IEEE International Conference on Multimedia and Expo 2024
Niagara Falls Marriott, Niagara Falls, Canada
July 15-19, 2024

Multimedia Deepfake Detection

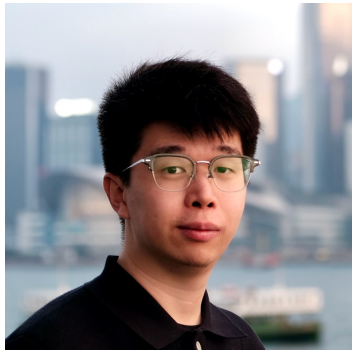
You (Neil) Zhang, Menglu Li, Luchuan Song
Zhiyao Duan, Xiao-Ping Zhang, Chenliang Xu

ICME 2024

2024-07-15

Speakers and Organizers

Speakers:

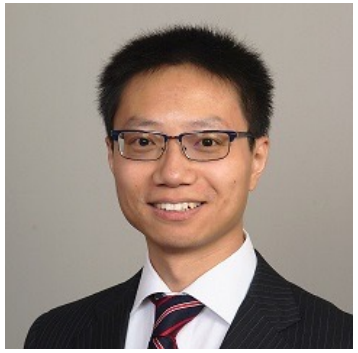


You (Neil) Zhang
University of Rochester

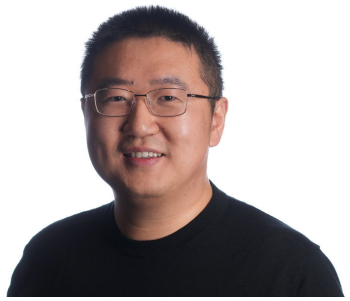
Luchuan Song
University of Rochester

Menglu Li
Toronto Metropolitan University

Organizers:



Zhiyao Duan
University of Rochester

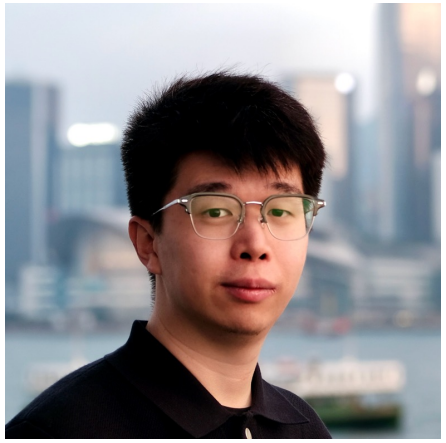


Chenliang Xu
University of Rochester

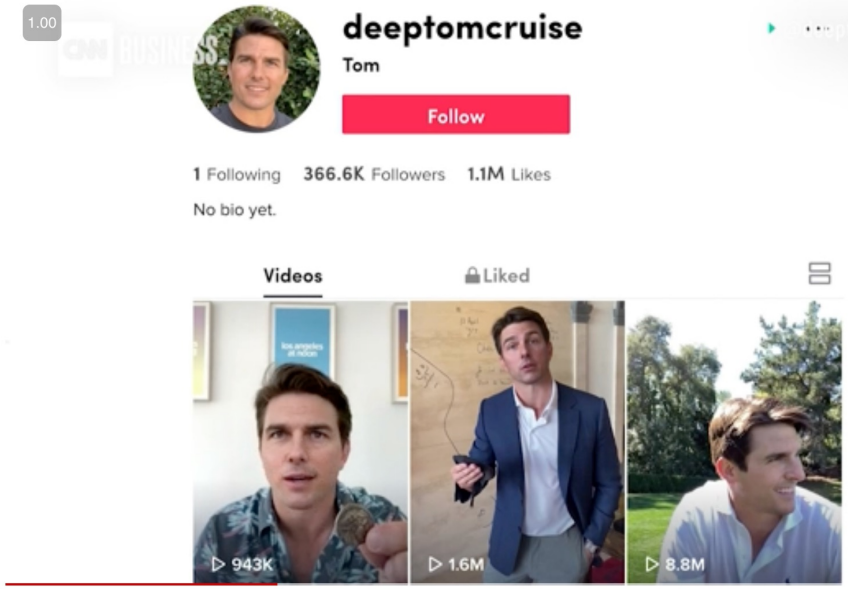
Xiao-Ping Zhang
Toronto Metropolitan University

Introduction

You (Neil) Zhang



Multimedia Deepfakes



No, Tom Cruise isn't on TikTok. It's a deepfake

https://www.youtube.com/watch?v=iyiOVUbsPcM

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

Martin Lewis felt 'sick' seeing deepfake scam ad on Facebook

7 July 2023

DIGITAL MUSIC NEWS CATEGORIES + SYNC NEWS JOBS + PODCASTS

Home > Music Industry News

AI Voice Tool Abused to Make Celebrity Deepfake Audio Clips

Ashley King February 1, 2023

Forbes

FORBES > INNOVATION > CYBERSECURITY

EDITORS' PICK

Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find

VIRAL TRENDS

AI clones teen girl's voice in $1M kidnapping scam: 'I've got your daughter'

