Truth, Trust, and Transparency in Synthetic Media

Undergraduate Students: Helen Chen, John Jack Lewis
Graduate Student: Imad Eddine Toubal
Faculty Advisors: Dr. Palaniappan, Dr. Prasad


abc
#NIGHTLINE


weta
DIGITAL

Deepfakes

Synthetic videos that contain altered faces and/or voices of a subject

Total number of video views across top four dedicated deepfake pornography websites

134,364,438

percentage of deepfake videos online by pornographic and non-pornographic content

96%

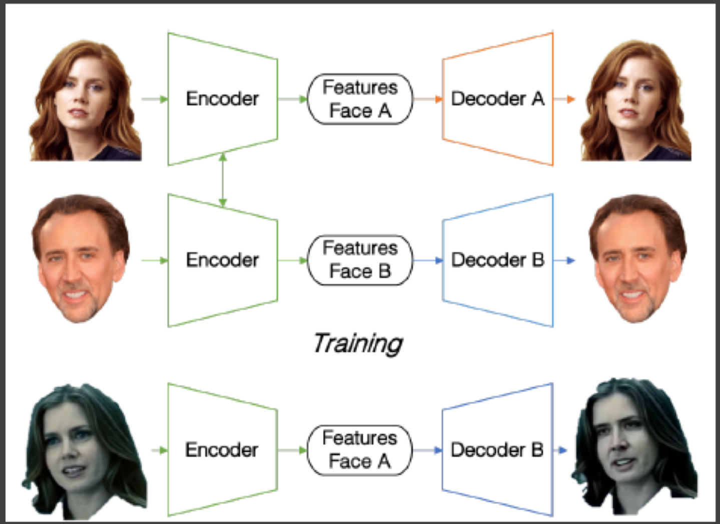
4%

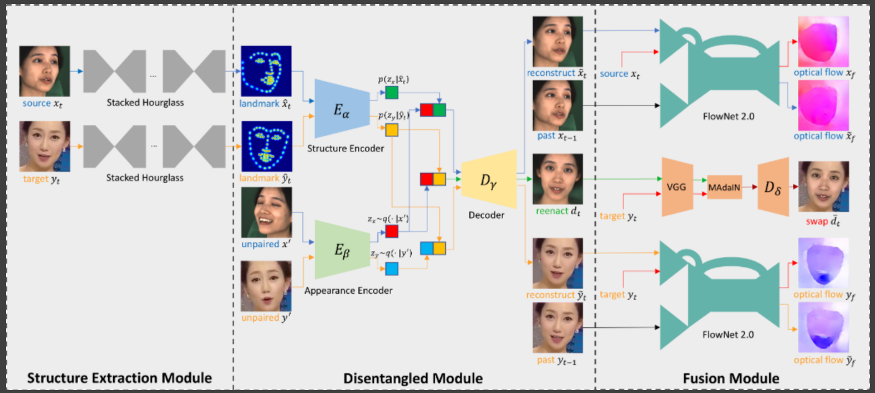

AZMIN

Problem Overview

Methods of “Deepfaking”

Variational Autoencoders



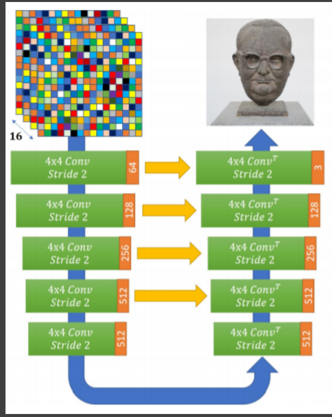


DeepFake Variational Auto-Encoder

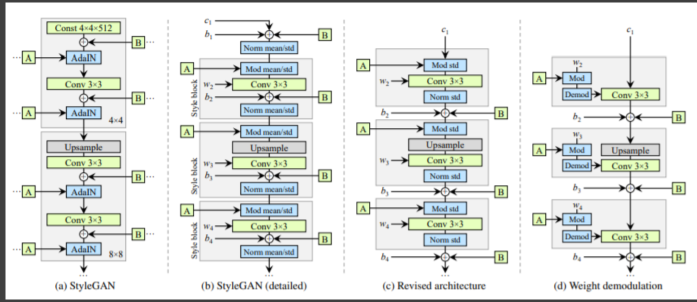
AUTOENCODERS

GENERATIVE ADVERSARIAL NETWORKS

Neural Textures



StyleGAN2



RecycleGAN

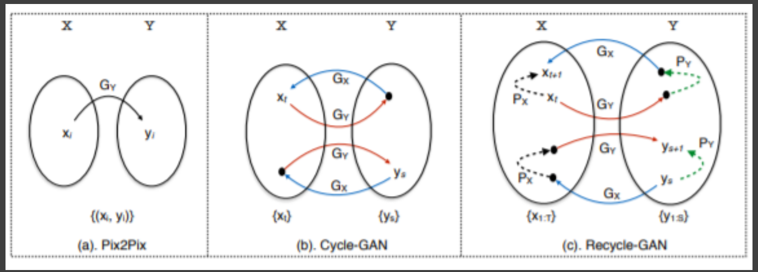


Figure 2. Samples of different methods displaying difference between color of the left and right eye. (Top to bottom: [18], [21], image taken from [39])

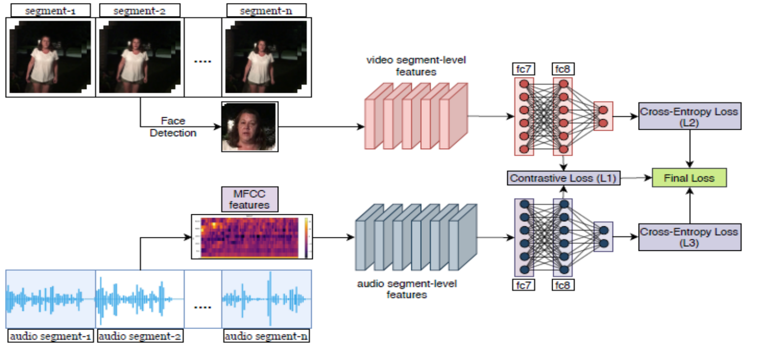


Figure 1: MDS-based fake video detection: Features extracted from 1-second audio-visual segments are input to the MDS network. The MDS network comprises the audio and visual sub-networks, whose description is provided in Table 1. Descriptors learned by the video and audio sub-networks are tuned via the cross-entropy loss, while the contrastive loss is employed to enforce higher dissimilarity between audio-visual chunks arising from *fake* videos. MDS is computed as the aggregate audio-visual dissonance over the video length, and employed as a figure of merit for labeling a video as *real*/*fake*.

