PRESENTATION ON
DEEPFAKES AND BEYOND: A SURVEY OF FACE MANIPULATION AND FAKE DETECTION
(ARXIV:2001.00179V1)

Pratikkumar.Prajapati@sjsu.edu

Mar/24/2020

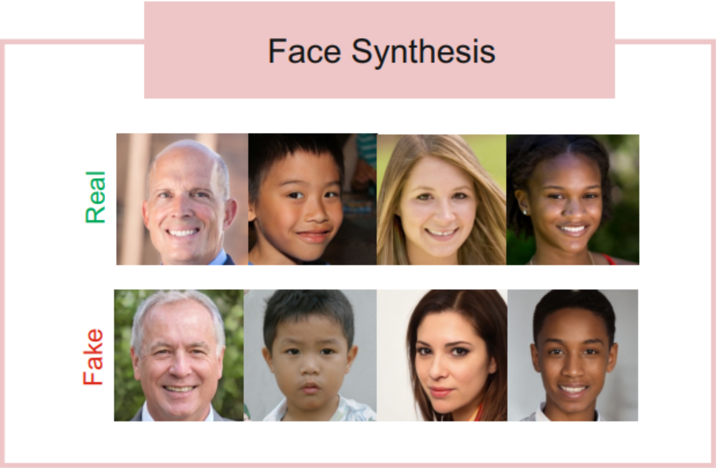
OVERVIEW

- A Comprehensive survey on Face manipulation and detection techniques by Tolosana, Ruben, et al [1]
- Posted on Jan 1 2020 on arxiv.org, (fairly recent paper)
- Covers four main techniques, results, datasets used, and more.
  1. entire face synthesis
  2. face identity swap (DeepFakes)
  3. facial attributes manipulation
  4. facial expression manipulation.

Note: All citation/references are indexed with respect to the original paper

1. ENTIRE FACE SYNTHESIS

- Generate none-existent faces
- Samples from http://www.whichfaceisreal.com/ and https://www.thispersondoesnotexist.com/
- Models
  - ProGAN, StyleGAN, StyleGANv2, SNGAN, CramerGAN, MMDGAN, CycleGAN, Xception Net, Autoencoders
- StyleGAN have achieved astonishing results



FACE SYNTHESIS –
MANIPULATION TECHNIQUES AND PUBLIC DATABASES

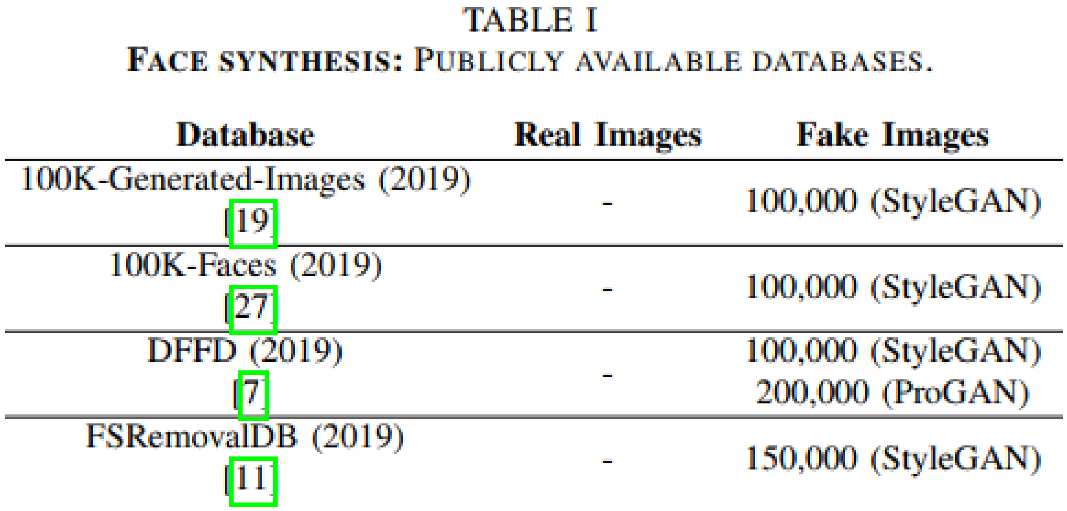


TABLE I
FACE SYNTHESIS: PUBLICLY AVAILABLE DATABASES.

Database	Real Images	Fake Images
100K-Generated-Images (2019) [19]	-	100,000 (StyleGAN)
100K-Faces (2019) [27]	-	100,000 (StyleGAN)
DFFD (2019) [7]	-	100,000 (StyleGAN) 200,000 (ProGAN)
FSRemovalDB (2019) [11]	-	150,000 (StyleGAN)

FACE SYNTHESIS – MANIPULATION DETECTION

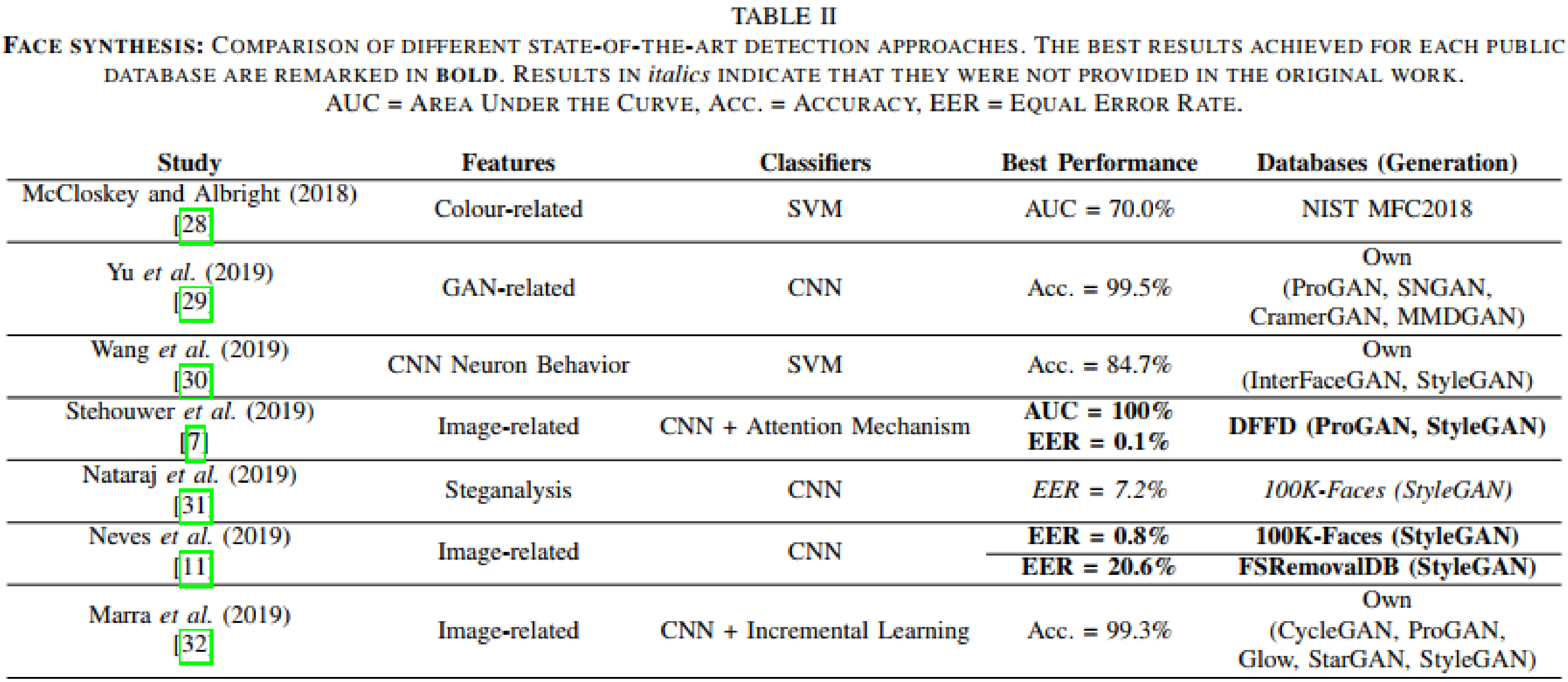


TABLE II

Face synthesis: Comparison of different state-of-the-art detection approaches. The best results achieved for each public database are remarked in **bold**. Results in *italics* indicate that they were not provided in the original work. AUC = Area Under the Curve, Acc. = Accuracy, EER = Equal Error Rate.

Study	Features	Classifiers	Best Performance	Databases (Generation)
McCloskey and Albright (2018) [28]	Colour-related	SVM	AUC = 70.0%	NIST MFC2018
Yu *et al.* (2019) [29]	GAN-related	CNN	Acc. = 99.5%	Own (ProGAN, SNGAN, CramerGAN, MMDGAN)
Wang *et al.* (2019) [30]	CNN Neuron Behavior	SVM	Acc. = 84.7%	Own (InterFaceGAN, StyleGAN)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 100%** **EER = 0.1%**	**DFFD (ProGAN, StyleGAN)**
Nataraj *et al.* (2019) [31]	Steganalysis	CNN	*EER = 7.2%*	*100K-Faces (StyleGAN)*
Neves *et al.* (2019) [11]	Image-related	CNN	**EER = 0.8%**	**100K-Faces (StyleGAN)**
			EER = 20.6%	**FSRemovalDB (StyleGAN)**
Marra *et al.* (2019) [32]	Image-related	CNN + Incremental Learning	Acc. = 99.3%	Own (CycleGAN, ProGAN, Glow, StarGAN, StyleGAN)

