# A Multimodal Method to Detect DeepFake Videos

Jinchen Wu · Weijian Zhang · Ruoye Wang

# Overall Project Goals

- Multimedia forgery threats
- DeepFake: face-swapping
- Our aim: learning-based deepfake video detection


https://www.youtube.com/watch?v=_q16aJTXVRE&t=87s

# Specific Aims

**Multimodal framework**

- Visual-audio features
- Deepfake video detection

**Performance**

- Comparable to state-of-the-art techniques on DFTIMIT and DFDC datasets

**Comparison**

- With different architectures and results
- Why their performance differ?

# Current State of the Art

**DeepFake detection categories**

- Intra-frame visual artifacts
- Inter-frame inconsistencies
- Multimodel features

# Current State of the Art: unimodal

## Approaches

- Intra-frame visual artifacts: DNN during face wrapping [1], discontinuity of head pose [2]
- Inter-frame inconsistencies: [3][4] combine CNN and LSTM for detection

## Limitations

- Low performance
- Various unutilized features
- Vulnerable to manipulated audio

# Current State of the Art: multimodal

**Mittal et al.[5]: congruence between emotions**

- Double Siamese network
- Extract features and vectors from emotion and audio respectively

**Hosler et al.[6]:**

- LSTM network: predict emotions from Low-level Descriptors
- Train real-fake detector from predicted emotion

# Novelty

## Multimodal architecture

- Measure similarities
- Use similarities

## Investigate losses

- Different combinations affect performance

## LSTM

- Combine with bimodal architecture
- Evaluate the results

## Performance

- Of different frameworks
- On DFTIMIT and DFDC datasets

# Importance

**Framework designs**

- Inspiration on audio-visual frameworks

**LSTM**

- Insight on its influence on similarity measurement

**Datasets**

- Deeper understanding on DFTIMIT and DFDC

## Datasets

**DFTIMIT**[11]

- 32 different people from VIDTIMIT[14]
- Manipulated with FS-GAN[13]
- Real audio channel
- Choose HQ over LQ
- Each video: 512*384 resolution, 25 fps, ≈4s duration

**DFDC**[12]

- Undisclosed manipulations
- Audio/visual/audio-visual
- Each video: 30 fps, ≈ 10s duration

**Principle of data selection**

- Requirement: audio and visual channels
- DFTIMIT:
  - Frontal angles
  - 3 out of 32 subjects
  - Correct learning outcome
- DFDC:
  - Computational limitation
  - Same size
  - Different folders