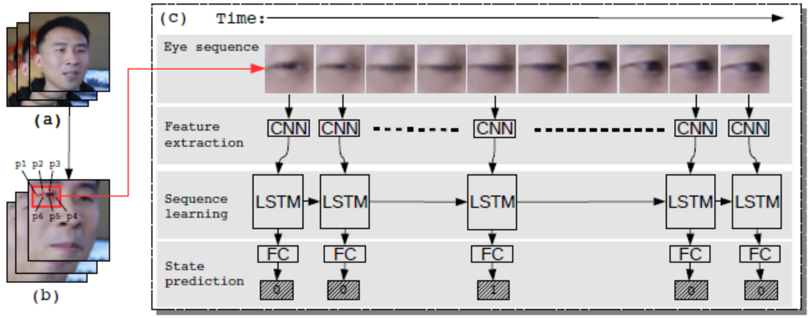


Figure 2. Overview of our LRCN method. (a) is the original sequence. (b) is the sequence after face alignment. We crop out eye region of each frame based on eye landmarks $p_{1\sim6}$ in (b) and pass it to (c) LRCN, which consists of three parts: feature extraction, sequence learning and state prediction.

Figure 3. Example from FaceForensics [33] showing shading artifacts arising from illumination estimation and imprecise geometry of the nose.

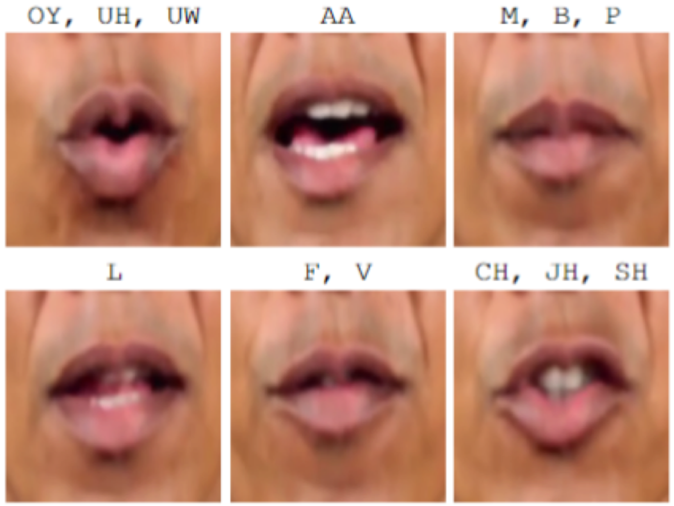


Figure 1. Six example visemes and their corresponding phonemes. The phonemes in the top-right (M, B, P), for example, correspond to the sound you make when you say "mother", "brother", or "parent". To make this sound, you must tightly press your lips together, leading to the shown viseme.

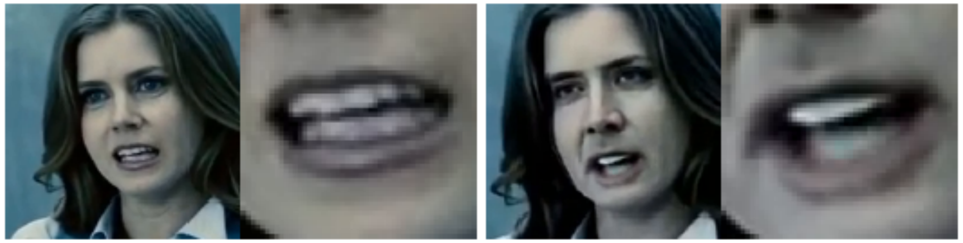


Figure 6. Missing geometry in Deepfakes. Teeth are generated as a structureless white blob. Samples from the dataset in Sec. 4.1.