FACE SYNTHESIS – MANIPULATION DETECTION

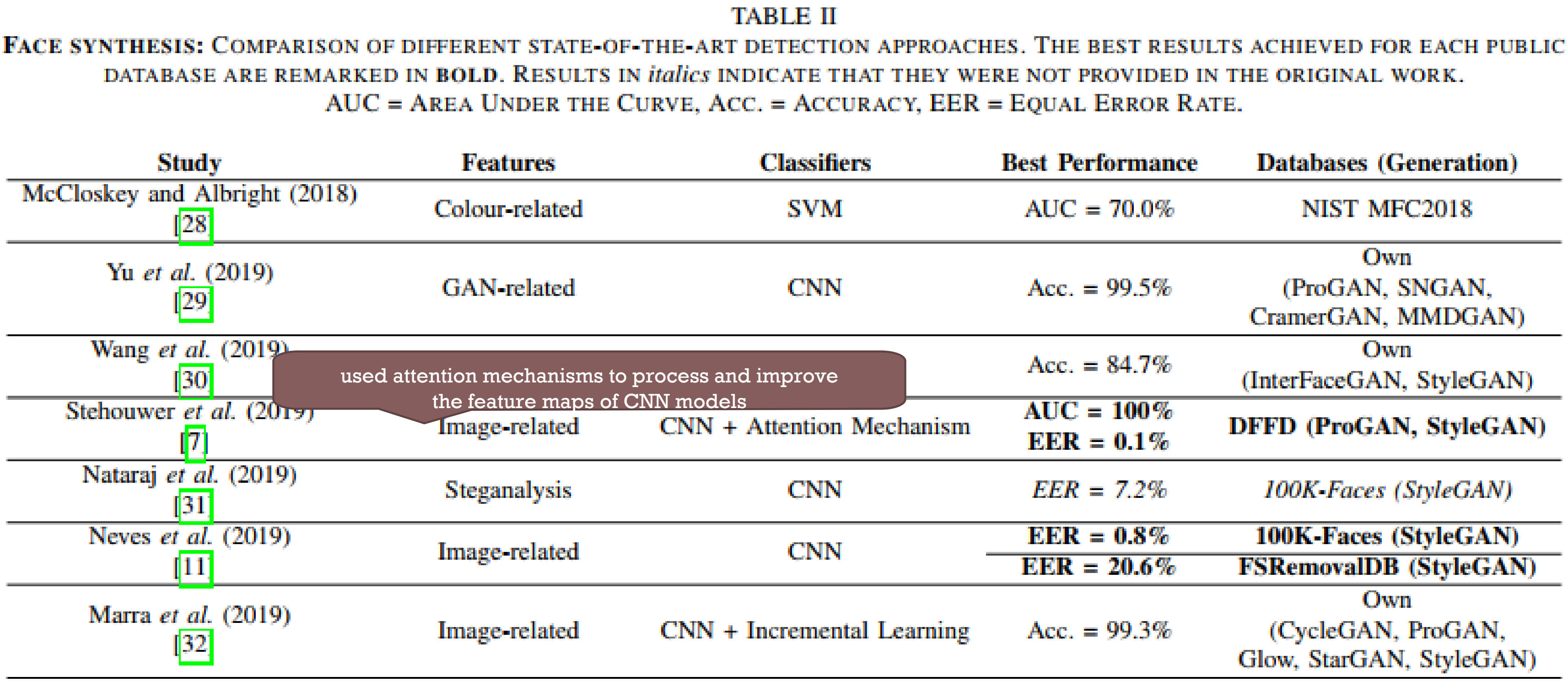


TABLE II

FACE SYNTHESIS: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. RESULTS IN *italics* INDICATE THAT THEY WERE NOT PROVIDED IN THE ORIGINAL WORK. AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases (Generation)
McCloskey and Albright (2018) [28]	Colour-related	SVM	AUC = 70.0%	NIST MFC2018
Yu *et al.* (2019) [29]	GAN-related	CNN	Acc. = 99.5%	Own (ProGAN, SNGAN, CramerGAN, MMDGAN)
Wang *et al.* (2019) [30]			Acc. = 84.7%	Own (InterFaceGAN, StyleGAN)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 100%** **EER = 0.1%**	**DFFD (ProGAN, StyleGAN)**
Nataraj *et al.* (2019) [31]	Steganalysis	CNN	*EER = 7.2%*	*100K-Faces (StyleGAN)*
Neves *et al.* (2019) [11]	Image-related	CNN	**EER = 0.8%**	**100K-Faces (StyleGAN)**
			EER = 20.6%	**FSRemovalDB (StyleGAN)**
Marra *et al.* (2019) [32]	Image-related	CNN + Incremental Learning	Acc. = 99.3%	Own (CycleGAN, ProGAN, Glow, StarGAN, StyleGAN)

used attention mechanisms to process and improve the feature maps of CNN models

FACE SYNTHESIS – MANIPULATION DETECTION

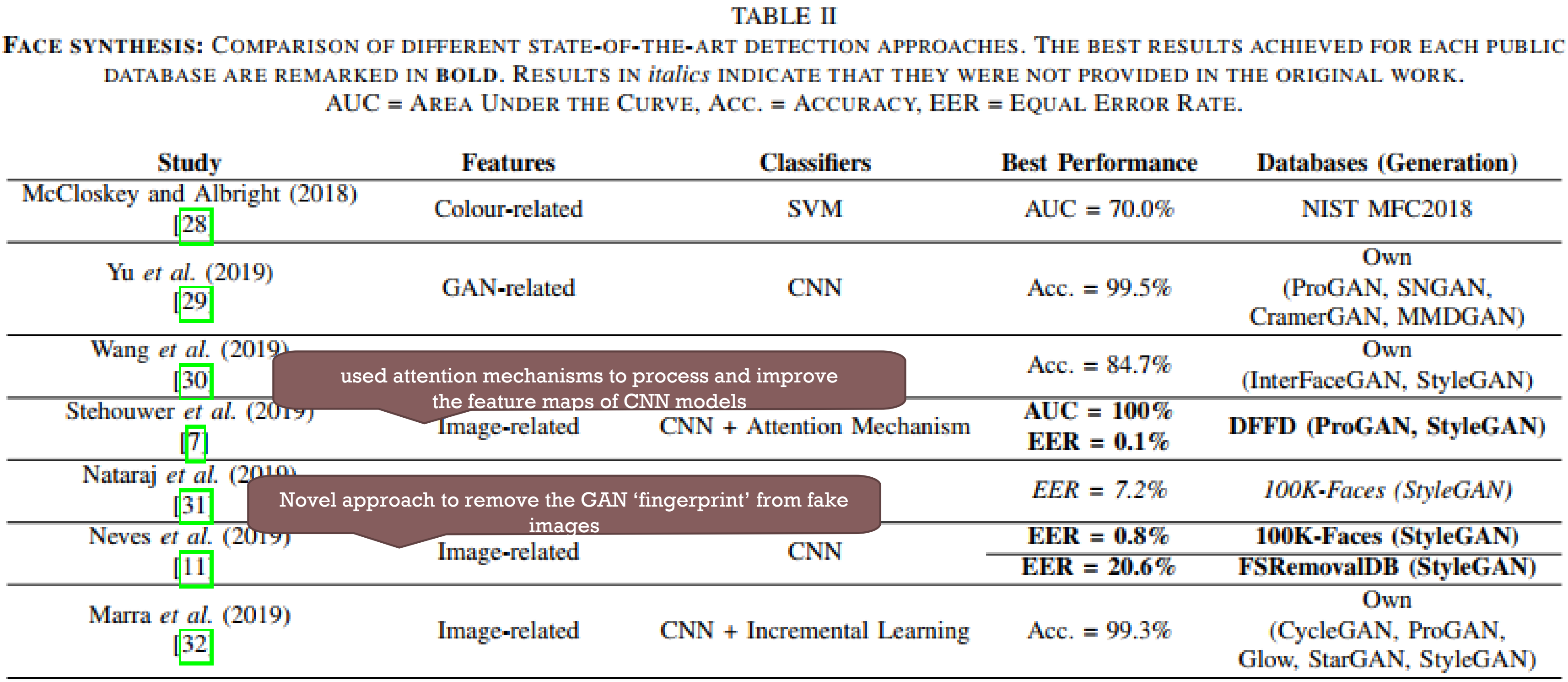


TABLE II

FACE SYNTHESIS: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. RESULTS IN *italics* INDICATE THAT THEY WERE NOT PROVIDED IN THE ORIGINAL WORK. AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

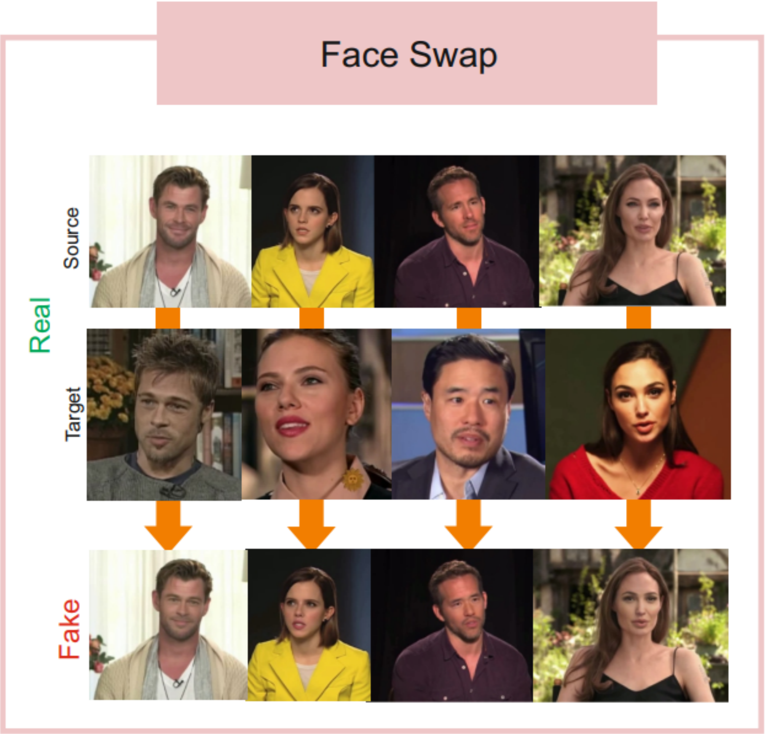
Study	Features	Classifiers	Best Performance	Databases (Generation)
McCloskey and Albright (2018) [28]	Colour-related	SVM	AUC = 70.0%	NIST MFC2018
Yu *et al.* (2019) [29]	GAN-related	CNN	Acc. = 99.5%	Own (ProGAN, SNGAN, CramerGAN, MMDGAN)
Wang *et al.* (2019) [30]	[illegible]	[illegible]	Acc. = 84.7%	Own (InterFaceGAN, StyleGAN)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 100%** **EER = 0.1%**	**DFFD (ProGAN, StyleGAN)**
Nataraj *et al.* (2019) [31]	[illegible]	[illegible]	*EER = 7.2%*	*100K-Faces (StyleGAN)*
Neves *et al.* (2019) [11]	Image-related	CNN	**EER = 0.8%**	**100K-Faces (StyleGAN)**
			EER = 20.6%	**FSRemovalDB (StyleGAN)**
Marra *et al.* (2019) [32]	Image-related	CNN + Incremental Learning	Acc. = 99.3%	Own (CycleGAN, ProGAN, Glow, StarGAN, StyleGAN)

used attention mechanisms to process and improve the feature maps of CNN models

Novel approach to remove the GAN 'fingerprint' from fake images

2. FACE SWAP (DEEPFAKES)

- Replace face of one person with another
- Two main methods
  - Classical computer graphics-based techniques e.g. FaceSwap App
  - novel deep learning techniques known as DeepFakes e.g. ZAO App
  - E.g. https://www.youtube.com/watch?v=UlvoEW7l5rs
- Models
  - FaceSwapGAN, CycleGAN, FaceNet, Autoencoders, CNN SVM etc



FACE SWAP –
MANIPULATION TECHNIQUES AND PUBLIC DATABASES

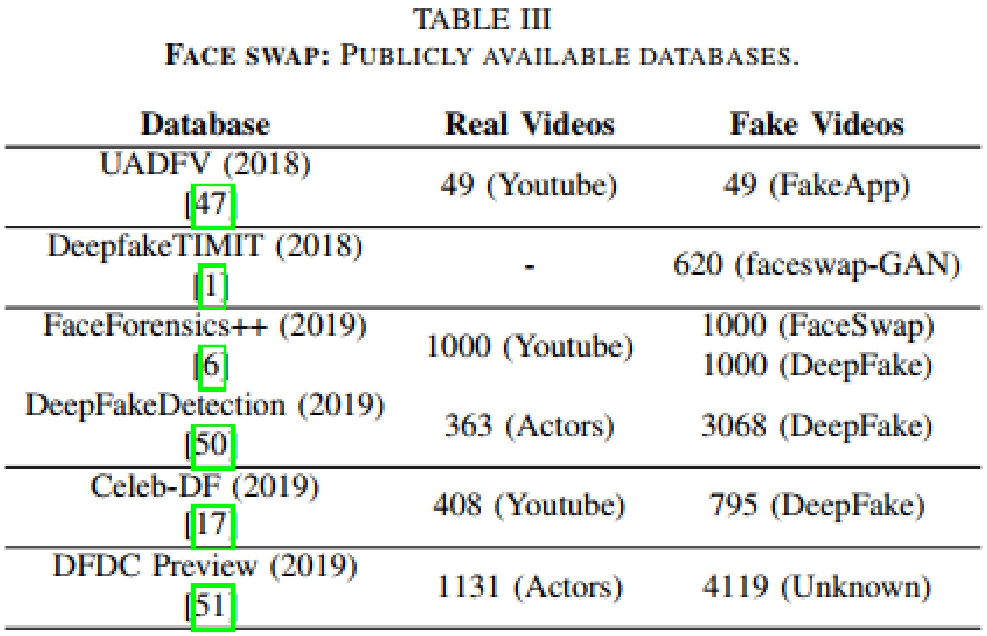


TABLE III
FACE SWAP: PUBLICLY AVAILABLE DATABASES.

