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

Computer Vision Lab

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Facial Forgery Detection

DFFD

Digital Face Manipulation

Problem Statement:

Given an image of a human face, determine if the face has been digitally manipulated or generated, while localizing the manipulated regions.

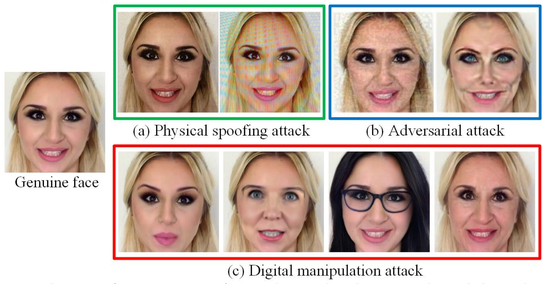


Fig 1: Three attack types for a genuine face: physical, adversarial, and digital manipulation

Insights and Contributions:

✧ Novel database (DFFD) for analysis of fake face detection methods.

✧ Attention mechanism for the localization of manipulated regions.

✧ Novel Inverse Intersection Non-Containment (IINC) metric for manipulated regions.

✧ SOTA fake face detection performance on DFFD and other Anti-Fake datasets.

Diverse Fake Faces Dataset

The DFFD dataset is collected by combining multiple prior datasets into a comprehensive collection. We utilize 3 separate real image sources and 8 algorithms to produce fake faces.

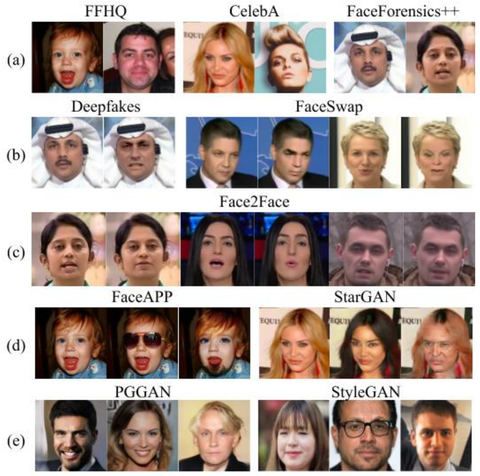


Fig 2: Sample images in DFFD.

Overall, the DFFD dataset includes ~60k real images, 240k fake images, and 5 image types.

(a) Genuine Images

(b) Identity Swap – Transferring the identity from one face image to another.

(c) Expression Swap – Transferring the expression while preserving the identity from 2 faces.

(d) Attribute Manipulation – Modifying 1+ facial attributes, while preserving the identity.

(e) Entire Synthesis – Synthesis of a novel face image and identity from random initialization.

Network Architecture and Attention Mechanism

Defining the Manipulation Mask:

A pixel is manipulated if it varies from the original pixel by at least 0.1 in the image range [0,1].

For real images, the mask is all 0s. For entirely synthetic images, the mask is all 1s.

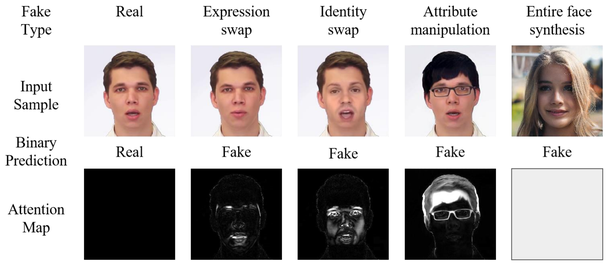


Fig 3: Example images and their corresponding manipulation masks for each of the image types.

Network Architecture:

We develop the attention mechanism using the Xception network, and later show that it also benefits the VGG network. The attention mechanism can be inserted between any inner layers in a CNN.

We design the attention mechanism with three criteria:

✧ Explainable – enhances understanding of the network operations

✧ Useful – enhances the final task of the network

✧ Modular – can be added easily into any network structure

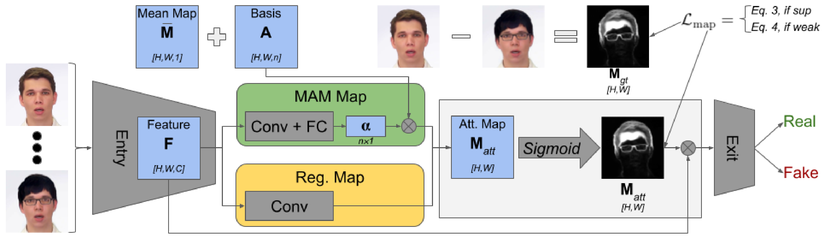


Fig 4: The framework for the proposed Attention Mechanism. We use a convolution layer (and a fully connected layer for MAM) to predict an attention map that highlights the manipulated regions in the face image. This is used to filter the network features before the binary classification task.