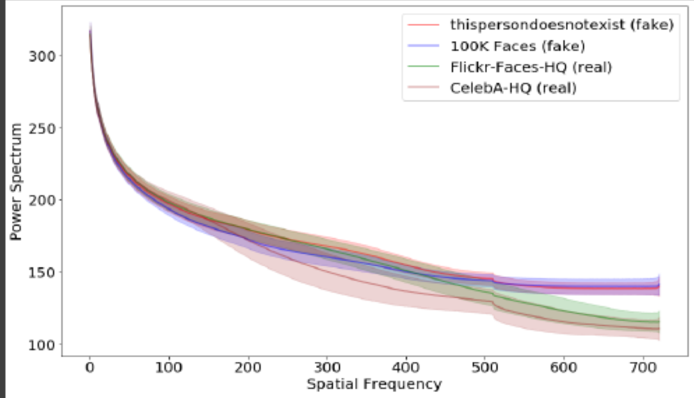
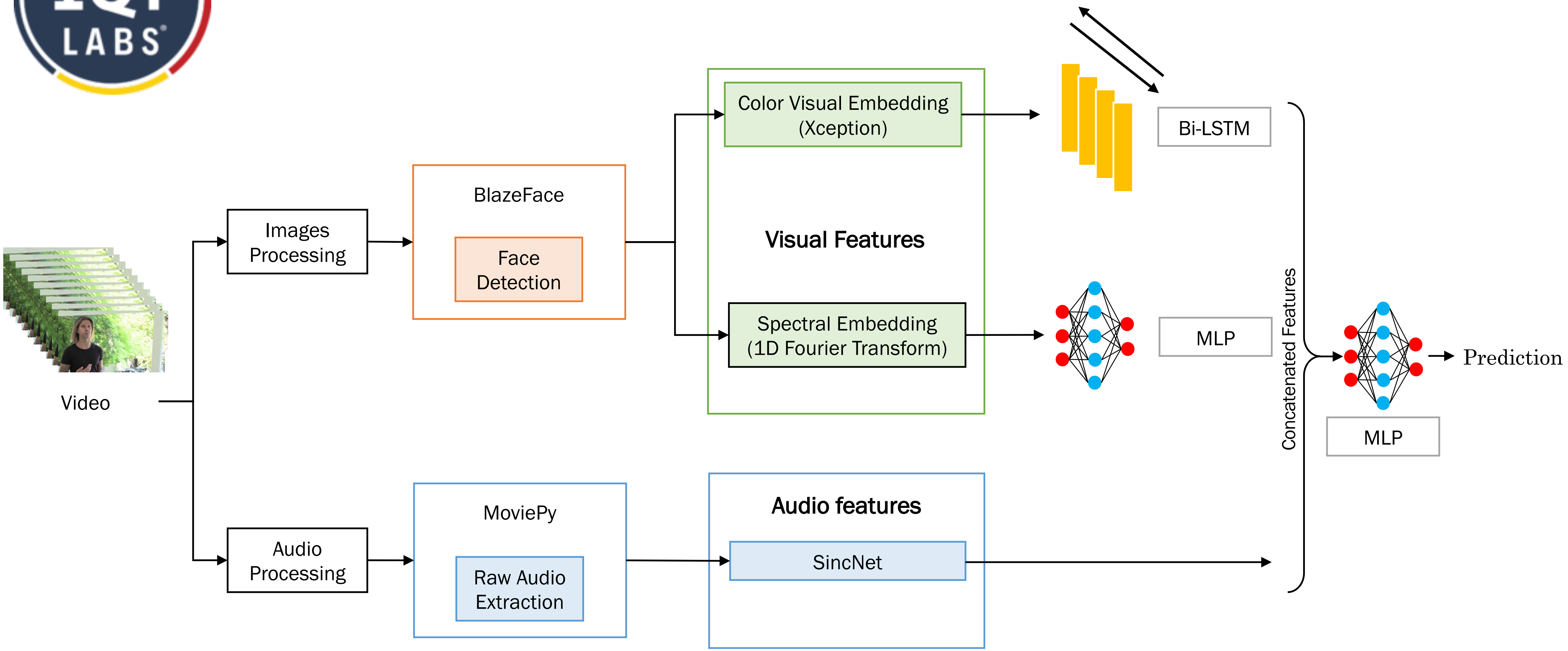
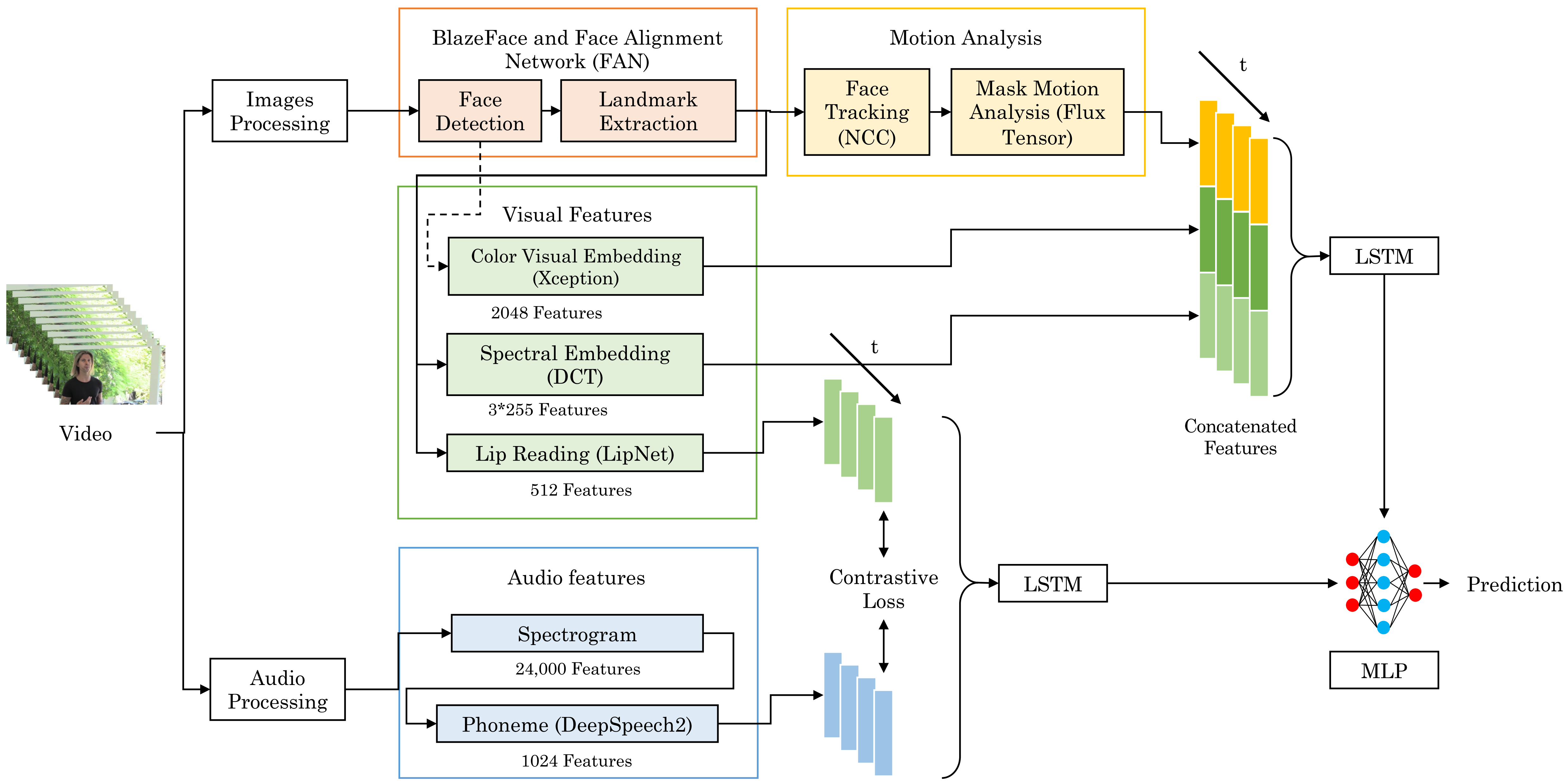


Fig. 1: 1D power spectrum statistics from each sub-data set from Faces-HQ. The higher the frequency, the bigger is the difference between real or fake data.

Can we build a multimodal network to optimize the detection of all types of deepfakes?



Video
Images Processing
Audio Processing
BlazeFace
Face Detection
MoviePy
Raw Audio Extraction
Color Visual Embedding (Xception)
Visual Features
Spectral Embedding (1D Fourier Transform)
Audio features
SincNet
Bi-LSTM
MLP
Concatenated Features
MLP
Prediction



Video
Images Processing
BlazeFace and Face Alignment Network (FAN)
Face Detection
Landmark Extraction
Motion Analysis
Face Tracking (NCC)
Mask Motion Analysis (Flux Tensor)
t
Visual Features
Color Visual Embedding (Xception)
2048 Features
Spectral Embedding (DCT)
3*255 Features
Lip Reading (LipNet)
512 Features
t
Concatenated Features
LSTM
Contrastive Loss
Audio Processing
Audio features
Spectrogram
24,000 Features
Phoneme (DeepSpeech2)
1024 Features
LSTM
MLP
Prediction