Database	Real Videos	Fake Videos
UADFV (2018) [47]	49 (Youtube)	49 (FakeApp)
DeepfakeTIMIT (2018) [1]	-	620 (faceswap-GAN)
FaceForensics++ (2019) [6]	1000 (Youtube)	1000 (FaceSwap) 1000 (DeepFake)
DeepFakeDetection (2019) [50]	363 (Actors)	3068 (DeepFake)
Celeb-DF (2019) [17]	408 (Youtube)	795 (DeepFake)
DFDC Preview (2019) [51]	1131 (Actors)	4119 (Unknown)

FACE SWAP – MANIPULATION DETECTION

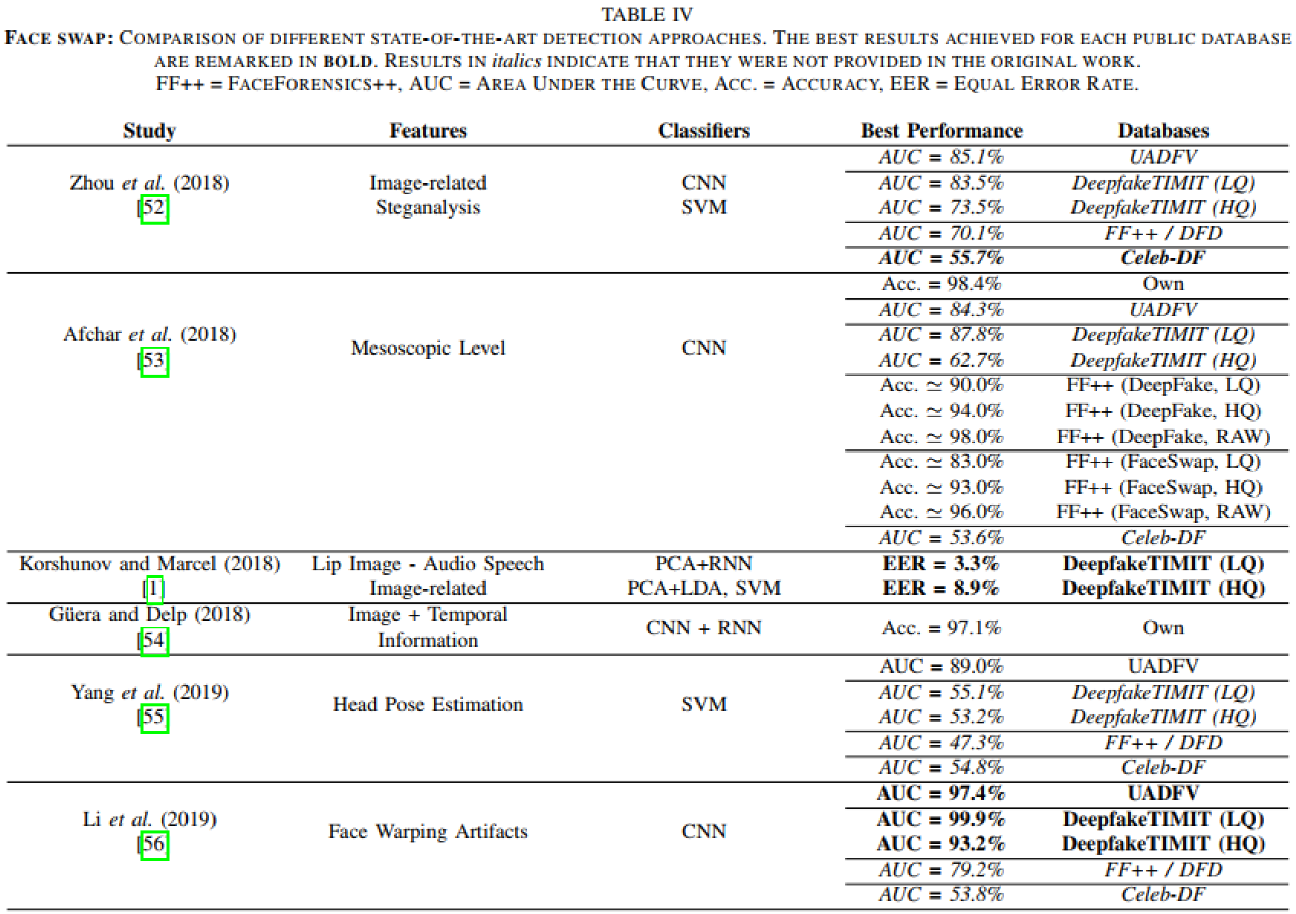


TABLE IV

FACE SWAP: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. RESULTS IN *italics* INDICATE THAT THEY WERE NOT PROVIDED IN THE ORIGINAL WORK. FF++ = FACEFORENSICS++, AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases
Zhou *et al.* (2018) [52]	Image-related Steganalysis	CNN SVM	*AUC = 85.1%*	*UADFV*
			AUC = 83.5%	*DeepfakeTIMIT (LQ)*
			AUC = 73.5%	*DeepfakeTIMIT (HQ)*
			AUC = 70.1%	*FF++ / DFD*
			AUC = 55.7%	***Celeb-DF***
Afchar *et al.* (2018) [53]	Mesoscopic Level	CNN	Acc. = 98.4%	Own
			AUC = 84.3%	*UADFV*
			AUC = 87.8%	*DeepfakeTIMIT (LQ)*
			AUC = 62.7%	*DeepfakeTIMIT (HQ)*
			Acc. ≃ 90.0%	FF++ (DeepFake, LQ)
			Acc. ≃ 94.0%	FF++ (DeepFake, HQ)
			Acc. ≃ 98.0%	FF++ (DeepFake, RAW)
			Acc. ≃ 83.0%	FF++ (FaceSwap, LQ)
			Acc. ≃ 93.0%	FF++ (FaceSwap, HQ)
			Acc. ≃ 96.0%	FF++ (FaceSwap, RAW)
			AUC = 53.6%	*Celeb-DF*
Korshunov and Marcel (2018) [1]	Lip Image - Audio Speech	PCA+RNN	**EER = 3.3%**	**DeepfakeTIMIT (LQ)**
	Image-related	PCA+LDA, SVM	**EER = 8.9%**	**DeepfakeTIMIT (HQ)**
Güera and Delp (2018) [54]	Image + Temporal Information	CNN + RNN	Acc. = 97.1%	Own
Yang *et al.* (2019) [55]	Head Pose Estimation	SVM	AUC = 89.0%	UADFV
			AUC = 55.1%	*DeepfakeTIMIT (LQ)*
			AUC = 53.2%	*DeepfakeTIMIT (HQ)*
			AUC = 47.3%	*FF++ / DFD*
			AUC = 54.8%	*Celeb-DF*
Li *et al.* (2019) [56]	Face Warping Artifacts	CNN	**AUC = 97.4%**	**UADFV**
			AUC = 99.9%	**DeepfakeTIMIT (LQ)**
			AUC = 93.2%	**DeepfakeTIMIT (HQ)**
			AUC = 79.2%	*FF++ / DFD*
			AUC = 53.8%	*Celeb-DF*

FACE SWAP – MANIPULATION DETECTION

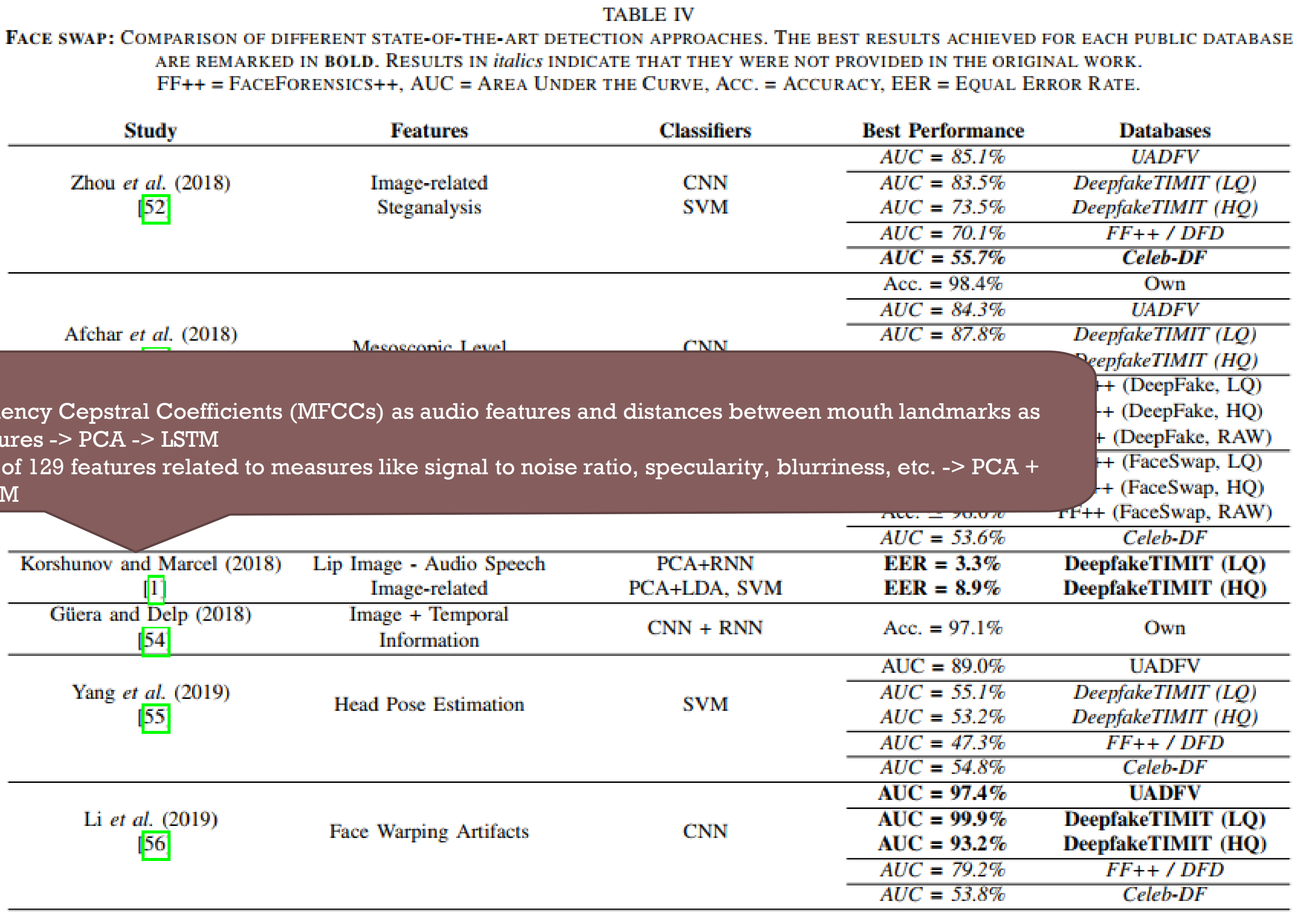


TABLE IV

FACE SWAP: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. RESULTS IN *italics* INDICATE THAT THEY WERE NOT PROVIDED IN THE ORIGINAL WORK. FF++ = FACEFORENSICS++, AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases
Zhou *et al.* (2018) [52]	Image-related Steganalysis	CNN SVM	*AUC = 85.1%*	*UADFV*
			AUC = 83.5% *AUC = 73.5%*	*DeepfakeTIMIT (LQ)* *DeepfakeTIMIT (HQ)*
			AUC = 70.1%	*FF++ / DFD*
			AUC = 55.7%	***Celeb-DF***
Afchar *et al.* (2018)	Mesoscopic Level	CNN	Acc. = 98.4%	Own
			AUC = 84.3%	*UADFV*
			AUC = 87.8% [illegible]	*DeepfakeTIMIT (LQ)* *eepfakeTIMIT (HQ)*
			[illegible] [illegible] [illegible]	++ (DeepFake, LQ) + (DeepFake, HQ) + (DeepFake, RAW)
			[illegible] [illegible] [illegible]	++ (FaceSwap, LQ) + (FaceSwap, HQ) FF++ (FaceSwap, RAW)
			AUC = 53.6%	*Celeb-DF*
Korshunov and Marcel (2018) [1]	Lip Image - Audio Speech Image-related	PCA+RNN PCA+LDA, SVM	**EER = 3.3%** **EER = 8.9%**	**DeepfakeTIMIT (LQ)** **DeepfakeTIMIT (HQ)**
Güera and Delp (2018) [54]	Image + Temporal Information	CNN + RNN	Acc. = 97.1%	Own
Yang *et al.* (2019) [55]	Head Pose Estimation	SVM	AUC = 89.0%	UADFV
			AUC = 55.1% *AUC = 53.2%*	*DeepfakeTIMIT (LQ)* *DeepfakeTIMIT (HQ)*
			AUC = 47.3%	*FF++ / DFD*
			AUC = 54.8%	*Celeb-DF*
Li *et al.* (2019) [56]	Face Warping Artifacts	CNN	**AUC = 97.4%**	**UADFV**
			AUC = 99.9% **AUC = 93.2%**	**DeepfakeTIMIT (LQ)** **DeepfakeTIMIT (HQ)**
			AUC = 79.2%	*FF++ / DFD*
			AUC = 53.8%	*Celeb-DF*