# Technical Approach

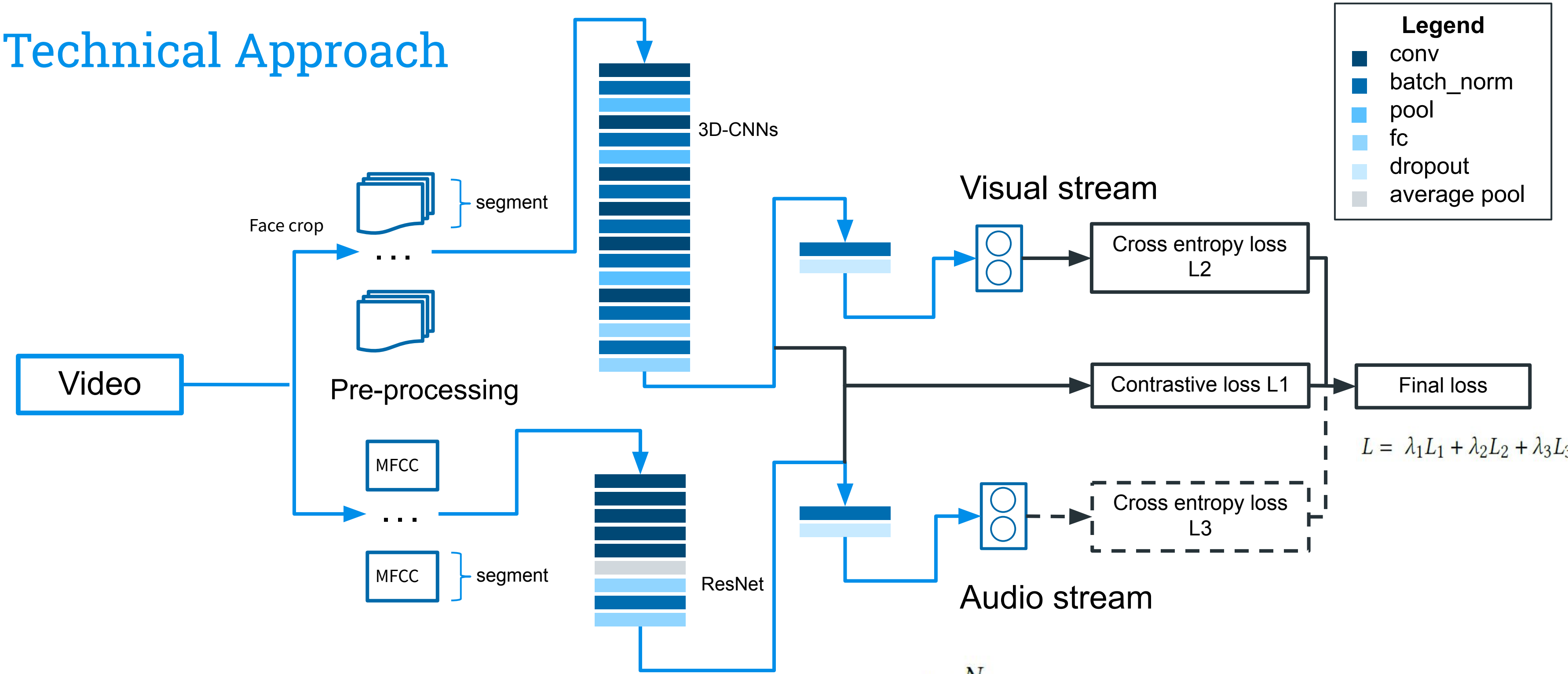
## Transition from midterm presentation

- Initially, Emotions Don’t Lie[5]
- Prob 1: Require the inputs of a real and fake video pair
- Prob 2: Framework too complicated

## Inspirations

- Directly model the similarity between visual and audio features
- Contrastive loss from Chung and Zisserman et al.[7]
- Imposition of cross-entropy loss from Chugh et al.[8]
- Extract visual features using 3D-CNNs[9]
- Extract audio MFCC[10] features using ResNet architectures