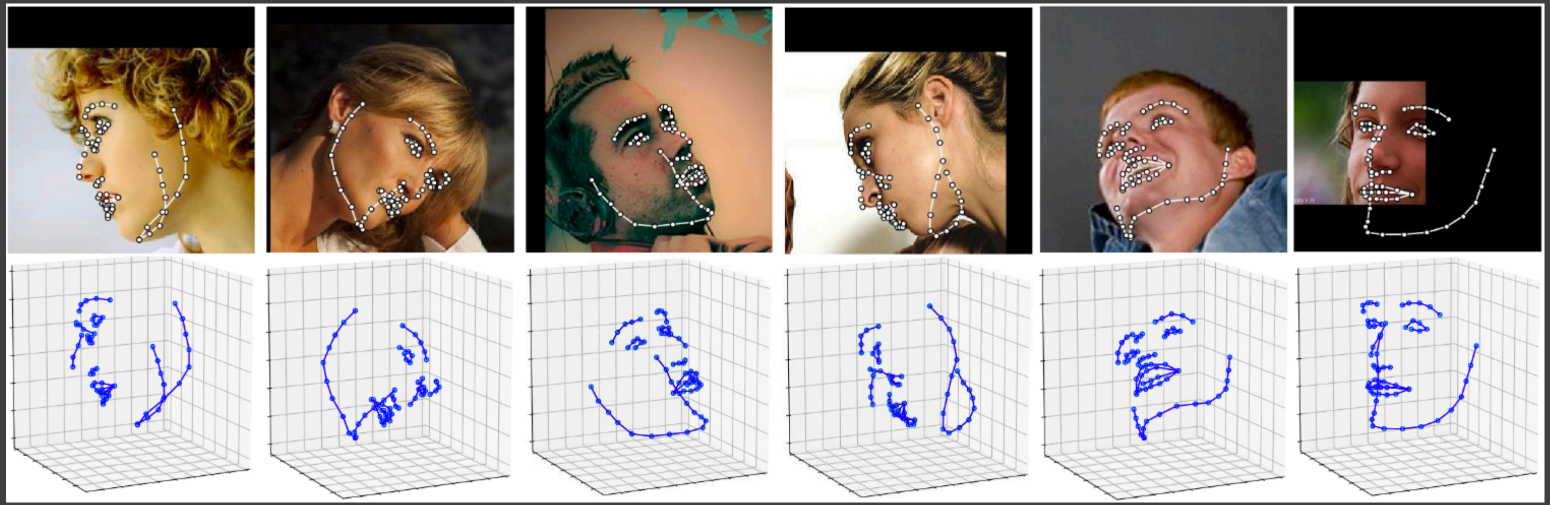
Results

Thank you!

John K. Lewis jklc9f@mail.missouri.edu
Helen Chen hc72t@mail.missouri.edu
Imad Eddine Toubal itoubal@mail.missouri.edu

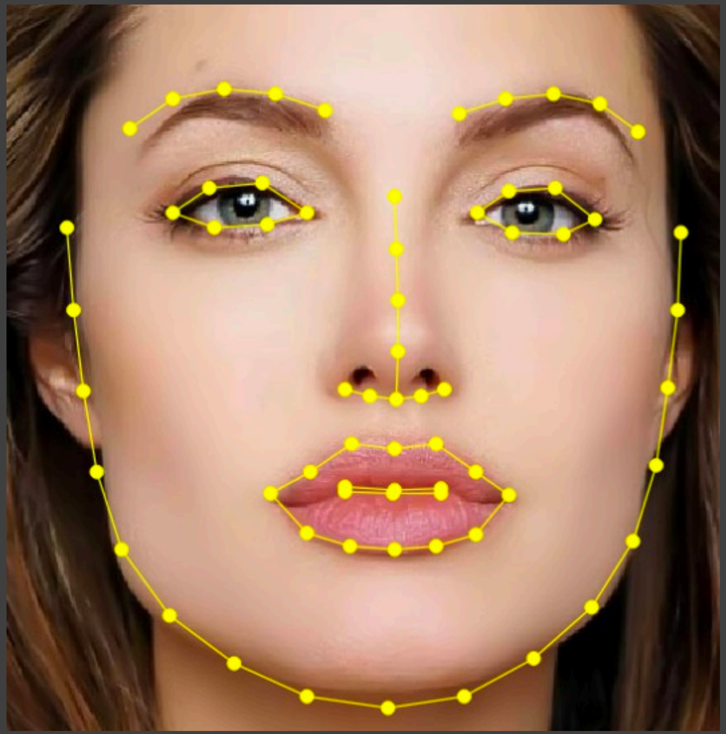
Backup Slides

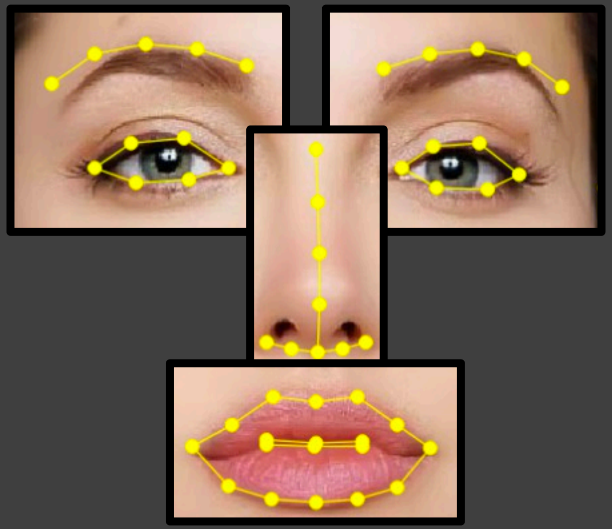
Face detection: FANet



- Alternatives:
  - BlazeFace (200-1000 fps)
  - S^3FD: Single Shot Scale-invariant Face Detector (36 fps)
  - FANet: Face Alignment Network (5 fps)
  - Super-FAN: Enhanced FANet

Landmark Extraction





Audio features: Deep Speech

- Alternatives:
  - SyncNet
    - Ravanelli, M., & Bengio, Y. (2018, December). Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1021-1028). IEEE.
  - Deep Speech
    - Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.

Color Visual features: Xception

- Inspired by InceptionV3 where Inception modules have been replaced with depth-wise separable convolutions

- Alternatives:
  - InceptionV3
  - MobileNetV2
  - ResNext

Table 1. Classification performance comparison on ImageNet (single crop, single model). VGG-16 and ResNet-152 numbers are only included as a reminder. The version of Inception V3 being benchmarked does not include the auxiliary tower.

	Top-1 accuracy	Top-5 accuracy
VGG-16	0.715	0.901
ResNet-152	0.770	0.933
Inception V3	0.782	0.941
Xception	**0.790**	**0.945**

Table 3. Size and training speed comparison.

	Parameter count	Steps/second
Inception V3	23,626,728	31
Xception	22,855,952	28

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).