By Ben Cost

Published April 12, 2023 | Updated April 12, 2023, 1:00 p.m. ET

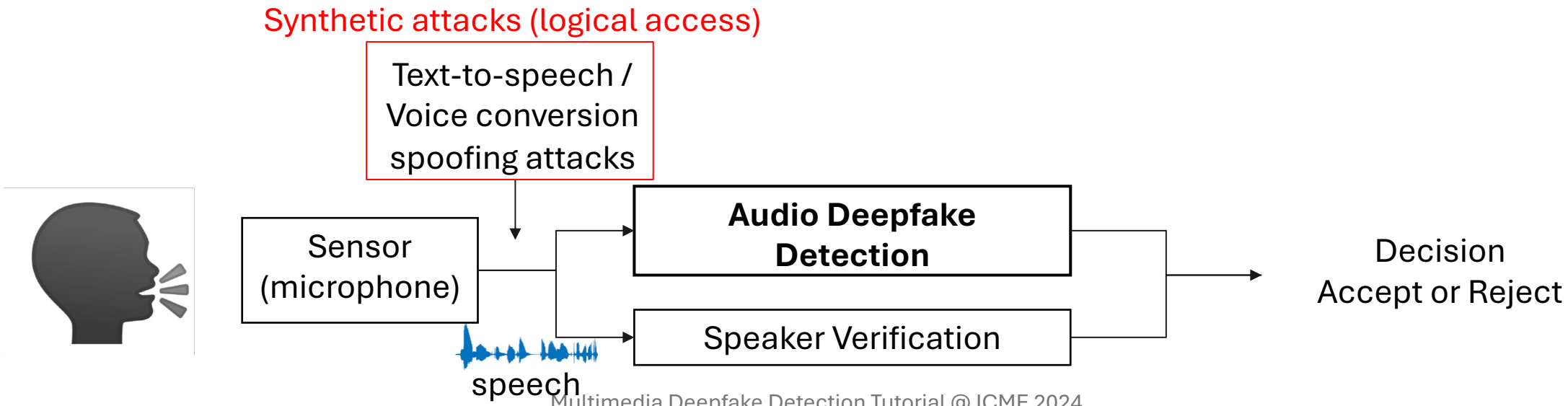
Background: Audio Deepfakes

Text-to-speech (TTS)

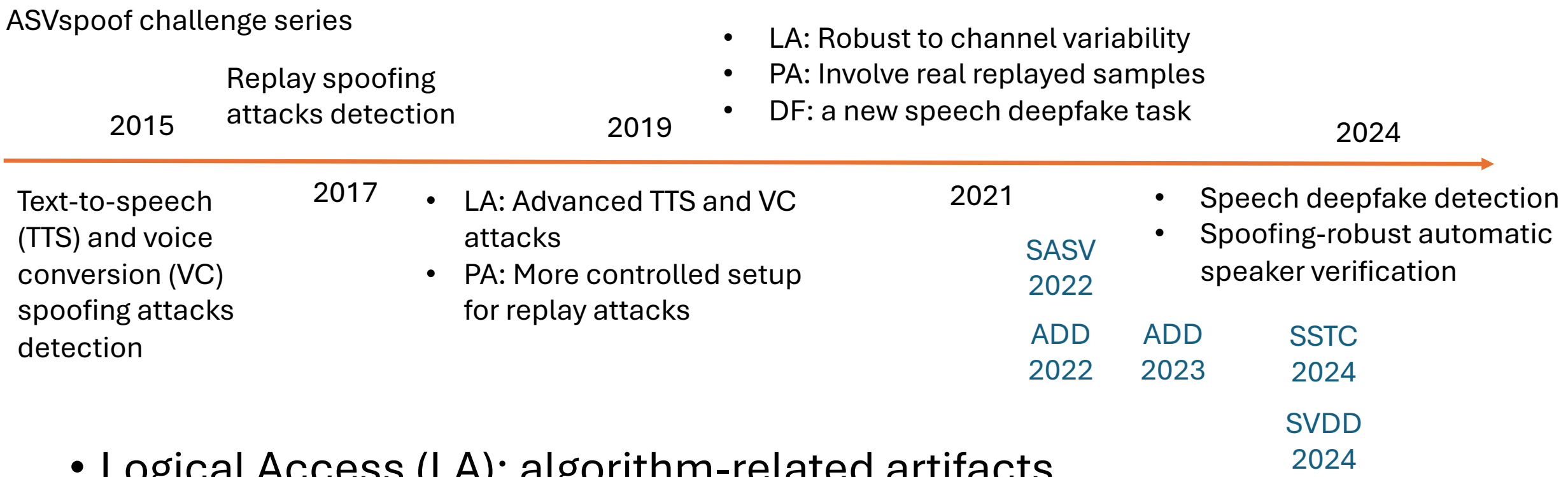
-- Convert written text into spoken words with speech synthesis

Voice conversion (VC)

-- Convert speech from source speaker to target speaker's voice



Background: Audio Deepfake Detection



- Logical Access (LA): algorithm-related artifacts
- Physical Access (PA): device-related artifacts