# Technical Approach




$$L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3$$

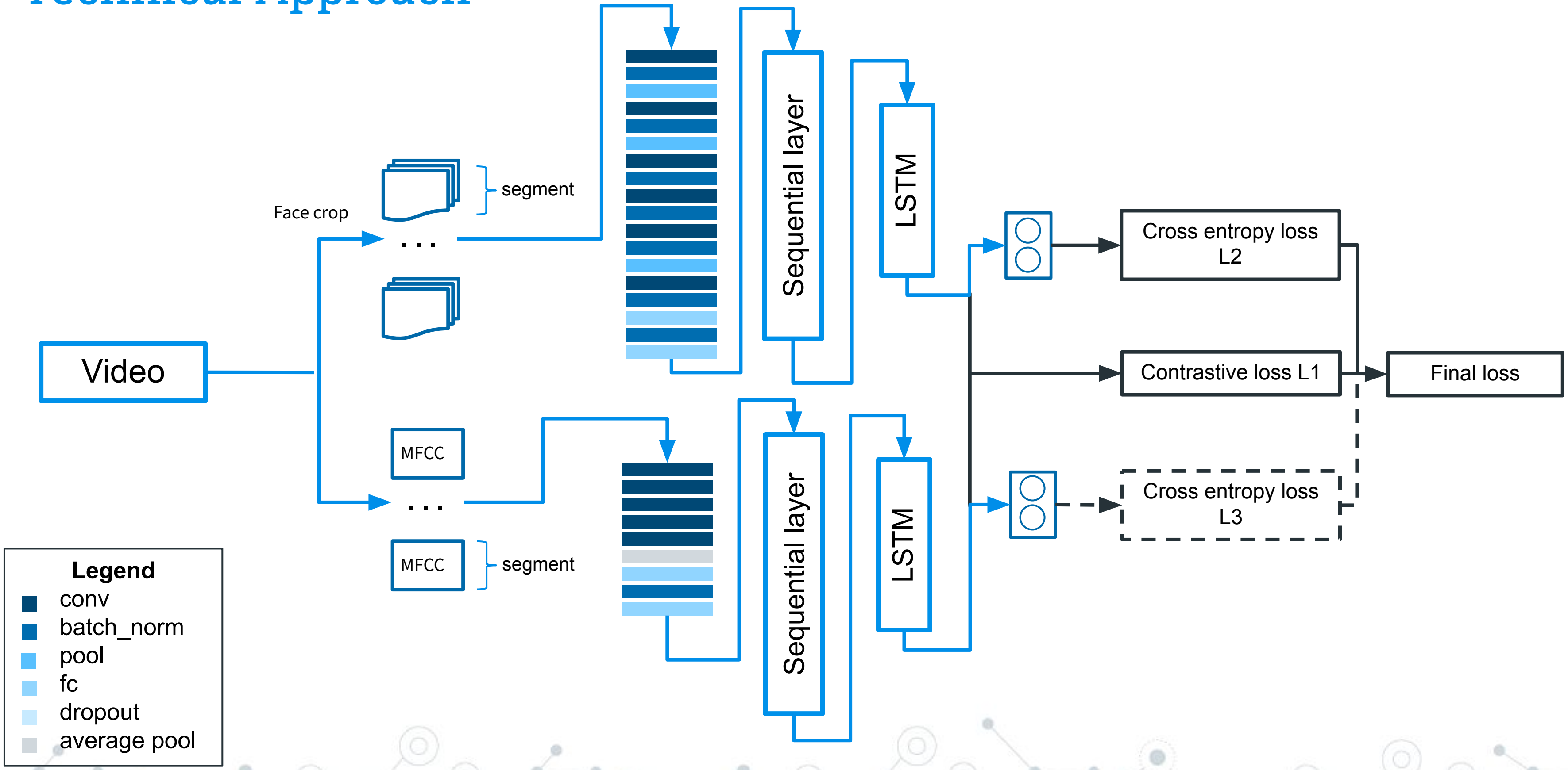
$$L_1 = \frac{1}{N}\sum_{i=1}^{N}(y^i)\,(d_t^i)^2 + (1-y^i)\,max(margin - d_t^i, 0)^2$$

$$d_t^i = \|f_v - f_a\|_2$$

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N} y^i \log \hat{y}_v^i + (1-y^i)\log(1-\hat{y}_v^i)$$

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N} y^i \log \hat{y}_a^i + (1-y^i)\log(1-\hat{y}_a^i)$$

# Technical Approach



# Platform

## Preprocessing

- 1-second segments (FFMpeg)
- S3FD to crop faces
- Python_speech_features for MFCC features