Attention Mechanism:

Given a face image, determine the manipulation mask for localization of the manipulated regions.

We estimate the attention map in two ways:

✧ Regression – Directly regress the manipulation mask from network features

✧ Manipulation Appearance Model (MAM) – Regress a vector that linearly combines a set of template masks to produce the manipulation mask.

Fig 5: The map bases that are used in the MAM model. These are produced using PCA on the known GT masks of 100 partial fake images.

Anti-Fake Detection Performance

Problem Statement:

Given an image, determine if it has been digitally altered or synthesized, in whole or in part.

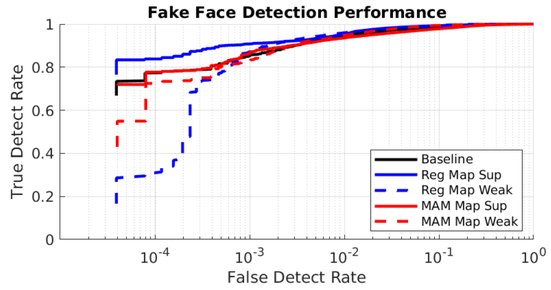


Fig 6: Receiver Operating Characteristic (ROC) curves for the detection of the digitally manipulated face images in the DFFD.

Map Supervision	AUC	EER	$TDR_{0.01\%}$	$TDR_{0.1\%}$	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

Method	Training Data	UADFV	Celeb-DF
Two-Stream	Private	85.1	55.7
MesoInception4	Private	82.1	49.6
HeadPose	UADFV	89.0	54.8
FWA	UADFV	97.4	53.8
Multi-Task	FaceForensics	68.5	36.5
Xception FF++	FaceForensics++	80.4	38.7
Xception + Reg.	DFFD	84.2	64.4
Xception + Reg.	UADFV	98.4	57.1
Xception + Reg.	UADFV, DFFD	98.4	71.2

Tab 1, 2: Anti-fake performance on the DFFD dataset using various types of map and map supervision. Cross dataset testing on other public datasets.

Manipulation Localization Performance

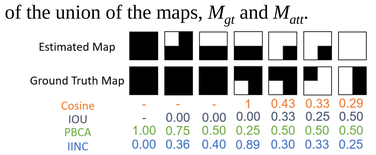
Problem Statement:

Given a face image, determine the manipulation mask for localization of the manipulated regions.

Inverse Intersection Non-Containment Metric:

Let I be the mean of the intersection and U be the mean of the union of the maps, M_{gt} and M_{att}.

$$IINC = \frac{1}{3-U} * \begin{cases} 0, & \text{if } \overline{M_{gt}} = 0 \text{ and } \overline{M_{att}} = 0 \\ 1, & \text{if } \overline{M_{gt}} = 0 \text{ xor } \overline{M_{att}} = 0 \\ 2 - \frac{I}{|M_{gt}|} - \frac{I}{|M_{att}|}, & \text{otherwise} \end{cases}$$



Eq 1 and Fig 7: Calculation of the IINC metric. Toy comparison between the IINC and previous metrics for the evaluation of attention maps.

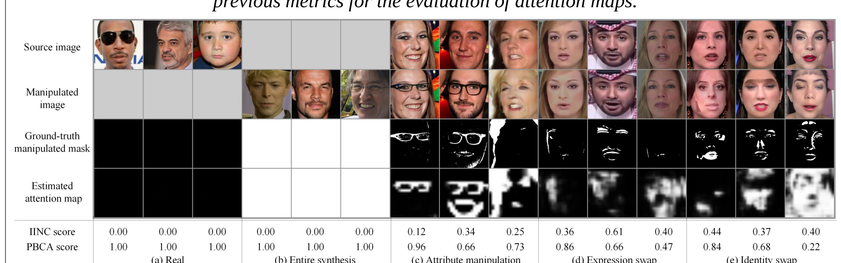


Fig 8: Estimation of the manipulation masks by the Reg. Map Sup. Xception model and their performance in terms of IINC and PBCA.