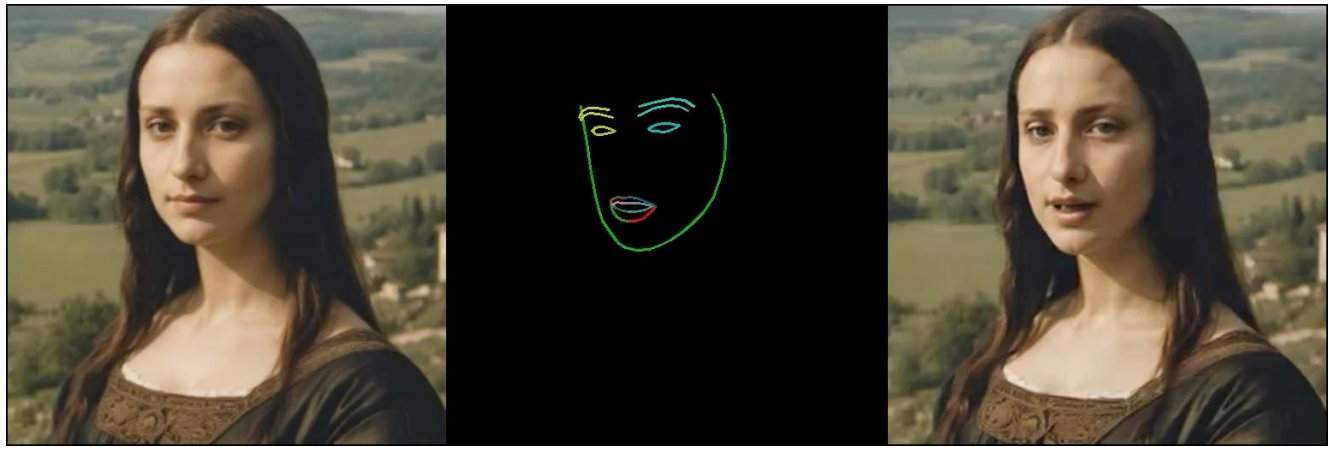
Spoofing-Aware Speaker Verification (SASV), Audio Deepfake Detection (ADD), Source Speaker Tracing Challenge (SSTC), Singing Voice Deepfake Detection (SVDD)

Background: Video Deepfakes

- Face swap

Video from https://www.media.io/faceswap.html.

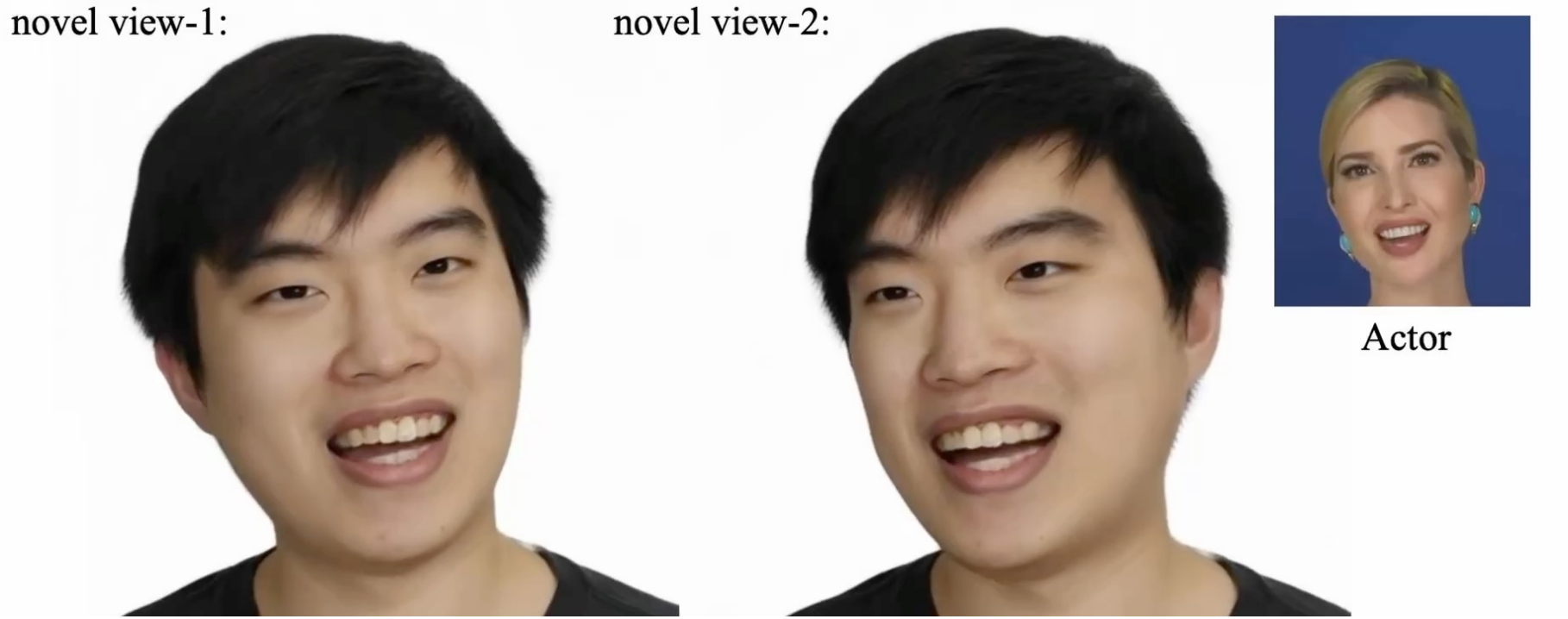
- Talking face generation



Wei H, Yang Z, Wang Z. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. arXiv 2024.

