict the attention maps is shown in Tab. 5. In Fig. 8, we show the IINC and PBCA for some test examples.

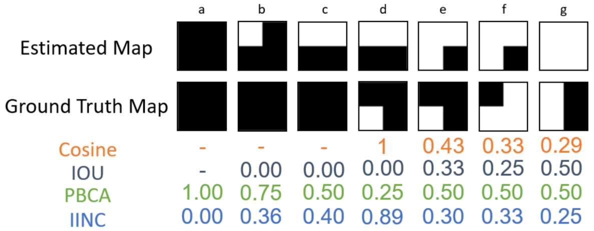
a
b
c
d
e
f
g
Estimated Map
Ground Truth Map
Cosine - - - 1 0.43 0.33 0.29
IOU - 0.00 0.00 0.00 0.33 0.25 0.50
PBCA 1.00 0.75 0.50 0.25 0.50 0.50 0.50
IINC 0.00 0.36 0.40 0.89 0.30 0.33 0.25

Figure 9. A toy-example comparing 4 metrics in evaluating attention maps. White is the manipulated pixel and black is the real pixel. IOU and Cosine metrics do not adequately reflect the differences in cases (a-c), while PBCA is not useful for cases (e-g). In contrast, the proposed IINC is discriminative in all cases.

The ordering of the IINC scores aligns with qualitative human analysis. The first cases in (d) and (e) are examples where the PBCA is high only because the majority of each map is non-activated. The IINC is more discriminative in these cases due to the non-overlap between the maps. For the third cases in (d) and (e), the IINC produces the same score because the maps display the same behavior (a large amount of over-activation), whereas the PBCA prefers the example in (d) because its maps have fewer activations.

6. Conclusion

We tackle the digitally manipulated face image detection and localization task. Our proposed method leverages an attention mechanism to process the feature maps of the detection model. The learned attention maps highlight the informative regions for improving the detection ability and also highlight the manipulated facial regions. In addition, we collect the first facial forgery dataset that contains diverse types of fake faces. Finally, we empirically show that the use of our attention mechanism improves facial forgery detection and manipulated facial region localization. This is the first unified approach that tackles a diverse set of face manipulation attacks, and also achieves the SOTA performance in comparison to previous solutions.

References

[1] https://www.patreon.com/ctrl_shift_face. Accessed: 2019-09-04. 5

[2] https://github.com/iperov/DeepFaceLab. Accessed: 2019-09-04. 5

[3] Deepfakes github. https://github.com/deepfakes/faceswap. Accessed: 2019-09-11. 2, 3, 5

[4] FaceApp. https://faceapp.com/app. Accessed: 2019-09-04. 2, 5

[5] ZAO. https://apps.apple.com/cn/app/zao/id1465199127. Accessed: 2019-09-16. 2

[6] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a compact facial video forgery detection network. In *WIFS*, 2018. 7

[7] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *ICCVW*, 2019. 1

[8] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based CNNs. In *IJCB*, 2017. 1

[9] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *PAMI*, 2017. 3

[10] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. GAN dissection: Visualizing and understanding generative adversarial networks. In *ICLR*, 2019. 2

[11] Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *ICCV*, 2015. 3

[12] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *CVPR*, 2015. 3

[13] Chang Chen, Zhiwei Xiong, Xiaoming Liu, and Feng Wu. Camera trace erasing. In *CVPR*, 2020. 3

[14] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, 2019. 3

[15] 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 *CVPR*, 2018. 2, 5

[16] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 6

[17] Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. In *TOG*, 2011. 1

[18] Debayan Deb, Jianbang Zhang, and Anil K Jain. AdvFaces: Adversarial face synthesis. *arXiv preprint arXiv:1908.05008*, 2019. 1

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[20] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2014. 1

[21] Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense U-Nets with dual transformers for robust face alignment. In *BMVC*, 2018. 5

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[24] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face despoofing: Anti-spoofing via noise modeling. In *ECCV*, 2018. 1

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 2, 3, 5

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3, 5

[27] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *TOG*, 2018. 2

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1

[29] Pavel Korshunov and Sébastien Marcel. DeepFakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 2, 5

[30] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *CVPRW*, 2019. 7

[31] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A new dataset for deepfake forensics. In *CVPR*, 2020. 7

[32] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3D face shapes for joint face reconstruction and recognition. In *CVPR*, 2018. 1

[33] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 3

[34] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018. 1

[35] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *CVPR*, 2019. 1

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1

[38] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *WACVW*, 2019. 7

[39] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *BTAS*, 2019. 3, 7

[40] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014. 1

[41] 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

[42] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 1, 2, 3, 5, 7

[43] 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*, 2017. 1

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6, 7

[45] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *TOG*, 2017. 2

[46] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *TOG*, 2015. 2

[47] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *CVPR*, 2016. 1, 2, 5

[48] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. *TPAMI*, 2019. 1, 2

[49] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, 2017. 2

[50] Luan Tran, Xi Yin, and Xiaoming Liu. Representation learning by rotating your faces. *TPAMI*, 2018. 1

[51] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017. 3

[52] 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. 5

[53] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *ICCV*, 2019. 2

[54] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[55] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018. 3

[56] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015. 3

[57] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 2020. 1

[58] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, 2019. 5, 7

[59] Donggeun Yoo, Sunggyun Park, Joon-Young Lee, Anthony S Paek, and In So Kweon. AttentionNet: Aggregating weak directions for accurate object detection. In *ICCV*, 2015. 3

[60] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 3

[61] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *CVPRW*, 2017. 2, 5, 7

[62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2

[63] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 2

[64] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. In *Computer Graphics Forum*, 2018. 1