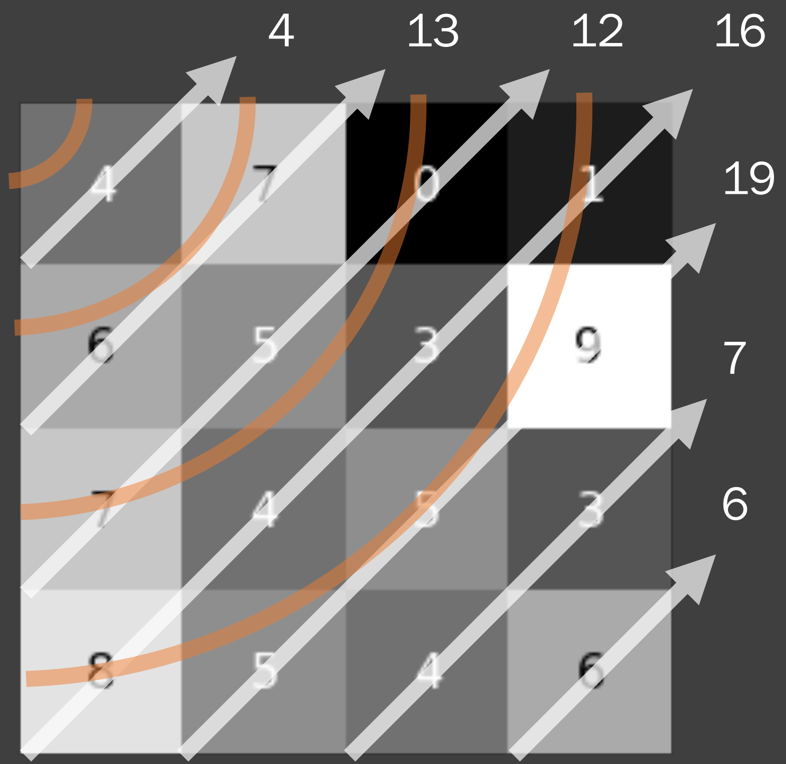
Lip Reading: LipNet

Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.

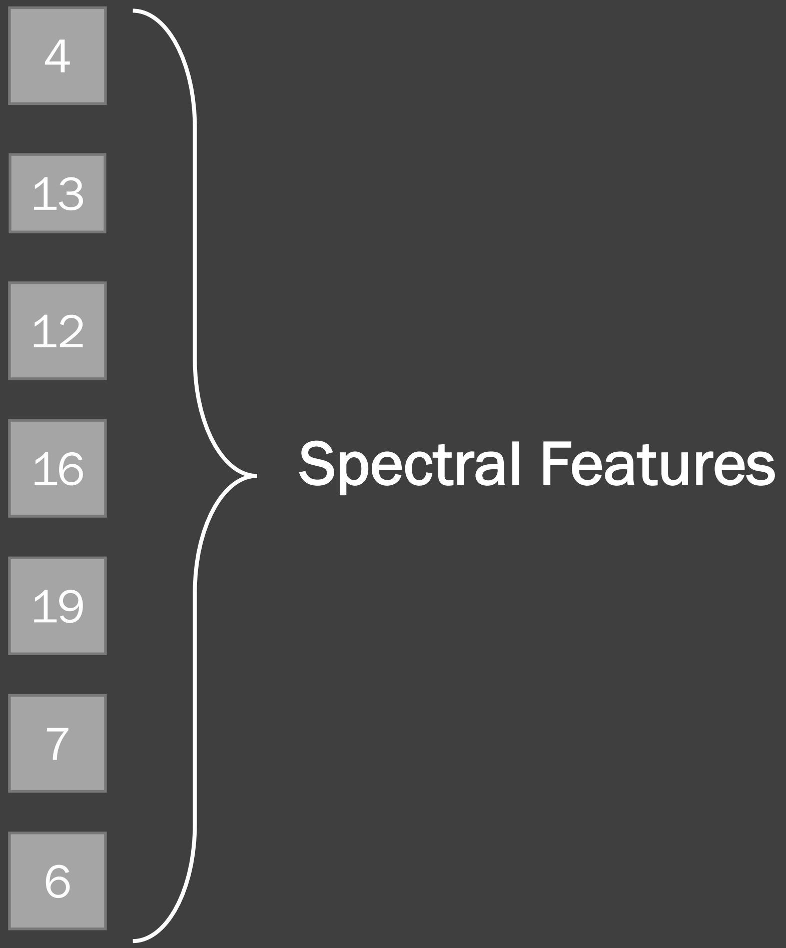
- Alternatives:

Spectral Features: DCT

Estimation of Azimuthal Average



Discrete Cosine Transform



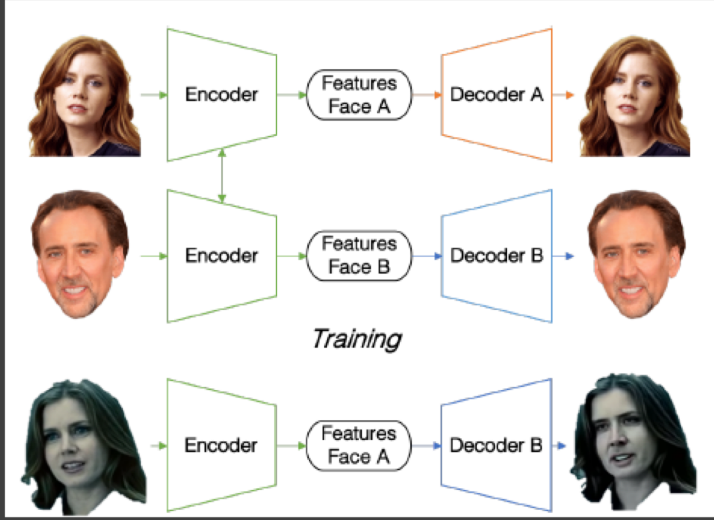

stable voices
VFXCHRISUME.COM
this is fake

Relevant Deepfake Detection Methods

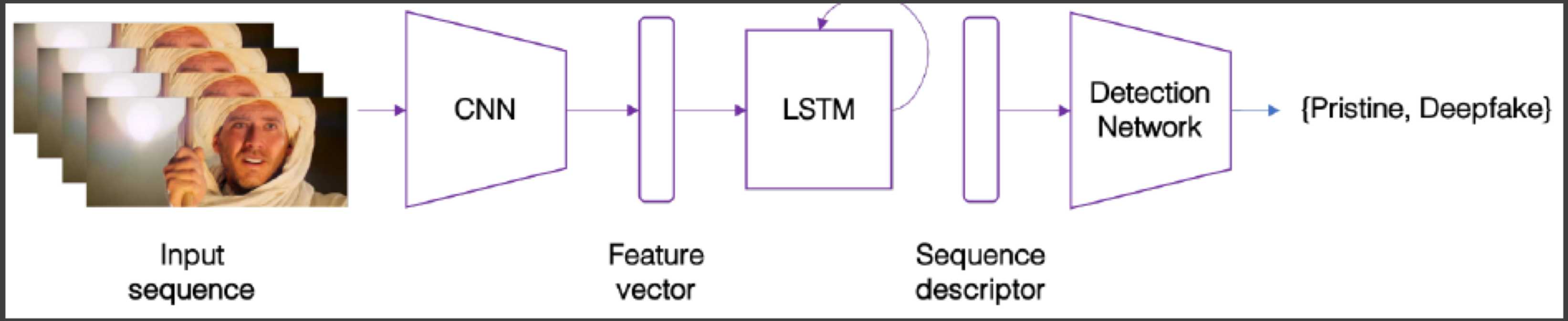
Deepfake Video Detection Using Recurrent Neural Networks