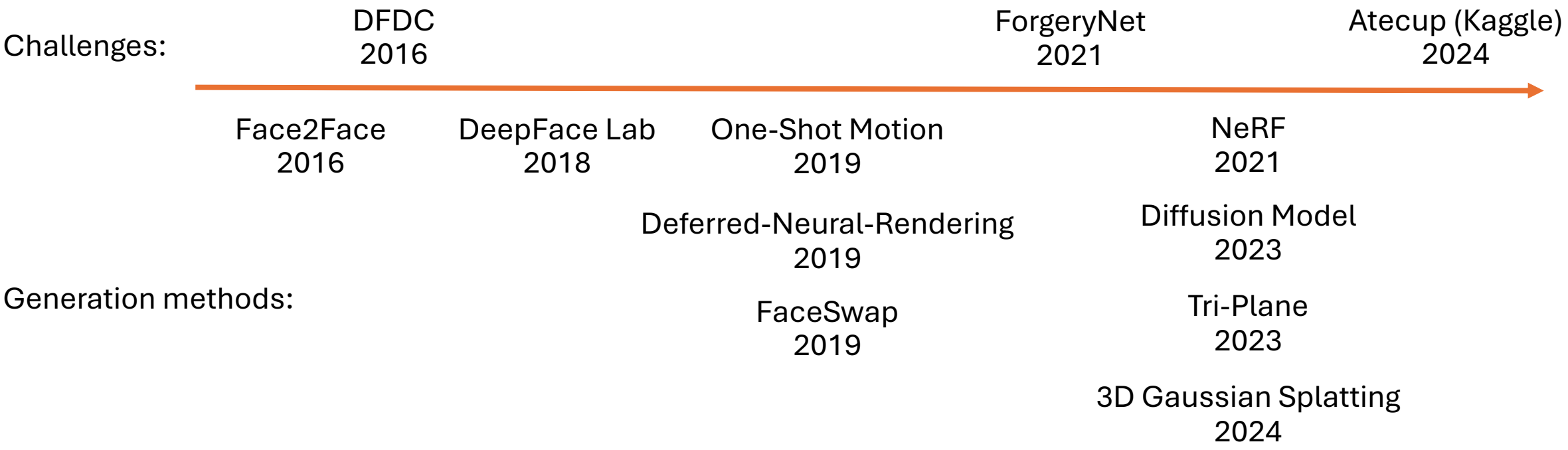
Background: Video Deepfakes

- Head-avatar generation (3D generation): Tri^{2}Plane [Song+2024]



Song L, Liu P, Chen L, et al. Tri^2-plane: Volumetric Avatar Reconstruction with Feature Pyramid. ECCV 2024.

Background: Video Deepfake Detection



Evaluation Metric: AUC vs. EER

Area Under the Curve (AUC)

- Widely used in Video Deepfake Detection

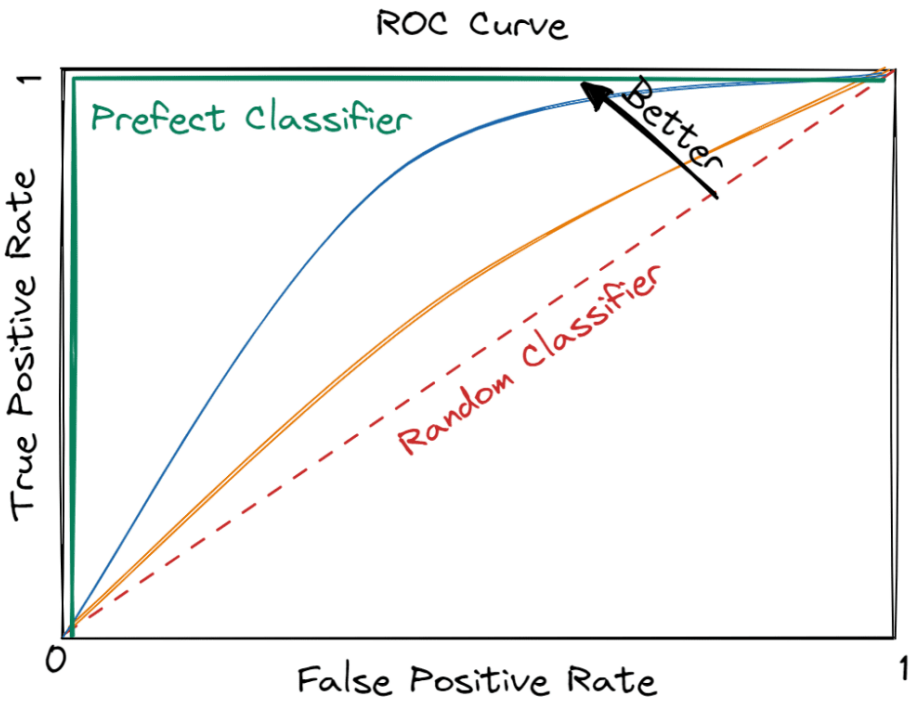


Figure from https://www.kdnuggets.com/2022/10/metric-accuracy-auc.html

Equal Error Rate (EER)