1. Mel-Frequency Cepstral Coefficients (MFCCs) as audio features and distances between mouth landmarks as visual features -> PCA -> LSTM
2. used a set of 129 features related to measures like signal to noise ratio, specularity, blurriness, etc. -> PCA + LDA -> SVM

FACE SWAP – MANIPULATION DETECTION

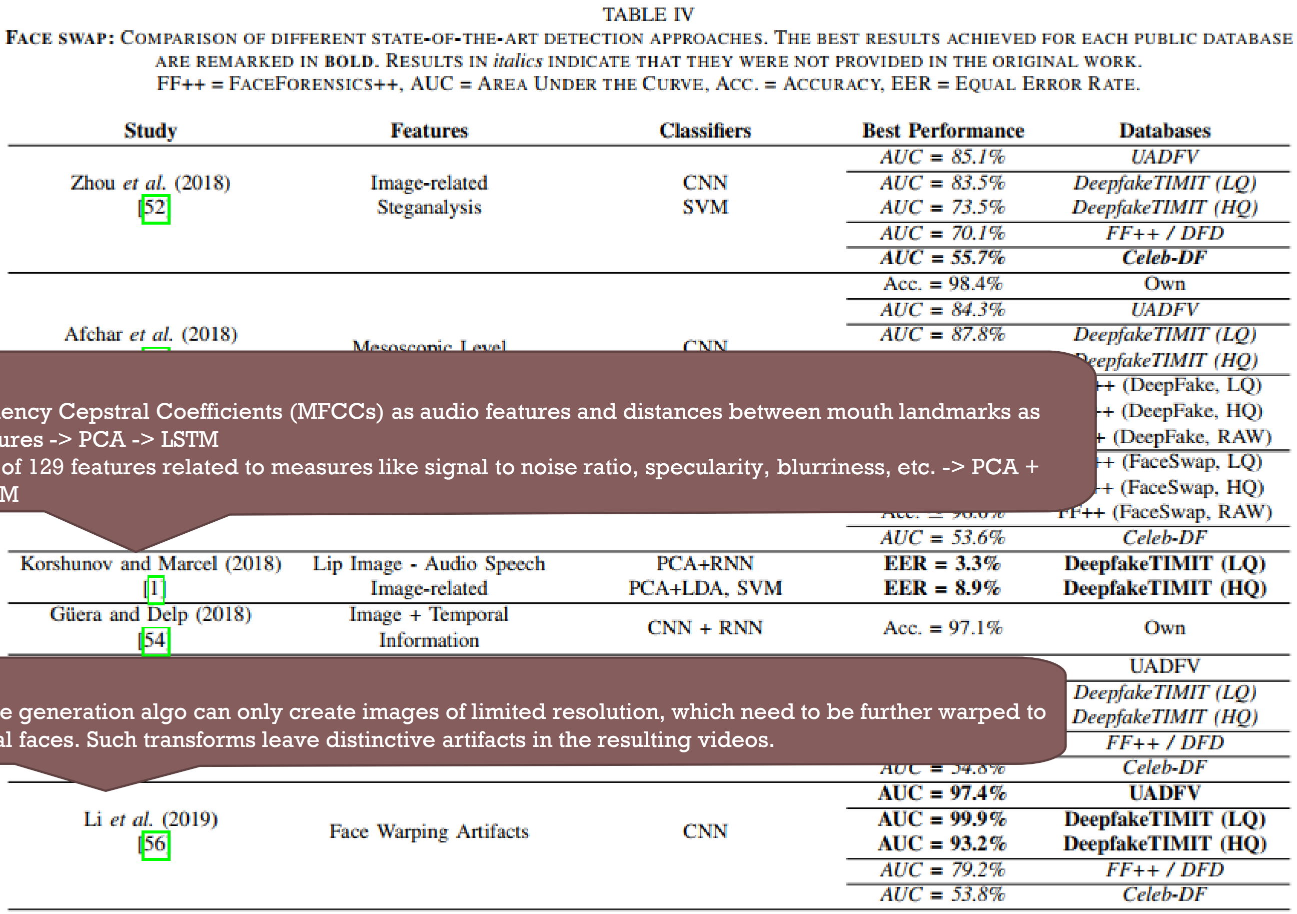


TABLE IV

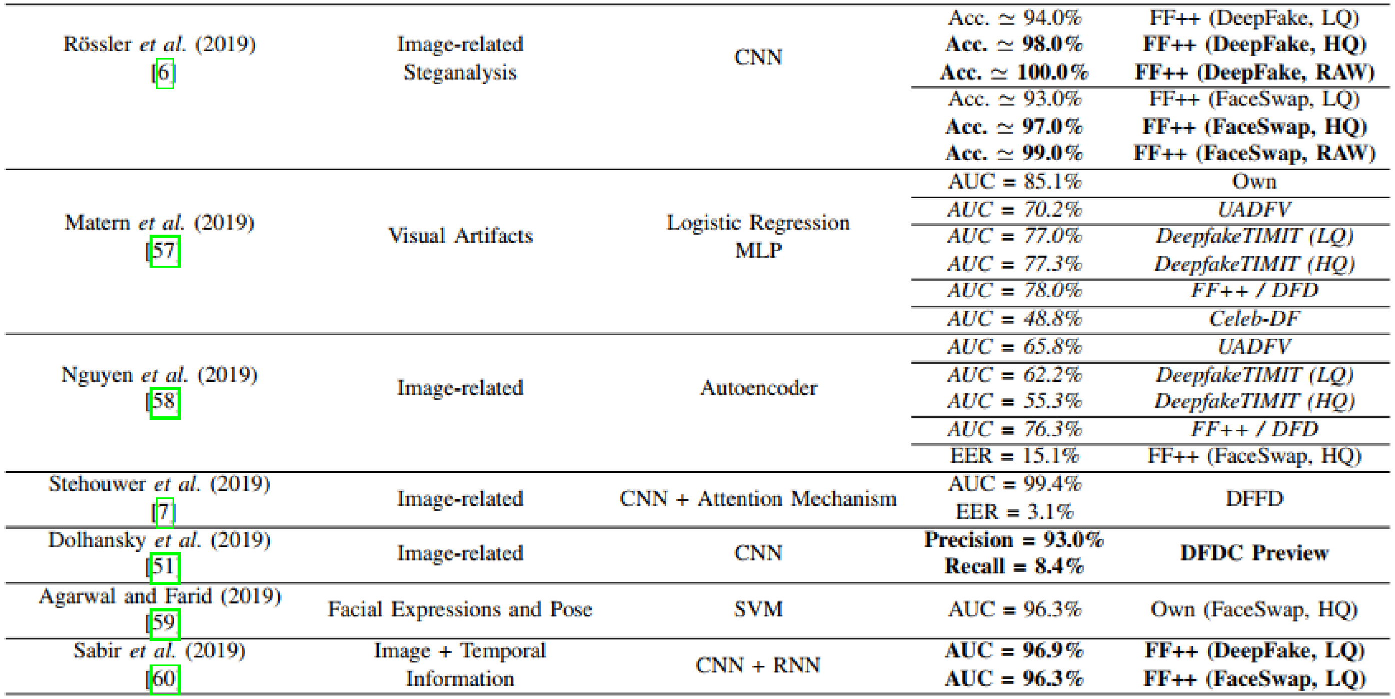
FACE SWAP: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. RESULTS IN *italics* INDICATE THAT THEY WERE NOT PROVIDED IN THE ORIGINAL WORK. FF++ = FACEFORENSICS++, AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases
Zhou *et al.* (2018) [52]	Image-related Steganalysis	CNN SVM	*AUC = 85.1%*	*UADFV*
			AUC = 83.5%	*DeepfakeTIMIT (LQ)*
			AUC = 73.5%	*DeepfakeTIMIT (HQ)*
			AUC = 70.1%	*FF++ / DFD*
			AUC = 55.7%	***Celeb-DF***
Afchar *et al.* (2018)	Mesoscopic Level	CNN	Acc. = 98.4%	Own
			AUC = 84.3%	*UADFV*
			AUC = 87.8%	*DeepfakeTIMIT (LQ)*
				DeepfakeTIMIT (HQ)
				FF++ (DeepFake, LQ)
				FF++ (DeepFake, HQ)
				FF++ (DeepFake, RAW)
				FF++ (FaceSwap, LQ)
				FF++ (FaceSwap, HQ)
			Acc. = [illegible]	FF++ (FaceSwap, RAW)
			AUC = 53.6%	*Celeb-DF*
Korshunov and Marcel (2018) [1]	Lip Image - Audio Speech Image-related	PCA+RNN PCA+LDA, SVM	**EER = 3.3%** **EER = 8.9%**	**DeepfakeTIMIT (LQ)** **DeepfakeTIMIT (HQ)**
Güera and Delp (2018) [54]	Image + Temporal Information	CNN + RNN	Acc. = 97.1%	Own
				UADFV
				DeepfakeTIMIT (LQ)
				DeepfakeTIMIT (HQ)
				FF++ / DFD
			AUC = 54.8%	*Celeb-DF*
Li *et al.* (2019) [56]	Face Warping Artifacts	CNN	**AUC = 97.4%**	**UADFV**
			AUC = 99.9%	**DeepfakeTIMIT (LQ)**
			AUC = 93.2%	**DeepfakeTIMIT (HQ)**
			AUC = 79.2%	*FF++ / DFD*
			AUC = 53.8%	*Celeb-DF*

1. Mel-Frequency Cepstral Coefficients (MFCCs) as audio features and distances between mouth landmarks as visual features -> PCA -> LSTM
2. used a set of 129 features related to measures like signal to noise ratio, specularity, blurriness, etc. -> PCA + LDA -> SVM

current DeepFake generation algo can only create images of limited resolution, which need to be further warped to match the original faces. Such transforms leave distinctive artifacts in the resulting videos.

FACE SWAP – MANIPULATION DETECTION (TABLE: IV CONT.)



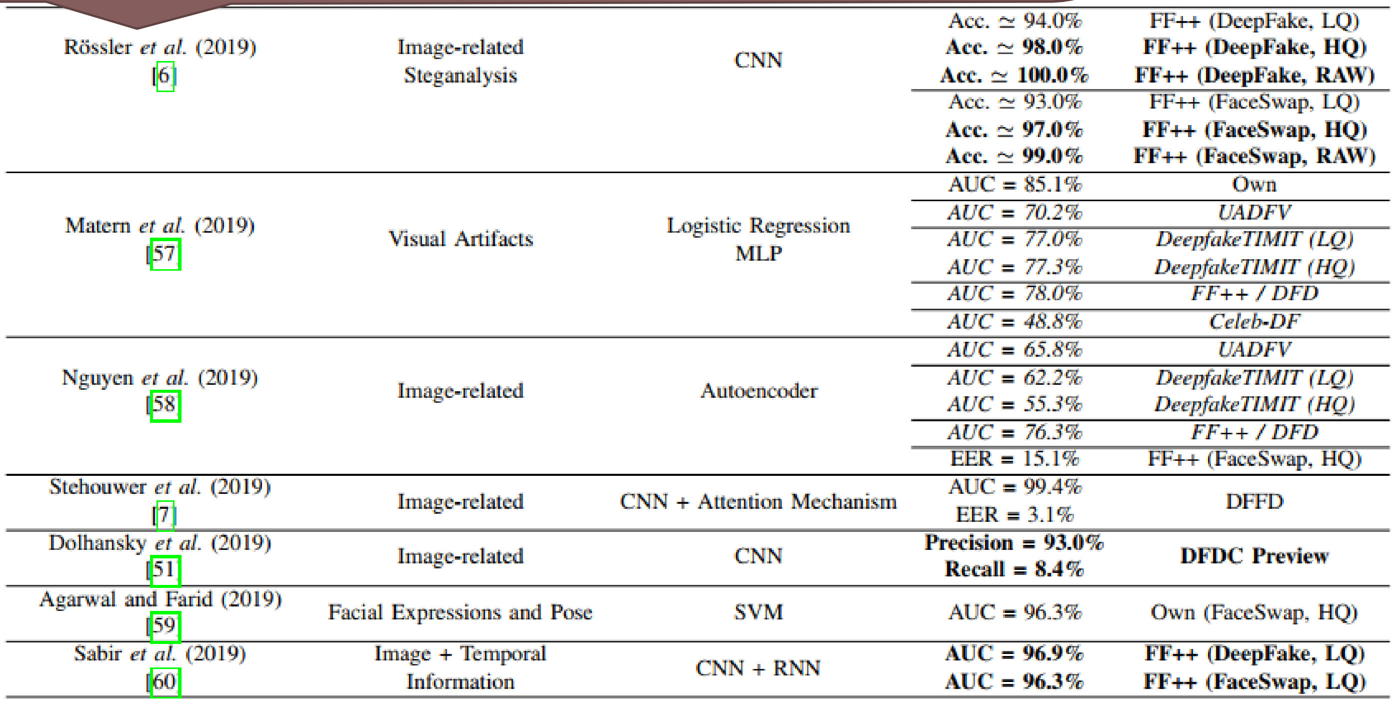
Rössler *et al.* (2019) [6]	Image-related Steganalysis	CNN	Acc. ≃ 94.0%	FF++ (DeepFake, LQ)
			Acc. ≃ 98.0%	**FF++ (DeepFake, HQ)**
			Acc. ≃ 100.0%	**FF++ (DeepFake, RAW)**
			Acc. ≃ 93.0%	FF++ (FaceSwap, LQ)
			Acc. ≃ 97.0%	**FF++ (FaceSwap, HQ)**
			Acc. ≃ 99.0%	**FF++ (FaceSwap, RAW)**
Matern *et al.* (2019) [57]	Visual Artifacts	Logistic Regression MLP	AUC = 85.1%	Own
			AUC = 70.2%	*UADFV*
			AUC = 77.0%	*DeepfakeTIMIT (LQ)*
			AUC = 77.3%	*DeepfakeTIMIT (HQ)*
			AUC = 78.0%	*FF++ / DFD*
			AUC = 48.8%	*Celeb-DF*
Nguyen *et al.* (2019) [58]	Image-related	Autoencoder	*AUC = 65.8%*	*UADFV*
			AUC = 62.2%	*DeepfakeTIMIT (LQ)*
			AUC = 55.3%	*DeepfakeTIMIT (HQ)*
			AUC = 76.3%	*FF++ / DFD*
			EER = 15.1%	FF++ (FaceSwap, HQ)

Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	AUC = 99.4%
			EER = 3.1%
				DFFD

Dolhansky *et al.* (2019) [51]	Image-related	CNN	**Precision = 93.0%** **Recall = 8.4%**	**DFDC Preview**
Agarwal and Farid (2019) [59]	Facial Expressions and Pose	SVM	AUC = 96.3%	Own (FaceSwap, HQ)
Sabir *et al.* (2019) [60]	Image + Temporal Information	CNN + RNN	**AUC = 96.9%**	**FF++ (DeepFake, LQ)**
			AUC = 96.3%	**FF++ (FaceSwap, LQ)**

FACE SWAP – MANIPULATION DETECTION (TABLE: IV CONT.)

Evaluated four different detection systems. The best one was using Xception Net pretrained with ImageNet Dataset and then re-trained for Fake datasets. Lower accuracy on Low-Quality samples.



Rössler *et al.* (2019) [6]	Image-related Steganalysis	CNN	Acc. ≃ 94.0% **Acc. ≃ 98.0%** **Acc. ≃ 100.0%**	FF++ (DeepFake, LQ) **FF++ (DeepFake, HQ)** **FF++ (DeepFake, RAW)**
			Acc. ≃ 93.0% **Acc. ≃ 97.0%** **Acc. ≃ 99.0%**	FF++ (FaceSwap, LQ) **FF++ (FaceSwap, HQ)** **FF++ (FaceSwap, RAW)**
Matern *et al.* (2019) [57]	Visual Artifacts	Logistic Regression MLP	AUC = 85.1%	Own
			AUC = 70.2%	*UADFV*
			AUC = 77.0% *AUC = 77.3%*	*DeepfakeTIMIT (LQ)* *DeepfakeTIMIT (HQ)*
			AUC = 78.0%	*FF++ / DFD*
			AUC = 48.8%	*Celeb-DF*
Nguyen *et al.* (2019) [58]	Image-related	Autoencoder	*AUC = 65.8%*	*UADFV*
			AUC = 62.2% *AUC = 55.3%*	*DeepfakeTIMIT (LQ)* *DeepfakeTIMIT (HQ)*
			AUC = 76.3%	*FF++ / DFD*
			EER = 15.1%	FF++ (FaceSwap, HQ)

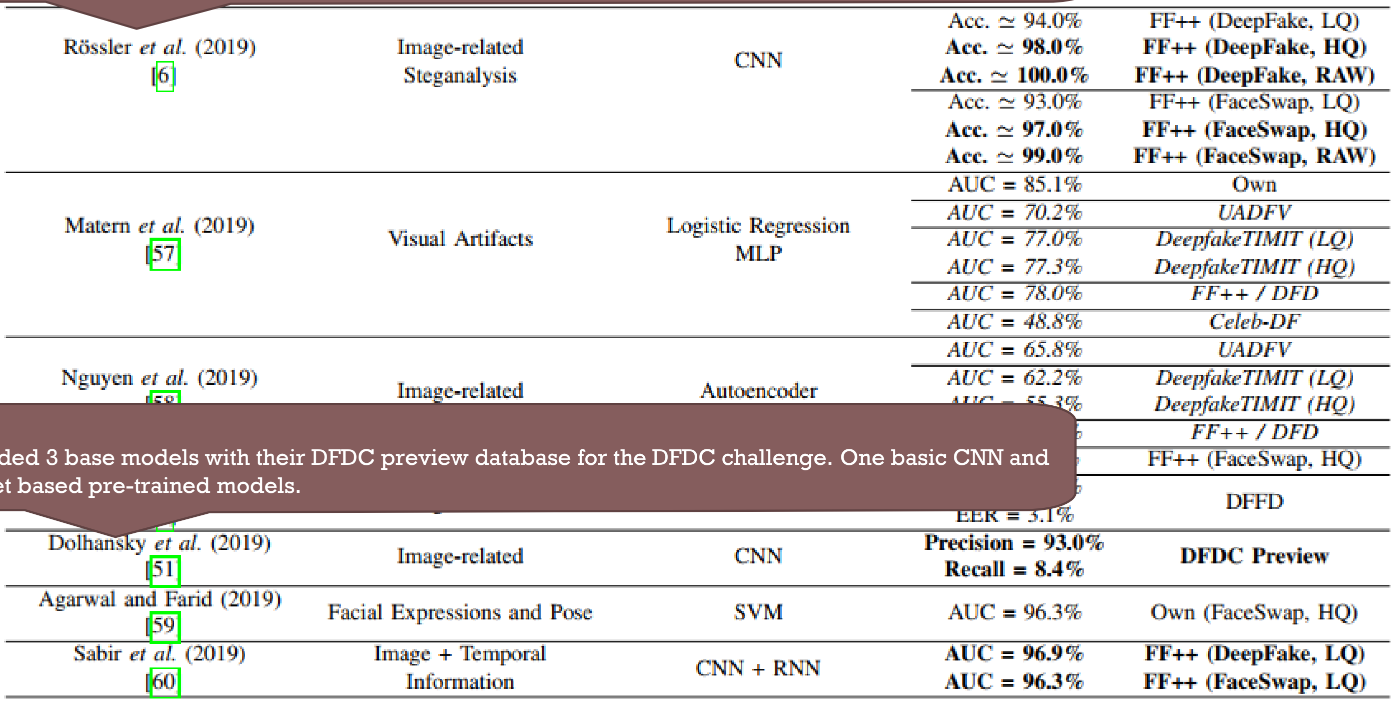
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	AUC = 99.4%
			EER = 3.1%
				DFFD

Dolhansky *et al.* (2019) [51]	Image-related	CNN	**Precision = 93.0%** **Recall = 8.4%**	**DFDC Preview**
Agarwal and Farid (2019) [59]	Facial Expressions and Pose	SVM	AUC = 96.3%	Own (FaceSwap, HQ)
Sabir *et al.* (2019) [60]	Image + Temporal Information	CNN + RNN	**AUC = 96.9%** **AUC = 96.3%**	**FF++ (DeepFake, LQ)** **FF++ (FaceSwap, LQ)**

FACE SWAP – MANIPULATION DETECTION (TABLE: IV CONT.)

Evaluated four different detection systems. The best one was using Xception Net pretrained with ImageNet Dataset and then re-trained for Fake datasets. Lower accuracy on Low-Quality samples.

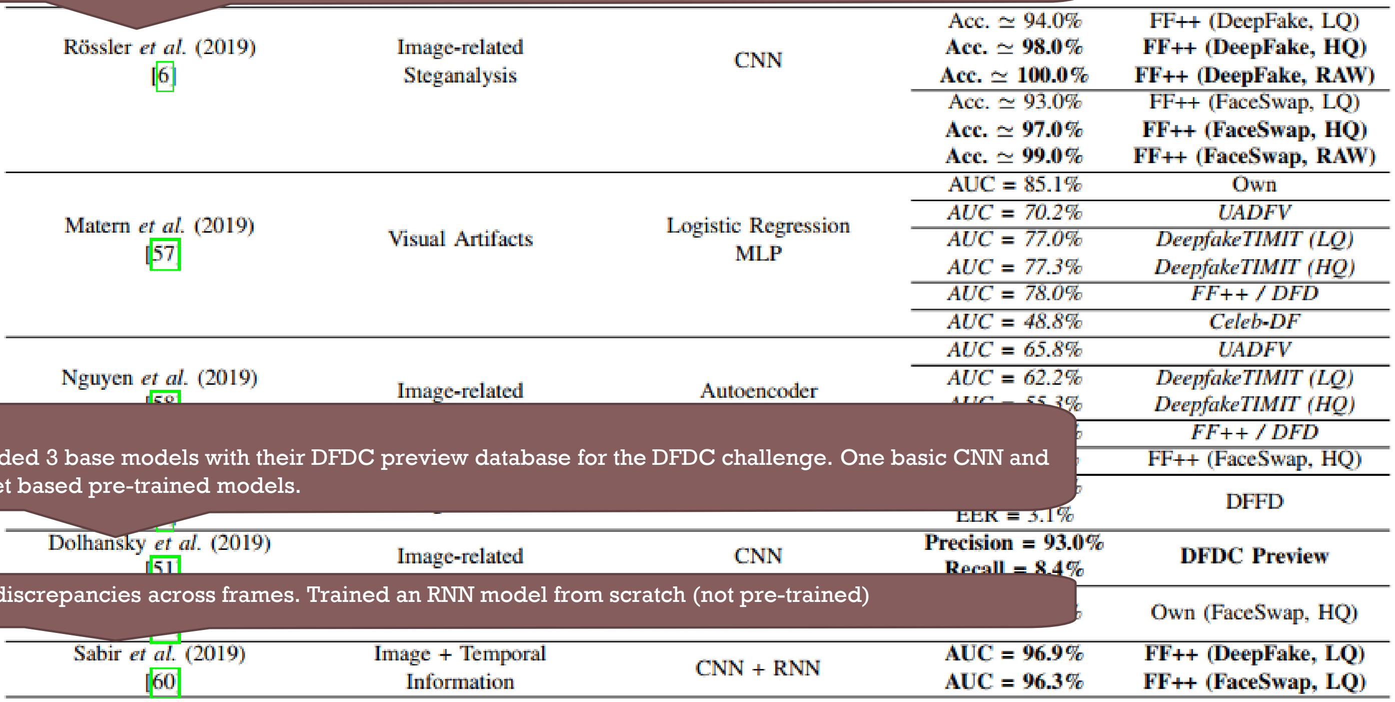
Facebook provided 3 base models with their DFDC preview database for the DFDC challenge. One basic CNN and two Xception Net based pre-trained models.