## Training and testing

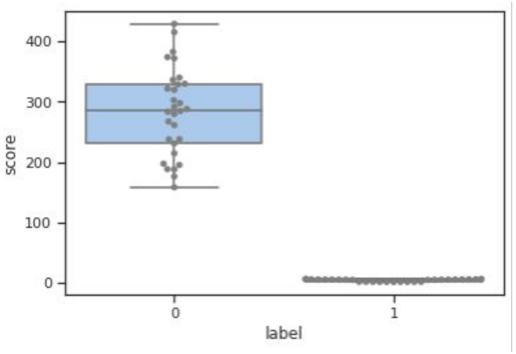
- PyTorch to build the network

## Runtime environment

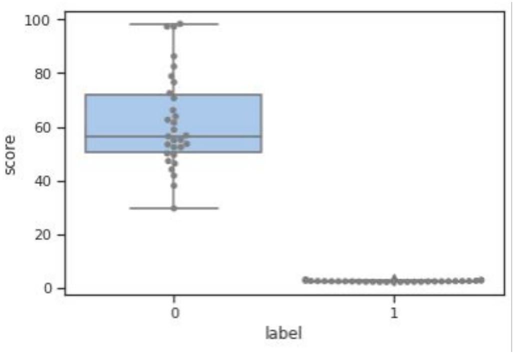
- Google CoLab using one GPU

# Evaluation

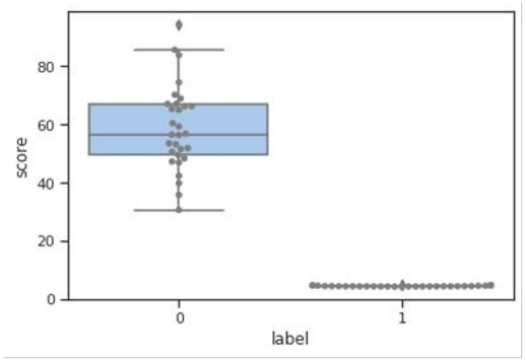
Results for DFTIMIT




L1+L2+L3



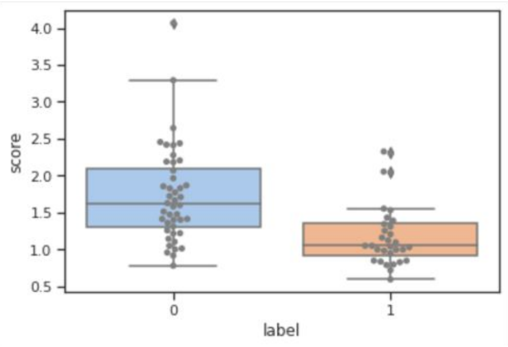

L1+L2




Only L1

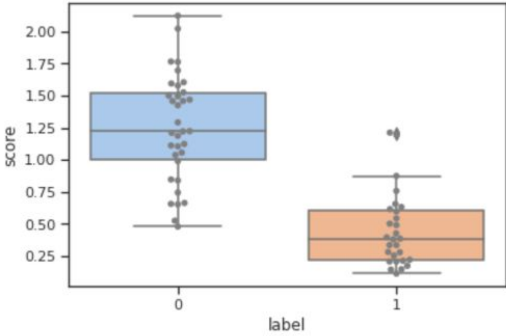
# Evaluation

Results for DFDC



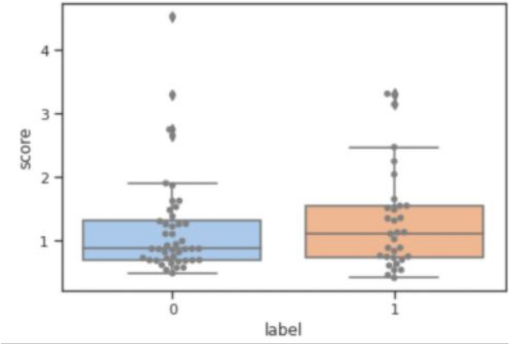

AUC:0.72

L1+L2+L3




AUC:0.83

L1+L2

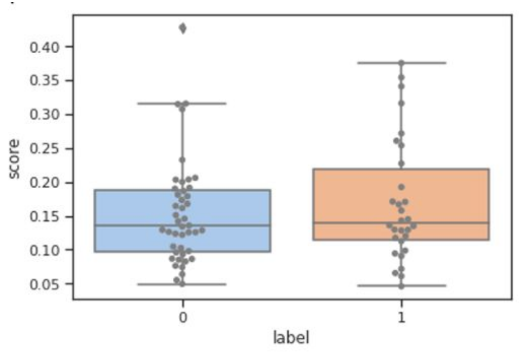


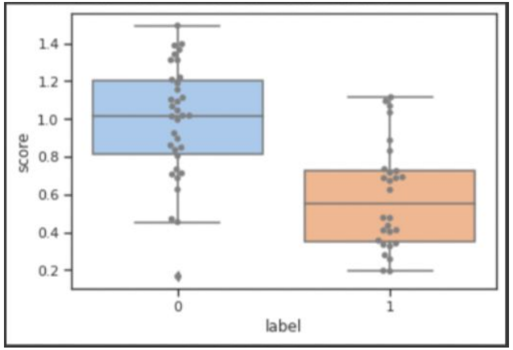

AUC:0.49

Only L1

# Evaluation

Results for DFDC (LSTM)