- Widely used in Audio Deepfake Detection

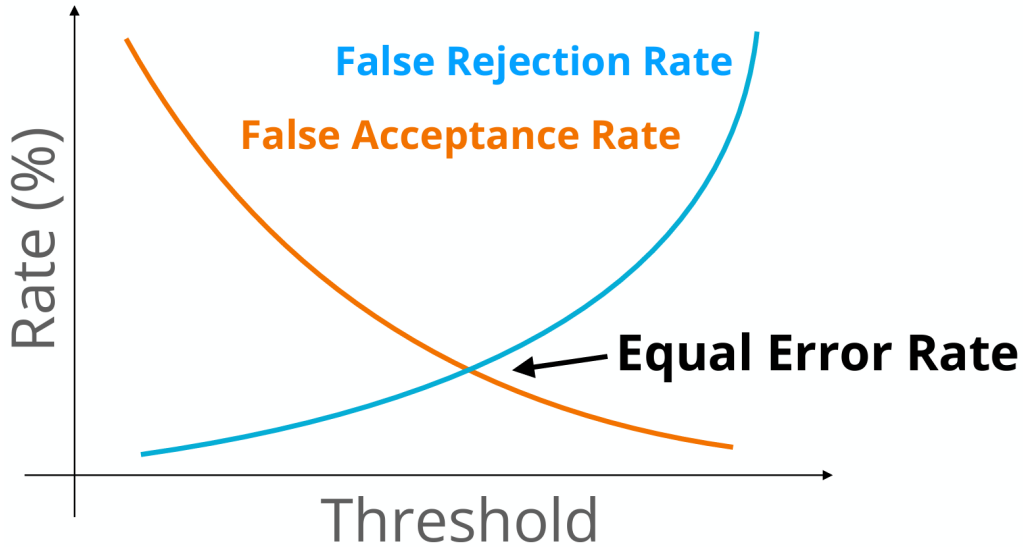


Figure from SVDD challenge evaluation plan

Advanced Evaluation Metrics

- T-DCF [Kinnunen+2018]: assess the influence of CM on the ASV system
- T-EER [Kinnunen+2023]: parameter-free tandem evaluation
- A-DCF [Shim+2024]: architecture agnostic metric for spoofing-robust speaker verification

Used in Audio Deepfake Detection but can be generalized to video and multimedia

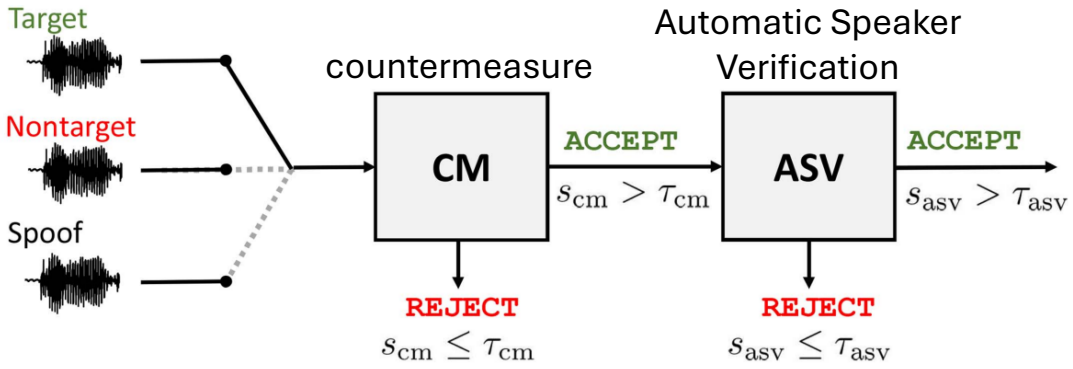


Figure from [Kinnunen+2023]

Kinnunen, Tomi, et al. "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification." *Speaker Odyssey* 2018.
Kinnunen, Tomi H., et al. “t-EER: Parameter-free tandem evaluation of countermeasures and biometric comparators.” *TPAMI* 2023.
Shim, Hye-jin, et al. “a-DCF: an architecture agnostic metric with application to spoofing-robust speaker verification.” *Speaker Odyssey* 2024.

Goal of this Tutorial

- Introduce the latest developments in audio, video, and audio-visual deepfake detection, understand the prevailing challenges, and highlight promising directions for future research.
- Bridge the gap among the research communities for single-modality deepfake detection
- Foster discussion and collaboration towards multimedia deepfake detection

Scope of this Tutorial

- Audio (speech) deepfake detection
- Video deepfake detection
- Audio-visual deepfake detection

- Outside the scope (emerging topics):
  - Singing voice deepfake detection [Zang+2024]
  - General audio deepfake detection [Xie+2024]
  - Text deepfake detection [Yang+2024]

Zang, Yongyi, et al. "Singfake: Singing voice deepfake detection." *ICASSP* 2024.
Xie, Zeyu, et al. "FakeSound: Deepfake General Audio Detection." *Interspeech* 2024.
Yang, Xianjun, et al. "DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text." *ICLR* 2024.