Rössler *et al.* (2019) [6]	Image-related Steganalysis	CNN	Acc. ≃ 94.0%	FF++ (DeepFake, LQ)
			Acc. ≃ 98.0%	**FF++ (DeepFake, HQ)**
			Acc. ≃ 100.0%	**FF++ (DeepFake, RAW)**
			Acc. ≃ 93.0%	FF++ (FaceSwap, LQ)
			Acc. ≃ 97.0%	**FF++ (FaceSwap, HQ)**
			Acc. ≃ 99.0%	**FF++ (FaceSwap, RAW)**
Matern *et al.* (2019) [57]	Visual Artifacts	Logistic Regression MLP	AUC = 85.1%	Own
			AUC = 70.2%	*UADFV*
			AUC = 77.0%	*DeepfakeTIMIT (LQ)*
			AUC = 77.3%	*DeepfakeTIMIT (HQ)*
			AUC = 78.0%	*FF++ / DFD*
			AUC = 48.8%	*Celeb-DF*
Nguyen *et al.* (2019) [58]	Image-related	Autoencoder	*AUC = 65.8%*	*UADFV*
			AUC = 62.2%	*DeepfakeTIMIT (LQ)*
			AUC = 55.3%	*DeepfakeTIMIT (HQ)*
			[illegible]	*FF++ / DFD*
			[illegible]	FF++ (FaceSwap, HQ)
[illegible]	[illegible]	[illegible]	[illegible] EER = 3.1%	DFFD
Dolhansky *et al.* (2019) [51]	Image-related	CNN	**Precision = 93.0%** **Recall = 8.4%**	**DFDC Preview**
Agarwal and Farid (2019) [59]	Facial Expressions and Pose	SVM	AUC = 96.3%	Own (FaceSwap, HQ)
Sabir *et al.* (2019) [60]	Image + Temporal Information	CNN + RNN	**AUC = 96.9%**	**FF++ (DeepFake, LQ)**
			AUC = 96.3%	**FF++ (FaceSwap, LQ)**

FACE SWAP – MANIPULATION DETECTION (TABLE: IV CONT.)

Evaluated four different detection systems. The best one was using Xception Net pretrained with ImageNet Dataset and then re-trained for Fake datasets. Lower accuracy on Low-Quality samples.



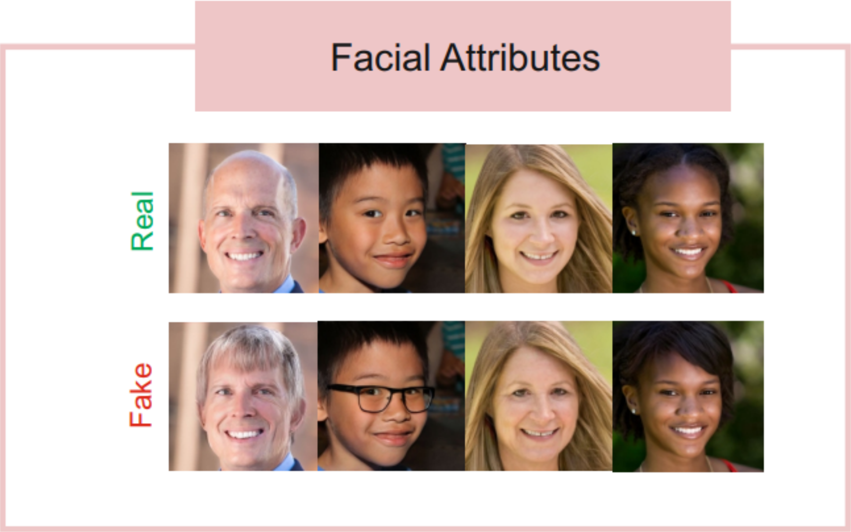
Rössler *et al.* (2019) [6]	Image-related Steganalysis	CNN	Acc. ≃ 94.0%	FF++ (DeepFake, LQ)
			Acc. ≃ 98.0%	**FF++ (DeepFake, HQ)**
			Acc. ≃ 100.0%	**FF++ (DeepFake, RAW)**
			Acc. ≃ 93.0%	FF++ (FaceSwap, LQ)
			Acc. ≃ 97.0%	**FF++ (FaceSwap, HQ)**
			Acc. ≃ 99.0%	**FF++ (FaceSwap, RAW)**
Matern *et al.* (2019) [57]	Visual Artifacts	Logistic Regression MLP	AUC = 85.1%	Own
			AUC = 70.2%	*UADFV*
			AUC = 77.0%	*DeepfakeTIMIT (LQ)*
			AUC = 77.3%	*DeepfakeTIMIT (HQ)*
			AUC = 78.0%	*FF++ / DFD*
			AUC = 48.8%	*Celeb-DF*
Nguyen *et al.* (2019) [58]	Image-related	Autoencoder	*AUC = 65.8%*	*UADFV*
			AUC = 62.2%	*DeepfakeTIMIT (LQ)*
			AUC = 55.3%	*DeepfakeTIMIT (HQ)*
				FF++ / DFD
				FF++ (FaceSwap, HQ)
			EER = 3.1%	DFFD
Dolhansky *et al.* (2019) [51]	Image-related	CNN	**Precision = 93.0%** **Recall = 8.4%**	**DFDC Preview**
				Own (FaceSwap, HQ)
Sabir *et al.* (2019) [60]	Image + Temporal Information	CNN + RNN	**AUC = 96.9%**	**FF++ (DeepFake, LQ)**
			AUC = 96.3%	**FF++ (FaceSwap, LQ)**

Facebook provided 3 base models with their DFDC preview database for the DFDC challenge. One basic CNN and two Xception Net based pre-trained models.

Used temporal discrepancies across frames. Trained an RNN model from scratch (not pre-trained)

3. FACIAL ATTRIBUTES MANIPULATION

- Modify some attributes of the face such as the color of the hair or the skin, the gender, the age, adding glasses, etc.
- Models
  - StarGAN, IcGANs , cGAN, Autoencoders, attGAN, STGAN,
- FaceApp mobile application
- Public Database
  - Diverse Fake Face Dataset (DFFD)



FACIAL ATTRIBUTES – MANIPULATION DETECTION

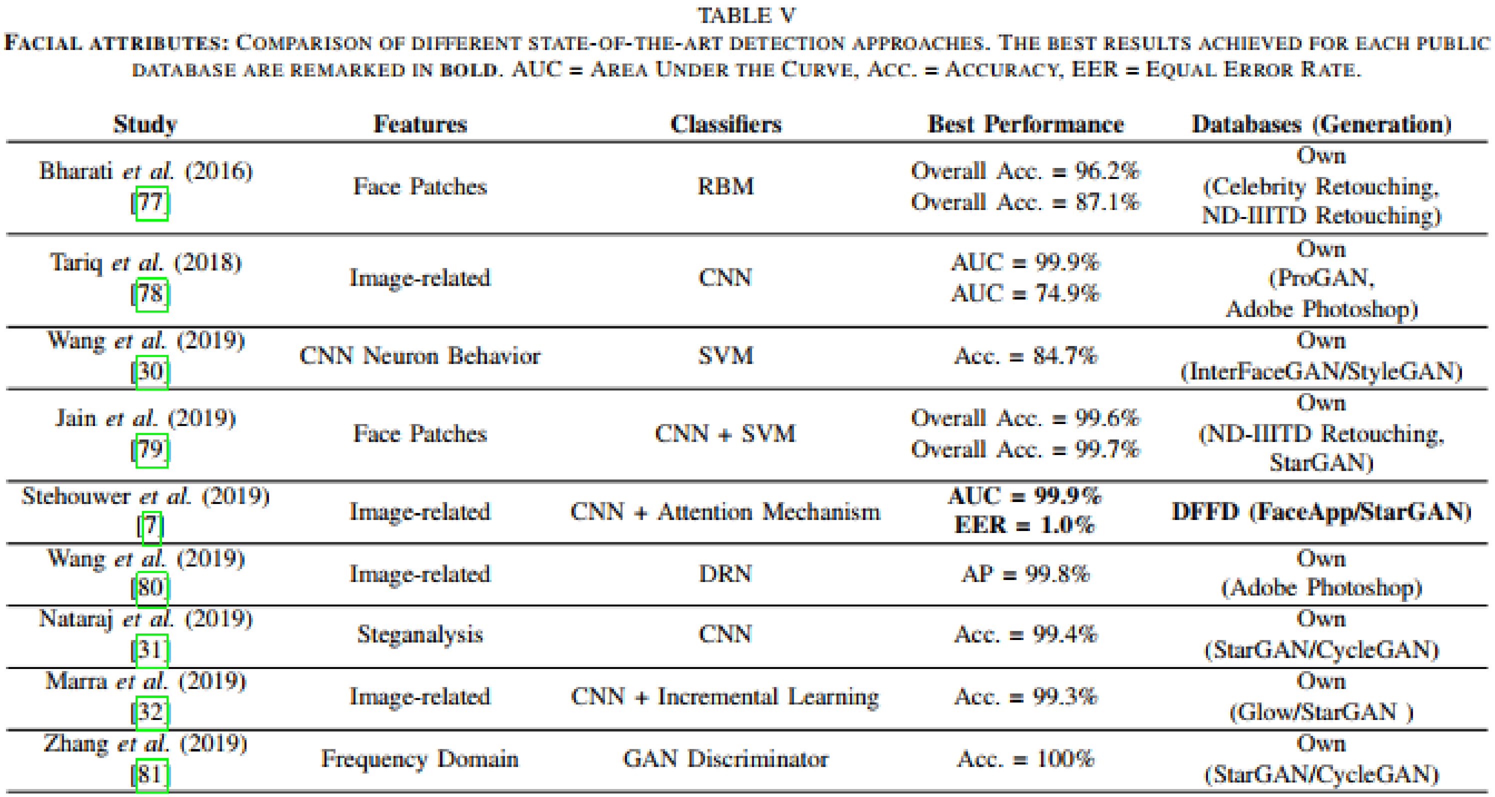


TABLE V

Facial attributes: Comparison of different state-of-the-art detection approaches. The best results achieved for each public database are remarked in bold. AUC = Area Under the Curve, Acc. = Accuracy, EER = Equal Error Rate.

Study	Features	Classifiers	Best Performance	Databases (Generation)
Bharati *et al.* (2016) [77]	Face Patches	RBM	Overall Acc. = 96.2% Overall Acc. = 87.1%	Own (Celebrity Retouching, ND-IIITD Retouching)

Tariq *et al.* (2018) [78]	Image-related	CNN	AUC = 99.9%
			AUC = 74.9%
				Own
				(ProGAN,
				Adobe Photoshop)

Wang *et al.* (2019) [30]	CNN Neuron Behavior	SVM	Acc. = 84.7%	Own (InterFaceGAN/StyleGAN)
Jain *et al.* (2019) [79]	Face Patches	CNN + SVM	Overall Acc. = 99.6% Overall Acc. = 99.7%	Own (ND-IIITD Retouching, StarGAN)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 99.9%** **EER = 1.0%**	**DFFD (FaceApp/StarGAN)**
Wang *et al.* (2019) [80]	Image-related	DRN	AP = 99.8%	Own (Adobe Photoshop)
Nataraj *et al.* (2019) [31]	Steganalysis	CNN	Acc. = 99.4%	Own (StarGAN/CycleGAN)
Marra *et al.* (2019) [32]	Image-related	CNN + Incremental Learning	Acc. = 99.3%	Own (Glow/StarGAN)
Zhang *et al.* (2019) [81]	Frequency Domain	GAN Discriminator	Acc. = 100%	Own (StarGAN/CycleGAN)