David Güera and Edward J. Delp

- 3 exploitations
  - Production Inconsistencies
  - Facial Boundaries
  - Temporal Awareness



- Model structure



Güera, D. and E. J. Delp. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Nov. 2018, pp. 1–6.

Exposing Deepfakes By Detecting Face Warping Artifacts

Yuezun Li and Siwei Lyu

- Artifact Detection with CNN
- Affine Transformation

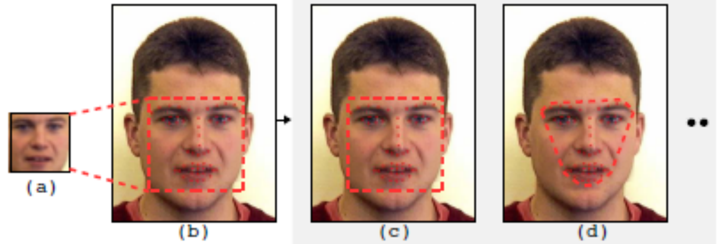


Figure 3. *Illustration of face shape augmentation of negative examples. (a) is the aligned and blurred face, which then undergoes an affine warped back to (b). (c, d) are post-processing for refining the shape of face area. (c) denotes the whole warped face is retained and (d) denotes only face area inside the polygon is retained.*

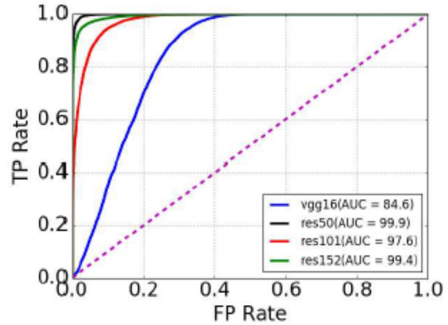


Figure 6. *Performance of each CNN model on all frames in LQ set of DeepFakeTIMIT [18].*

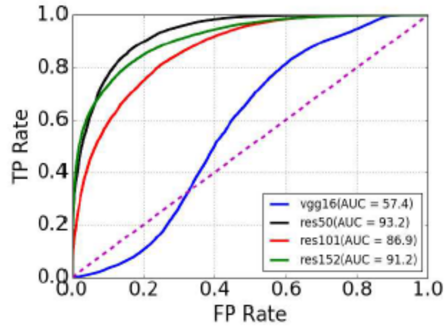


Figure 7. *Performance of each CNN model on all frames in HQ set of DeepFakeTIMIT [18].*

Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," CVPRW, pp. 46-52, 2018.

In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking

Yuezun Li et al.

- LRCN (Long Term Recurrent CNN)

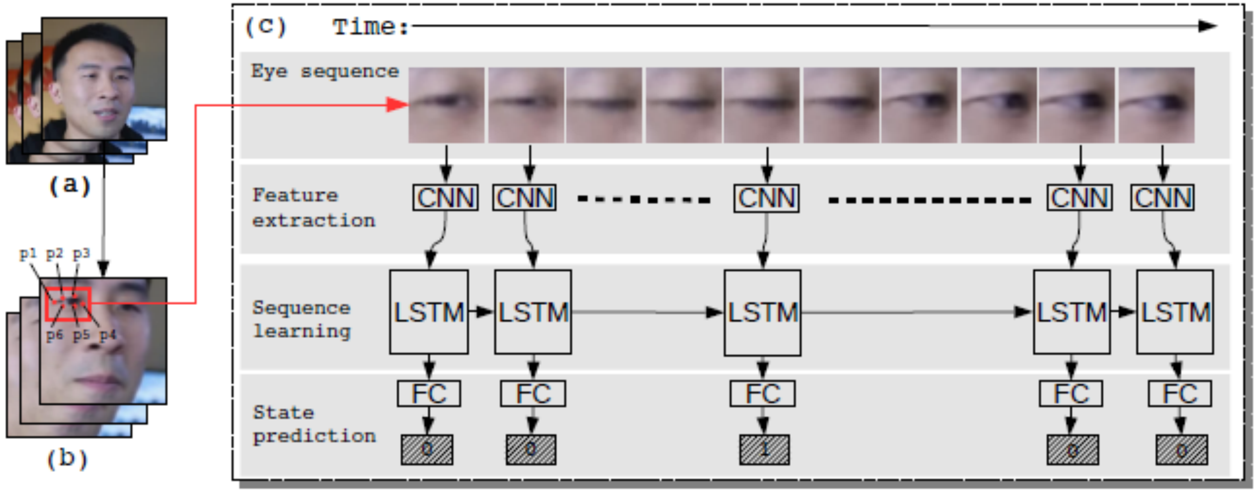


Figure 2. *Overview of our LRCN method. (a) is the original sequence. (b) is the sequence after face alignment. We crop out eye region of each frame based on eye landmarks $p_{1\sim 6}$ in (b) and pass it to (c) LRCN, which consists of three parts: feature extraction, sequence learning and state prediction.*

Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630787.

Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations

Falko Matern et al.

- Eye color
- Shadow
- Teeth

Figure 3. Example from FaceForensics [33] showing shading artifacts arising from illumination estimation and imprecise geometry of the nose.

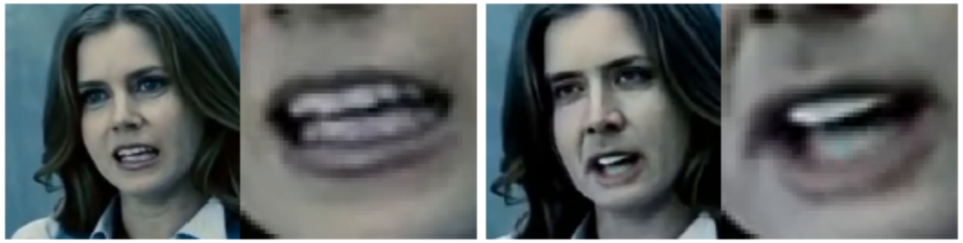


Figure 6. Missing geometry in Deepfakes. Teeth are generated as a structureless white blob. Samples from the dataset in Sec. 4.1.