Outline

- Introduction (20 min)
- Audio Deepfake Detection (45 min)
- Break (25 min)
- Video Deepfake Detection (45 min)
- Audio-Visual Deepfake Detection (30 min)
- Q&A (15 min)

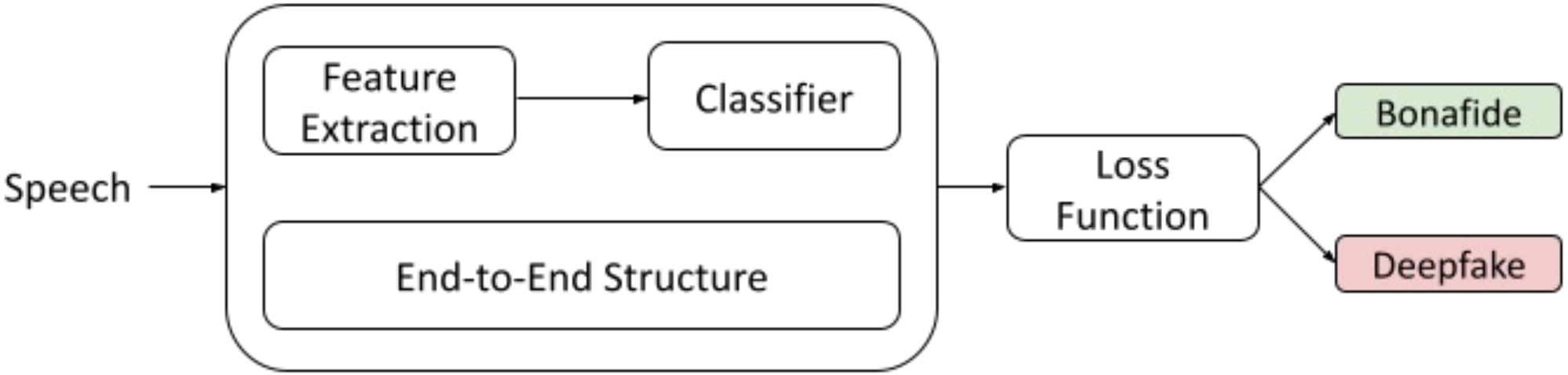
Audio Deepfake Detection

Menglu Li

Audio Deepfake Detection

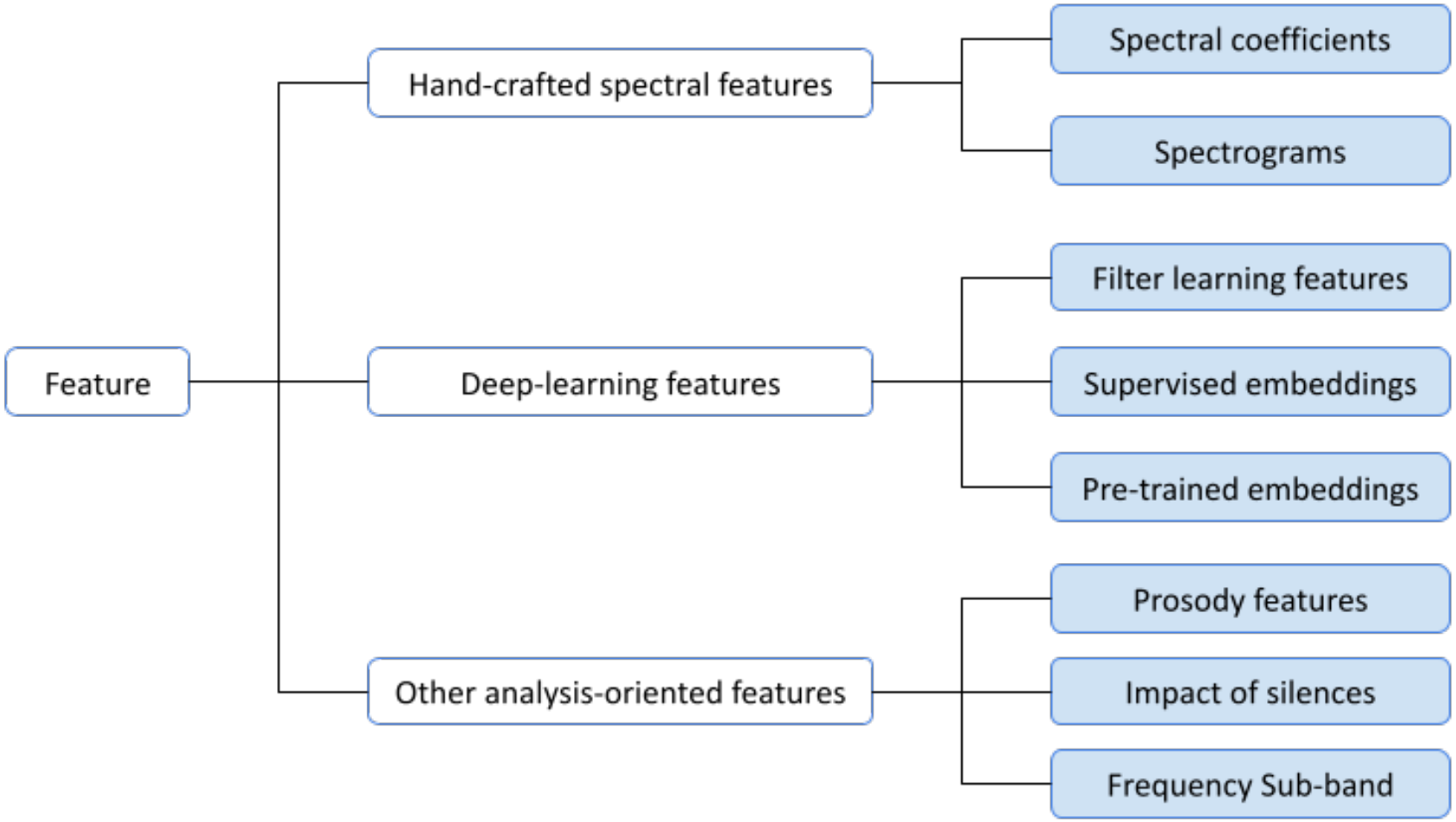
- Key Components of Detection Models
- Advanced Topics in Detection Model Developments
- Summary and Future Directions