AUC:0.44

L1+L2+L3

AUC:0.78

L1+L2

# Evaluation

| | Capsule[15] | HeadPose[2] | VA-MLP[16] | FWA[1] | Siamese-based[5] | AV-dissonance[6] | Our method |
|---|---|---|---|---|---|---|---|
| DFDC | **53.3** | **55.9** | **61.9** | **72.7** | **84.4** | **90.55** | **83.19**<br>**77.52(LSTM)** |
| DFTIMIT-HQ | **74.4** | **53.2** | **62.1** | **93.2** | **94.9** | **96.8** | **99.9** |
| Modality | **V** | **V** | **V** | **V** | **AV** | **AV** | **AV** |

# Discussions

**Advantages**

- Achieved perfect performance on DFTIMIT and decent results on DFDC
- Analyzed the effects to the performance using different loss combinations

**Disadvantages**

- The size of training samples is relatively small due to computation limit
- The performance of the LSTM based network is not good as expected

## Future directions

- Test our models on bigger and more SOTA datasets
- Improve the architecture of the LSTM based network
- Try different advanced feature extraction networks and analyze the results
- Improve the techniques used in data preprocessing, e.g. face alignment, overlapping, etc.

# Team member contributions

- **Jinchen Wu:**
- Coding of entire projects
- Data cleaning
- Training and evaluating DFDC dataset
- Slides

- **Weijian Zhang:**
- Idea exploring
- Co-coding of LSTM network
- Training DFTIMIT dataset
- Training DFDC dataset using LSTM network
- Slides and report

- **Ruoye Wang:**
- Evaluating DFTIMIT dataset
- Slides and report

# Reference


[1] Yuezun Li and Siwei Lyu. 2018. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018). https://arxiv.org/abs/1811.00656

[2] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 8261–8265. https://arxiv.org/abs/1811.00661

[3] David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 1–6.

[4] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3 (2019), 1.

[5] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions Don't Lie: A Deepfake Detection Method using Audio-Visual Affective Cues. ArXiv, abs/2003.06711. https://arxiv.org/abs/2003.06711

[6] Hosler, Brian, et al. "Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[7] Joon Son Chung and Andrew Zisserman. 2017. Out of Time: Automated Lip Sync in the Wild. 251–263. https://doi.org/10.1007/978-3-319-54427-419

[8] Chugh, Komal, et al. "Not made for each other-audio-visual dissonance-based deepfake detection and localization." Proceedings of the 28th ACM international conference on multimedia. 2020.

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? CoRR abs/1711.09577 (2017). arXiv:1711.09577 http://arxiv.org/abs/1711.09577

[10] Nelson Mogran, Hervé Bourlard, and Hynek Hermansky. 2004. Automatic Speech Recognition: An Auditory Perspective. Springer New York, New York, NY, 309–338. https://doi.org/10.1007/0-387-21575-1_6

[11] Pavel Korshunov and Sébastien Marcel. 2018. Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018).

[12] BrianDolhansky,RussHowes,BenPflaum,NicoleBaram,andCristianCanton Ferrer. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv preprint arXiv:1910.08854 (2019).

[13] GitHub-shaoanlu/faceswap-GAN:Adenoisingautoencoder+adversarial losses and attention mechanisms for face swapping. https://github.com/shaoanlu/ faceswap-GAN. (Accessed on 02/16/2020).

[14] Conrad Sanderson. 2002. The vidtimit database. Technical Report. IDIAP.

[15] H. H. Nguyen, J. Yamagishi, and I. Echizen. 2019. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2307–2311.

[16] F. Matern, C. Riess, and M. Stamminger. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). 83–92.

# Thank you

Questions?