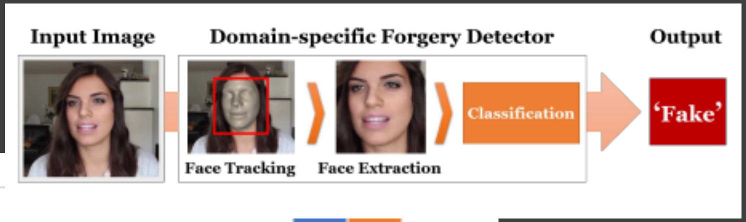
Figure 2. Samples of different methods displaying difference between color of the left and right eye. (Top to bottom: [18], [21], image taken from [39])

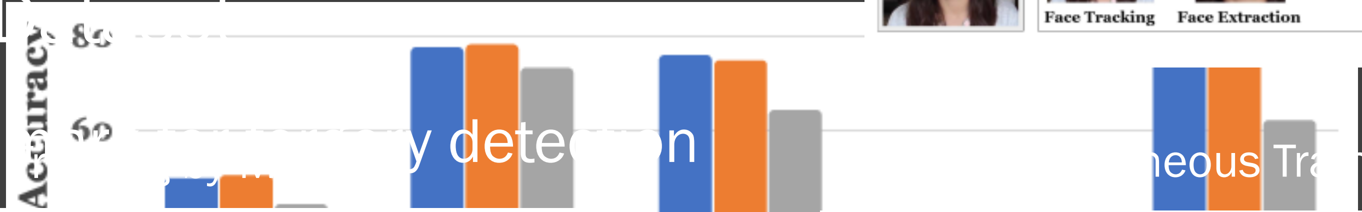
F. Matern, C. Riess, and M. Stamminger, “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations,” in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 83–92, doi: 10.1109/WACVW.2019.00020.

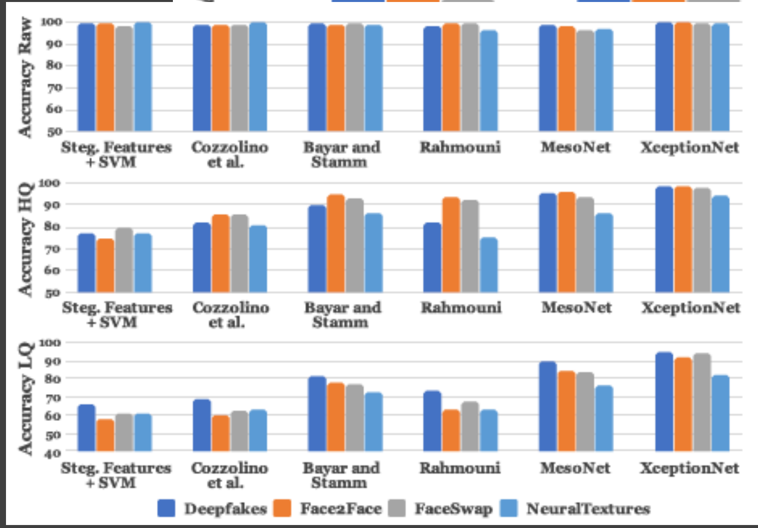
FaceForensics++: Learning to Detect Manipulated Facial Images

Andreas Rossler et al.

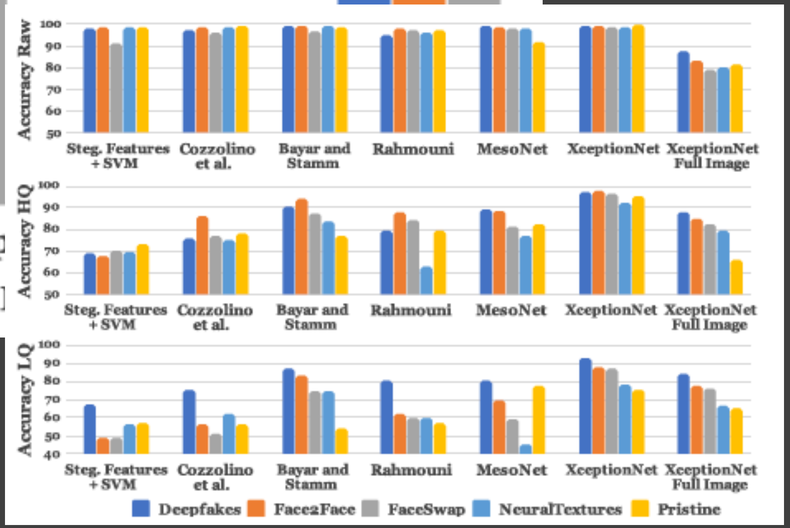
- Forgery Detection
- Diverse D
  - Bench y detection







neous Tra ning



A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 1–11, doi: 10.1109/ICCV.2019.00009.

Unmasking DeepFakes with Simple Features

Richard Durall et al.

- Fast Fourier Transform
  - Azimuthal Average
- Medium-High Resolution Success
  - Low resolution valley

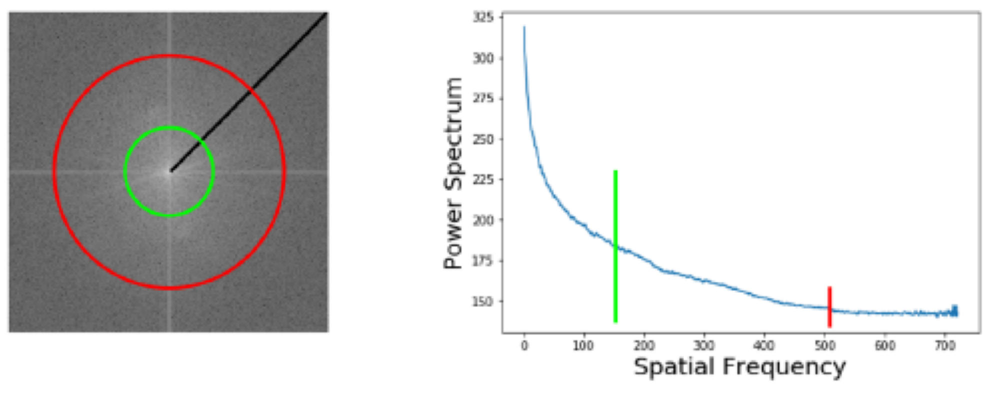
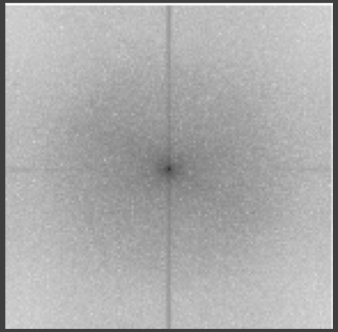


Fig. 4: Example of an azimuthal average. (Left) Power Spectrum 2D. (Right) Power Spectrum 1D. Each frequency component is the radial average from the 2D spectrum.