Key Components of Detection Models



- Feature Extraction: speech signals -> acoustic features
- Classifier: acoustic feature -> detection decision
- E2E: speech signals -> detection decision

Feature Extraction



Feature Extraction: Hand-crafted features

- Short-time Spectral Coefficients
  - Such as Linear Frequency Cepstral Coefficient (LFCC) [Alegre+, 2013] and Mel-Frequency Cepstral Coefficient (MFCC) [Sahidullah+, 2015]
  - Speech -> Windowing -> DFT -> Filter banks -> Log -> DCT -> Coefficients

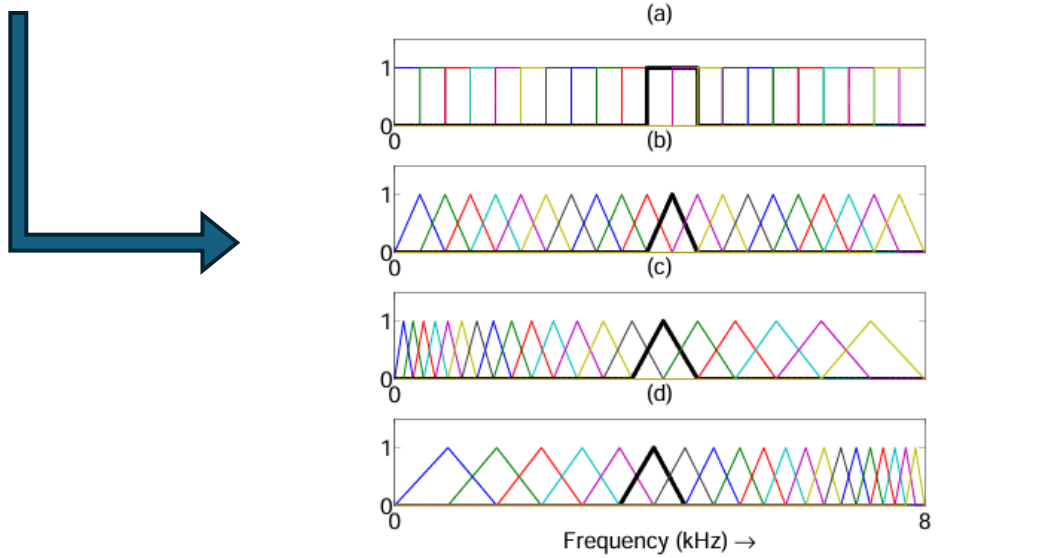
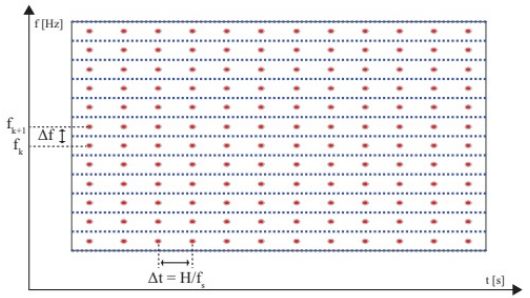


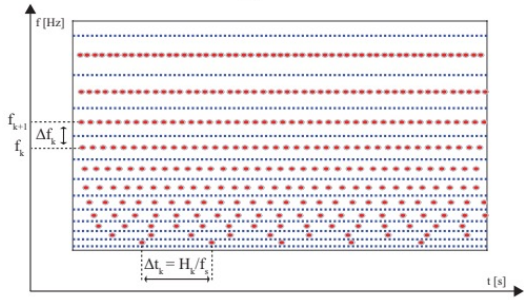
Figure 1: *Figure showing filter bank used in the computation of (a) RFCC, (b) LFCC, (c) MFCC, and (d) IMFCC.*