FACIAL ATTRIBUTES – MANIPULATION DETECTION

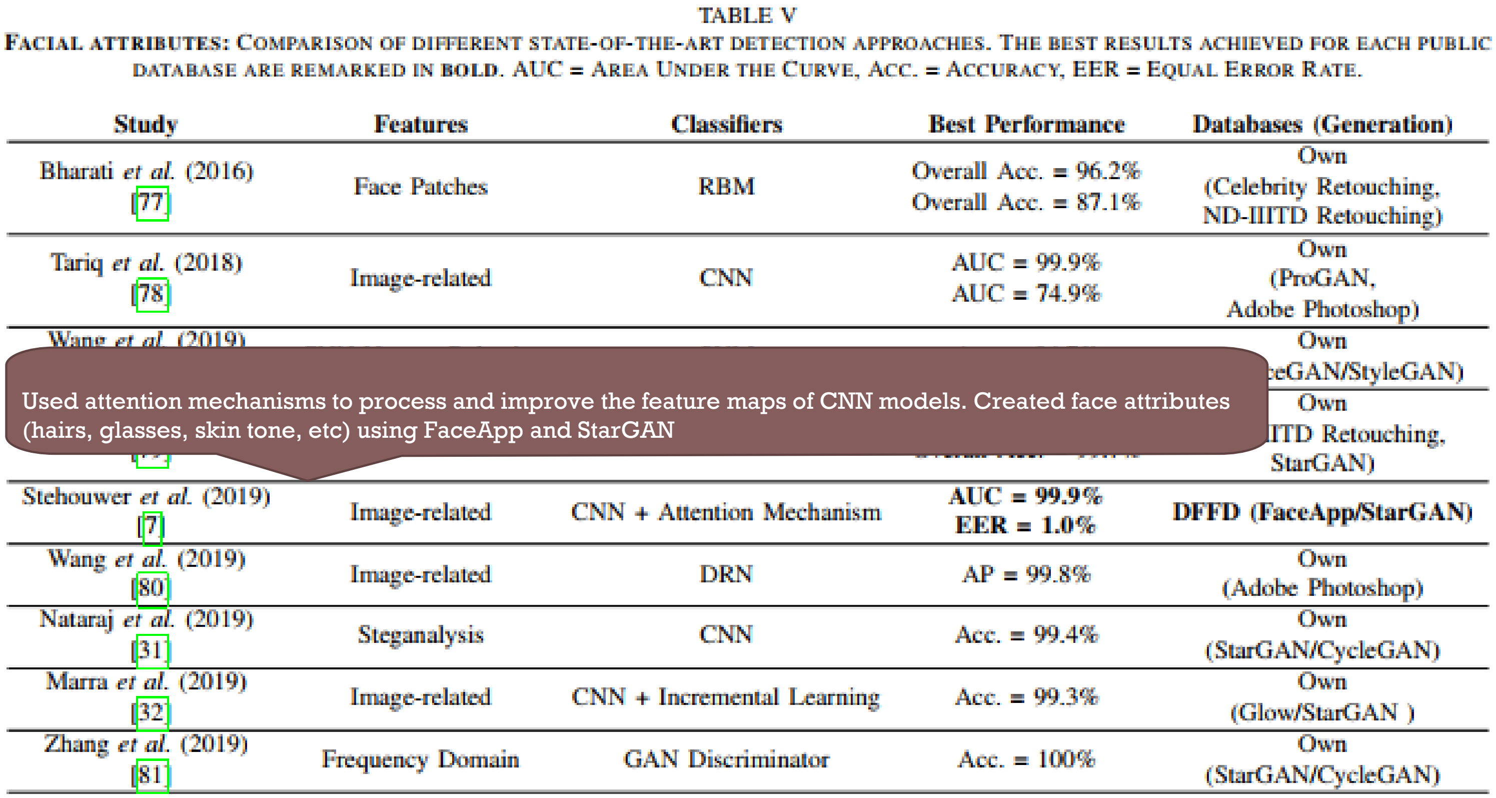


TABLE V

FACIAL ATTRIBUTES: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN BOLD. AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases (Generation)
Bharati *et al.* (2016) [77]	Face Patches	RBM	Overall Acc. = 96.2% Overall Acc. = 87.1%	Own (Celebrity Retouching, ND-IIITD Retouching)

Tariq *et al.* (2018)
[78]
	Image-related	CNN	AUC = 99.9%
			AUC = 74.9%
				Own
				(ProGAN,
				Adobe Photoshop)

Wang *et al.* (2019) [illegible]	[illegible]	[illegible]	[illegible]	Own [illegible]eGAN/StyleGAN)
[illegible]	[illegible]	[illegible]	[illegible]	Own [illegible]ITD Retouching, StarGAN)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 99.9%** **EER = 1.0%**	**DFFD (FaceApp/StarGAN)**
Wang *et al.* (2019) [80]	Image-related	DRN	AP = 99.8%	Own (Adobe Photoshop)
Nataraj *et al.* (2019) [31]	Steganalysis	CNN	Acc. = 99.4%	Own (StarGAN/CycleGAN)
Marra *et al.* (2019) [32]	Image-related	CNN + Incremental Learning	Acc. = 99.3%	Own (Glow/StarGAN)
Zhang *et al.* (2019) [81]	Frequency Domain	GAN Discriminator	Acc. = 100%	Own (StarGAN/CycleGAN)

Used attention mechanisms to process and improve the feature maps of CNN models. Created face attributes (hairs, glasses, skin tone, etc) using FaceApp and StarGAN

FACIAL ATTRIBUTES – MANIPULATION DETECTION

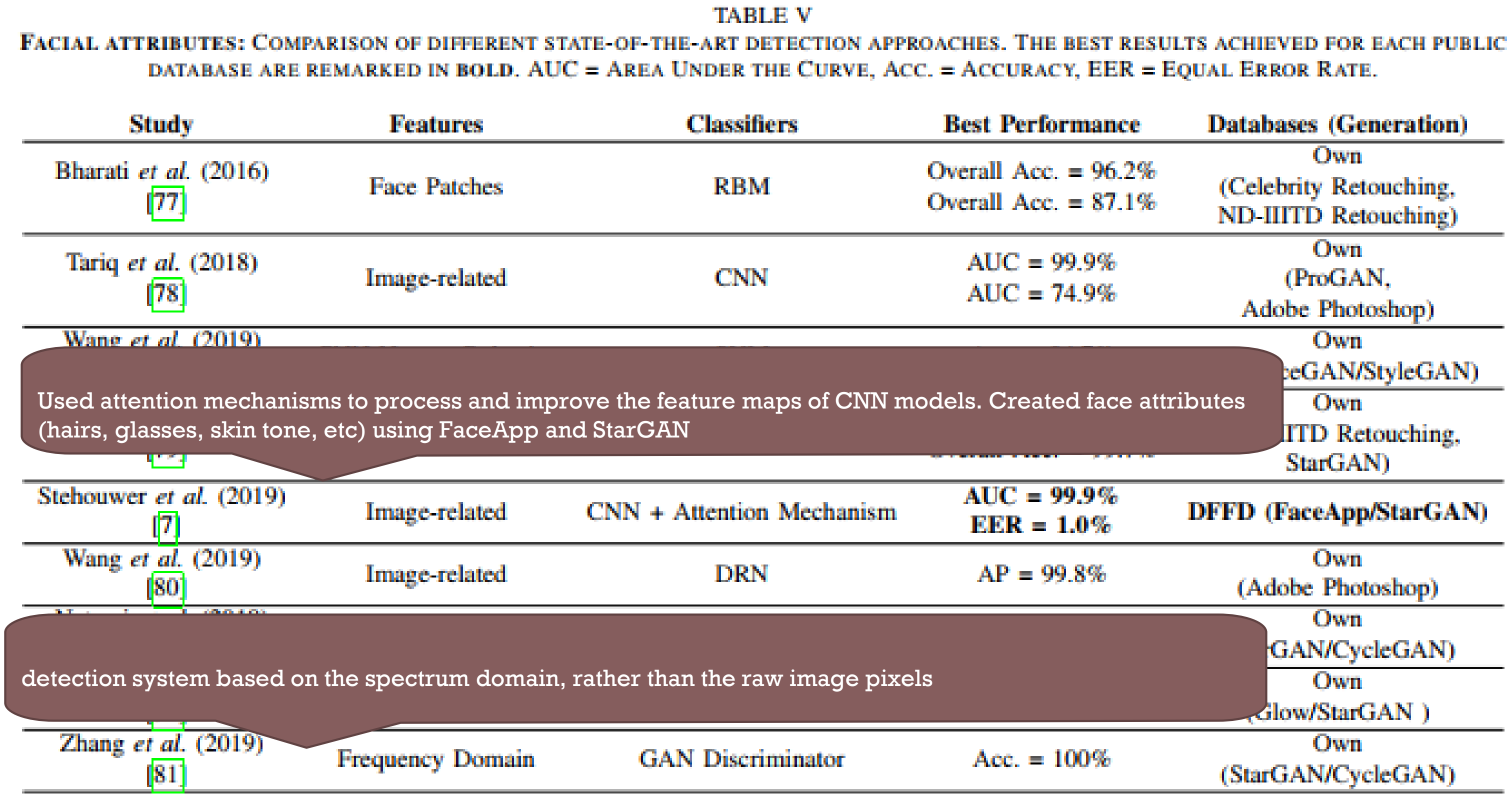


TABLE V

FACIAL ATTRIBUTES: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN BOLD. AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases (Generation)
Bharati *et al.* (2016) [77]	Face Patches	RBM	Overall Acc. = 96.2% Overall Acc. = 87.1%	Own (Celebrity Retouching, ND-IIITD Retouching)

Tariq *et al.* (2018)
[78]
	Image-related	CNN	AUC = 99.9%
			AUC = 74.9%
				Own
				(ProGAN,
				Adobe Photoshop)

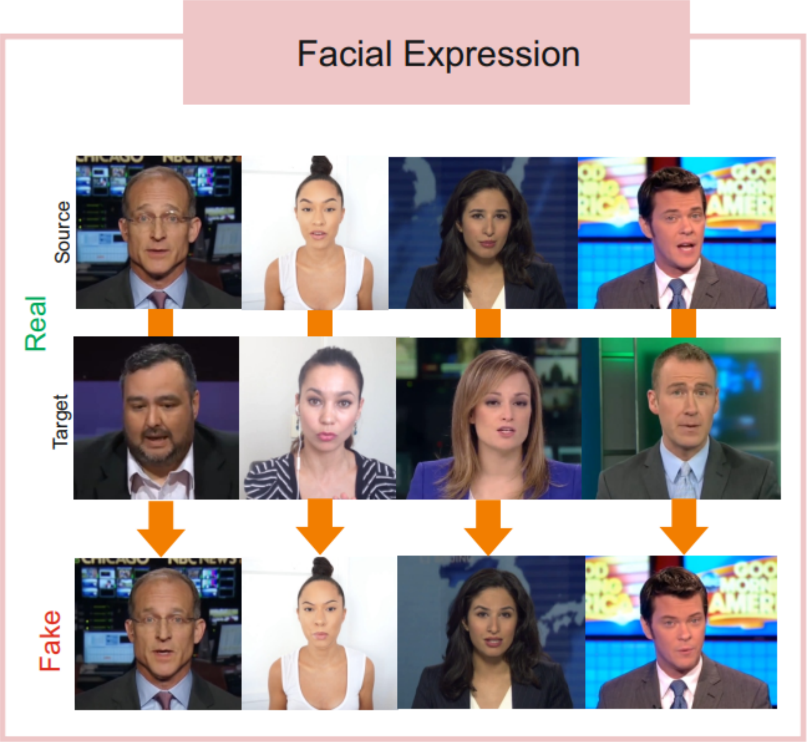
Wang *et al.* (2019)				Own eGAN/StyleGAN)
				Own ITD Retouching, StarGAN)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 99.9%** **EER = 1.0%**	**DFFD (FaceApp/StarGAN)**
Wang *et al.* (2019) [80]	Image-related	DRN	AP = 99.8%	Own (Adobe Photoshop)
				Own GAN/CycleGAN)
				Own Glow/StarGAN)
Zhang *et al.* (2019) [81]	Frequency Domain	GAN Discriminator	Acc. = 100%	Own (StarGAN/CycleGAN)

Used attention mechanisms to process and improve the feature maps of CNN models. Created face attributes (hairs, glasses, skin tone, etc) using FaceApp and StarGAN

detection system based on the spectrum domain, rather than the raw image pixels

4. FACIAL EXPRESSION MANIPULATION

- Modify the facial expression of the person, e.g., transferring the facial expression of one person to another person.
- Face2Face, FaceApp applications
- Models
  - StarGAN, InterFaceGAN, UGAN, STGAN, AttGAN, Autoencoders, GauGAN
- Sample: https://www.ted.com/talks/supasorn_suwajanakorn_fake_videos_of_real_people_and_how_to_spot_them?language=en
- Public Database
  - Face-Forensics++