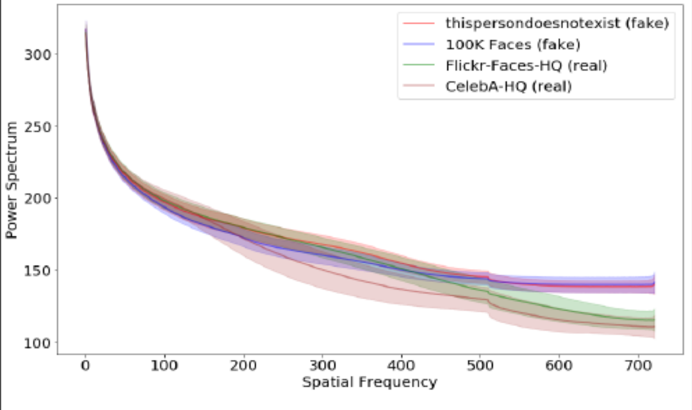


Fig. 1: 1D power spectrum statistics from each sub-data set from Faces-HQ. The higher the frequency, the bigger is the difference between real or fake data.

R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking DeepFakes with simple Features," *arXiv:1911.00686 [cs, stat]*, Mar. 2020, Accessed: Jun. 15, 2020. [Online]. Available: http://arxiv.org/abs/1911.00686.

Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches (2020)

Shruti Agarwal et al.

- Letters 'm', 'b', and 'p'
- Phoneme
- Viseme

dataset	profile	CNN
original	99.4%	99.6%
A2V	96.6%	96.9%
T2V-L	83.7%	71.1%
T2V-S	89.5%	80.7%
in-the-wild	93.9%	97.0%

Table 3. The accuracy of the two automatic techniques (profile and CNN) to detect if a mouth is open or closed. The accuracies are computed at a fixed threshold corresponding to average false alarm rate of 0.5% (i.e., misclassifying a closed mouth as open).

Figure 1. Six example visemes and their corresponding phonemes. The phonemes in the top-right (M, B, P), for example, correspond to the sound you make when you say "mother", "brother", or "parent". To make this sound, you must tightly press your lips together, leading to the shown viseme.

youtu.be/VWMEDacz3L4

Agarwal, S., Farid, H., Fried, O., and M. Agrawala. Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. In *Workshop on Media Forensics at CVPR*, Seattle, WA, 2020.

Not made for each other- Audio-Visual Dissonance-based Deepfake Detection and Localization

Komal Chugh et al.

- MDS (Modality Dissonance Score)

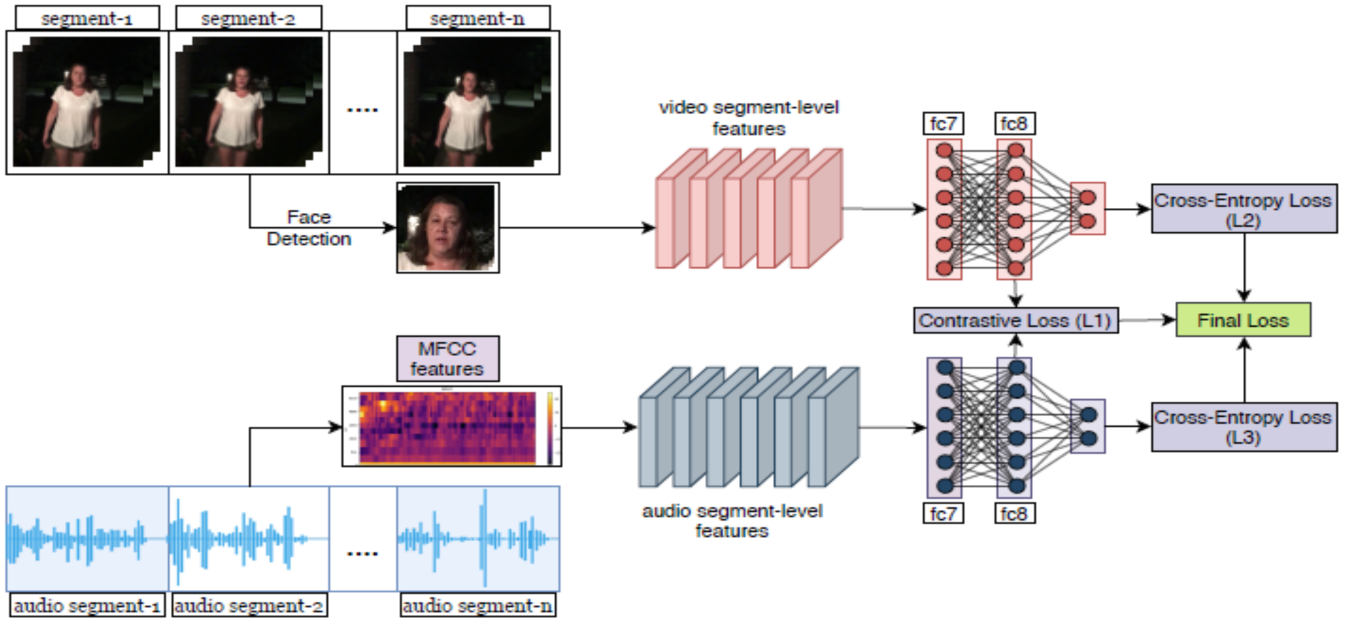


Figure 1: MDS-based fake video detection: Features extracted from 1-second audio-visual segments are input to the MDS network. The MDS network comprises the audio and visual sub-networks, whose description is provided in Table 1. Descriptors learned by the video and audio sub-networks are tuned via the cross-entropy loss, while the contrastive loss is employed to enforce higher dissimilarity between audio-visual chunks arising from *fake* videos. MDS is computed as the aggregate audio-visual dissonance over the video length, and employed as a figure of merit for labeling a video as *real*/*fake*.

K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other- Audio-Visual Dissonance-based Deepfake Detection and Localization," *arXiv:2005.14405 [cs]*, May 2020, Accessed: Jun. 19, 2020. [Online]. Available: http://arxiv.org/abs/2005.14405.

You said that https://arxiv.org/pdf/1705.02966.pdf

- Figure 1, 2 and 6

Sample Videos from DFDC Dataset