FACIAL EXPRESSION MANIPULATION – DETECTION

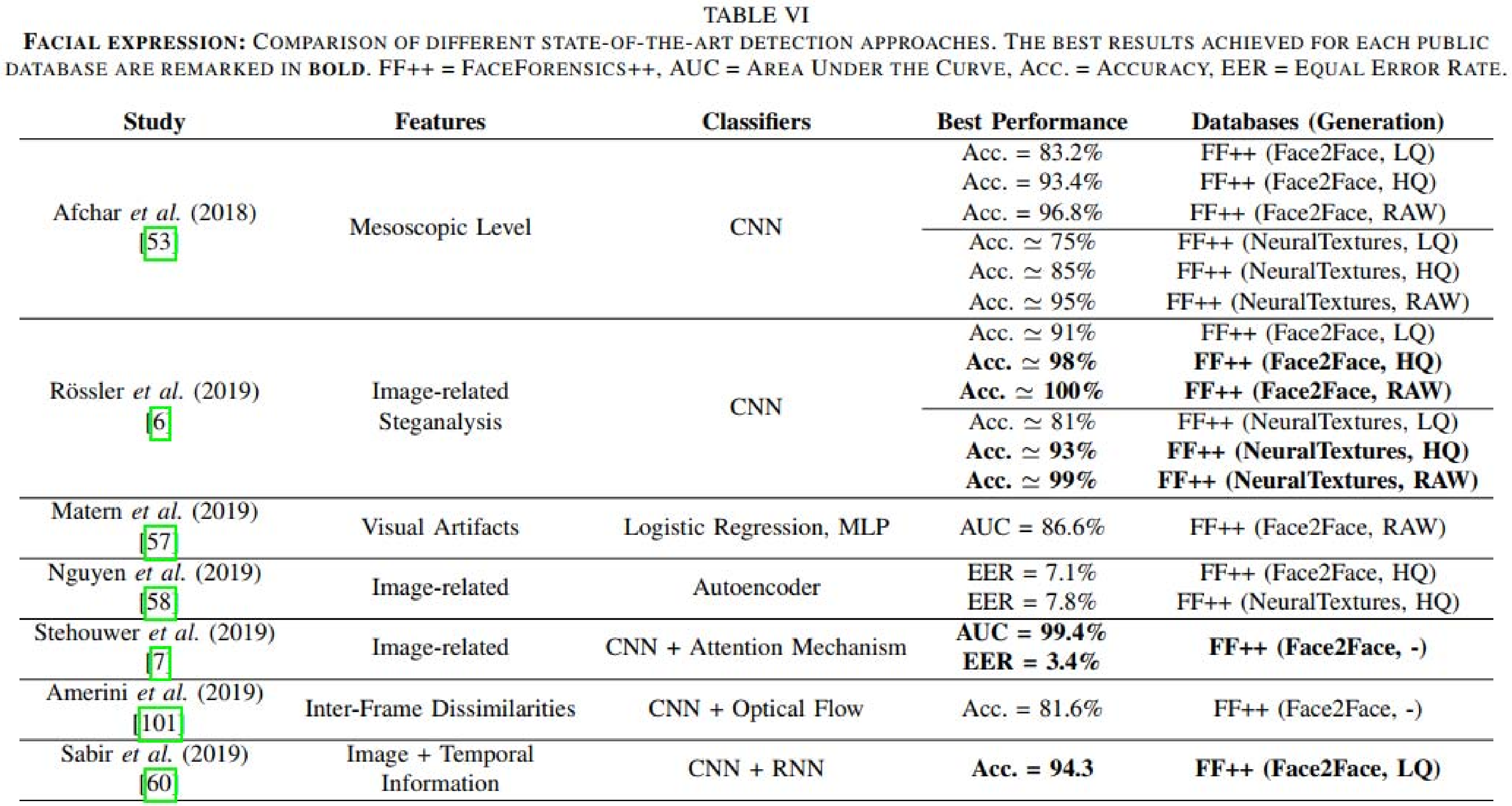


TABLE VI

FACIAL EXPRESSION: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. FF++ = FACEFORENSICS++, AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases (Generation)
Afchar *et al.* (2018) [53]	Mesoscopic Level	CNN	Acc. = 83.2%	FF++ (Face2Face, LQ)
			Acc. = 93.4%	FF++ (Face2Face, HQ)
			Acc. = 96.8%	FF++ (Face2Face, RAW)
			Acc. ≃ 75%	FF++ (NeuralTextures, LQ)
			Acc. ≃ 85%	FF++ (NeuralTextures, HQ)
			Acc. ≃ 95%	FF++ (NeuralTextures, RAW)
Rössler *et al.* (2019) [6]	Image-related Steganalysis	CNN	Acc. ≃ 91%	FF++ (Face2Face, LQ)
			Acc. ≃ 98%	**FF++ (Face2Face, HQ)**
			Acc. ≃ 100%	**FF++ (Face2Face, RAW)**
			Acc. ≃ 81%	FF++ (NeuralTextures, LQ)
			Acc. ≃ 93%	**FF++ (NeuralTextures, HQ)**
			Acc. ≃ 99%	**FF++ (NeuralTextures, RAW)**
Matern *et al.* (2019) [57]	Visual Artifacts	Logistic Regression, MLP	AUC = 86.6%	FF++ (Face2Face, RAW)
Nguyen *et al.* (2019) [58]	Image-related	Autoencoder	EER = 7.1%	FF++ (Face2Face, HQ)
			EER = 7.8%	FF++ (NeuralTextures, HQ)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 99.4%** **EER = 3.4%**	**FF++ (Face2Face, -)**
Amerini *et al.* (2019) [101]	Inter-Frame Dissimilarities	CNN + Optical Flow	Acc. = 81.6%	FF++ (Face2Face, -)
Sabir *et al.* (2019) [60]	Image + Temporal Information	CNN + RNN	**Acc. = 94.3**	**FF++ (Face2Face, LQ)**

FACIAL EXPRESSION MANIPULATION – DETECTION

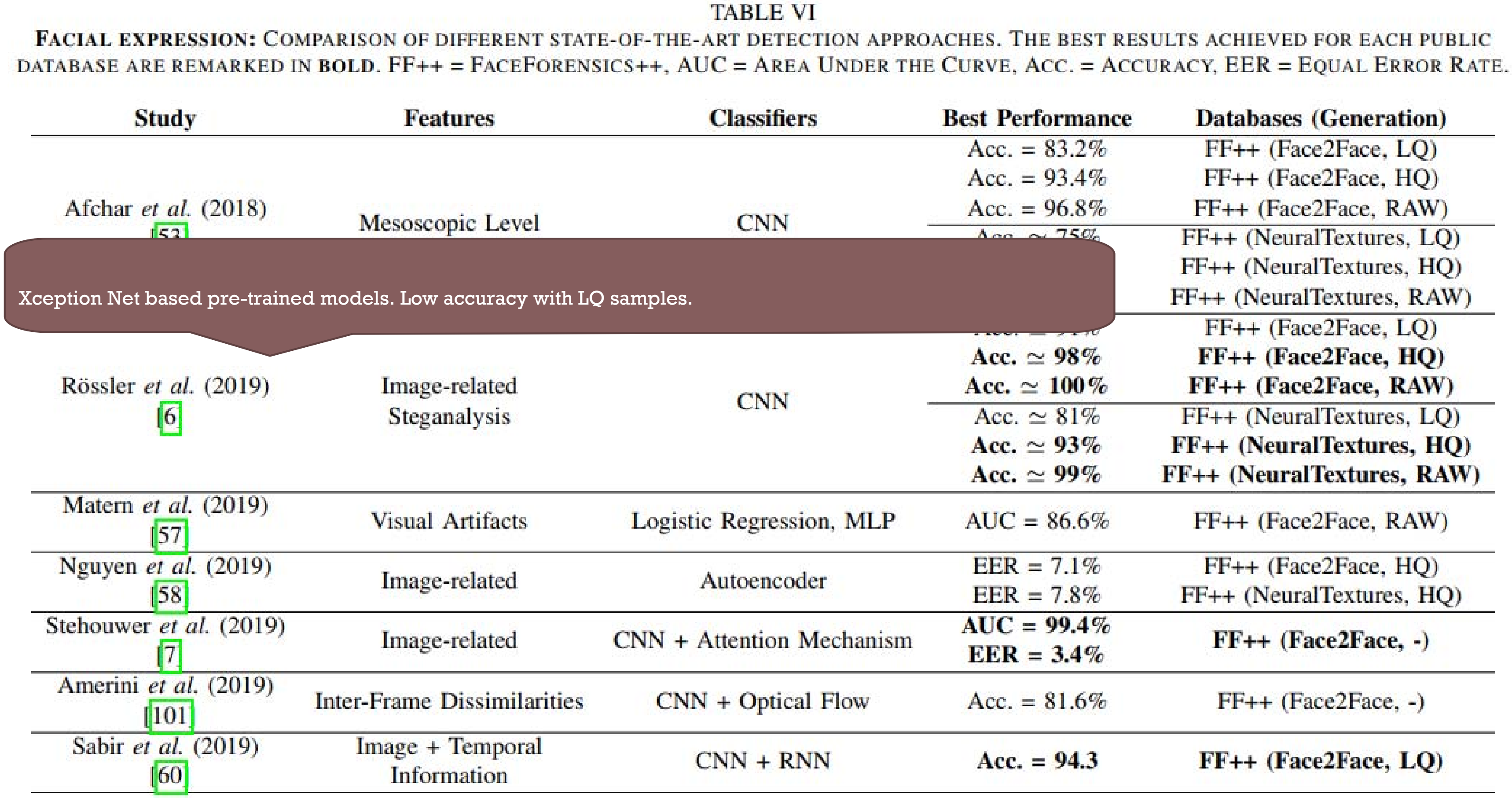


TABLE VI

FACIAL EXPRESSION: COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. FF++ = FACEFORENSICS++, AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

Study	Features	Classifiers	Best Performance	Databases (Generation)
Afchar *et al.* (2018) [53]	Mesoscopic Level	CNN	Acc. = 83.2%	FF++ (Face2Face, LQ)
			Acc. = 93.4%	FF++ (Face2Face, HQ)
			Acc. = 96.8%	FF++ (Face2Face, RAW)
			Acc. ≃ 75%	FF++ (NeuralTextures, LQ)
			[illegible]	FF++ (NeuralTextures, HQ)
			[illegible]	FF++ (NeuralTextures, RAW)
Rössler *et al.* (2019) [6]	Image-related Steganalysis	CNN	Acc. ≃ 91%	FF++ (Face2Face, LQ)
			Acc. ≃ 98%	**FF++ (Face2Face, HQ)**
			Acc. ≃ 100%	**FF++ (Face2Face, RAW)**
			Acc. ≃ 81%	FF++ (NeuralTextures, LQ)
			Acc. ≃ 93%	**FF++ (NeuralTextures, HQ)**
			Acc. ≃ 99%	**FF++ (NeuralTextures, RAW)**
Matern *et al.* (2019) [57]	Visual Artifacts	Logistic Regression, MLP	AUC = 86.6%	FF++ (Face2Face, RAW)
Nguyen *et al.* (2019) [58]	Image-related	Autoencoder	EER = 7.1%	FF++ (Face2Face, HQ)
			EER = 7.8%	FF++ (NeuralTextures, HQ)
Stehouwer *et al.* (2019) [7]	Image-related	CNN + Attention Mechanism	**AUC = 99.4%** **EER = 3.4%**	**FF++ (Face2Face, -)**
Amerini *et al.* (2019) [101]	Inter-Frame Dissimilarities	CNN + Optical Flow	Acc. = 81.6%	FF++ (Face2Face, -)
Sabir *et al.* (2019) [60]	Image + Temporal Information	CNN + RNN	**Acc. = 94.3**	**FF++ (Face2Face, LQ)**

Xception Net based pre-trained models. Low accuracy with LQ samples.

TRENDS AND COMPETITIONS

- **Media Forensics Challenge (MFC)**
  - launched by National Institute of Standards Technology (NIST)
  - 2018, 2019, 2020?
- **DeepFake Detection Challenge (DFDC)**
  - launched by Facebook and others. (https://deepfakedetectionchallenge.ai/)
  - Submission due Mar 31 2020.

CONCLUSION

- Most approaches for fake detection are focused on controlled scenarios, e.g., training and testing detection systems considering the same image compression level.
- DeepFake detection on real life scenarios are still challenging and need more work
  - Scenarios like, image/video compression levels, noise, blur, etc
- Robustness of the detection systems of unseen face manipulation attacks (e.g. use of future GANs or other models) is a big question!

REFERENCES

- [1] Tolosana, Ruben, et al. "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection." arXiv preprint arXiv:2001.00179 (2020).