Feature Extraction: Hand-crafted features

- Long-time Spectral Coefficients
  - Constant Q Cepstral Coefficient (CQCC) [Tak+, 2020]
  - Higher temporal resolution at higher frequencies
  - Higher frequency resolutions at lower frequencies

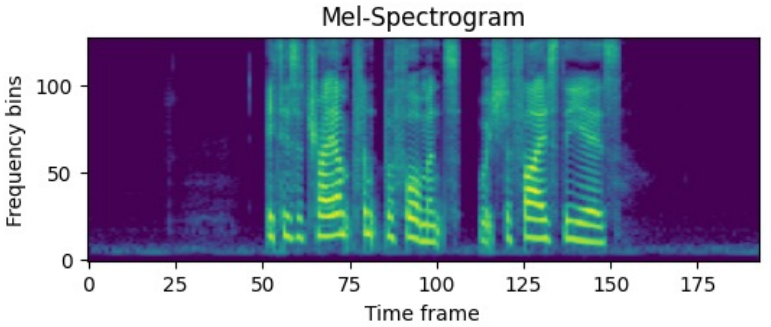


(a) FFT



(b) CQT

- Spectrograms
  - Treated as 2D images
  - Mel-spectrogram [Ray+, 2021], CQT-spectrogram [Abdzadeh and Veisi, 2023]
  - Include information regarding frequencies and intensities of the speech signal as it propagates in time.



Feature Extraction: Deep-learning features

- Filter Learning Features
  - Approximate the standard filtering process
  - Most widely-used: SincNet [Zeinali+, 2019]
  - Other works: nnAudio [Cheuk+, 2020] , FastAudio [Fu+, 2022]
- Supervised Embeddings
  - CNN [Wu+, 2020] , ResNet [Shim+, 2022] , Bi-LSTM [Khan+, 2024]
- Pre-trained Embeddings
  - Self-supervised models: wav2vec2.0 [Wang and Yamagishi, 2022] , wavLM [Zhu+, 2023] , Hubert [Li+, 2023]
  - Finetuning along with the classifier

Feature Extraction: Other Directions

- Prosody / Semantic Embeddings [Conti+, 2022] [Wang+, 2023]
  - More effective for TTS-generated Deepfake speech, rather than VC-based Deepfakes
- Impact of Silence [Zhang+ 2023]
  - Duration proportion of silence -> TTS Deepfakes
  - Content of silence -> VC Deepfakes
- Frequency Sub-band Features [Pillai+, 2022]
  - Low-frequency band of 0-4kHz -> voiced segments
  - High-frequency band of 4-8kHz -> silence and unvoiced segment

Classifier

Category	Advantages	Disadvantages	Methods
Traditional ML	Light-weight; facilitating easier interpretation of the distribution outcomes	Poor generalization performance on unseen attacks	GMM [269], RF [91], SVM [29]
CNN	Light-weight; Producing promising detection performance	Causing information loss in the frequency domain due to the translation invariant property	LCNN [113], Non-OFD [39], CapsuleNet [147]
ResNet	Enabling architectural adjustments for modifying receptive fields; enhancing generalizability to unseen attacks; accommodating deeper networks	High computational cost; The performance can be highly varied by feature selection	ResNet [7], SE-Net [112], ResMax [110], ResNext [296], Res2Net [128], DenseNet [234], xResNet [25]
GNN	Aggregating all note features for message passing; enhancing the formulation of inter-relationships among frame-level features	Challenging to construct a deep network; high time and space complexity	RawGAT [204], AASIST [92], GCN [32]
Transformer	Effectively capturing long-term dependencies	Potential for overfitting; high computational costs	CCT [18], OCT [117], TFT [235], Rawformer [139]
TDNN	Lightweight; allowing varying input lengths	Unsatisfactory detection performance	ECAPA-TDNN [34], AF-TDNN [243]
DART	Enabling architecture optimization during back-propagation	Performance may be influenced by predefined hyperparameters	PC-PARTS [70], Raw PC-PARTS [71], light-DARTS [217]

Classifier: CNN-based

- Light-CNN [Lavrentyeva+, 2019]
  - Replace ReLU with Max-Feature-Map activation
- Translation invariance property of CNN
  - Sub-band CNN: Split the spectrogram inputs along the frequency axis [Choi+, 2022]

Type	Filter / Stride	Output	Params
Conv_1	5 × 5 / 1 × 1	863 × 600 × 64	1.6K
MFM_2	–	864 × 600 × 32	–
MaxPool_3	2 × 2 / 2 × 2	431 × 300 × 32	–
Conv_4	1 × 1 / 1 × 1	431 × 300 × 64	2.1K
MFM_5	–	431 × 300 × 32	–
BatchNorm_6	–	431 × 300 × 32	–
Conv_7	3 × 3 / 1 × 1	431 × 300 × 96	27.7K
MFM_8	–	431 × 300 × 48	–
MaxPool_9	2 × 2 / 2 × 2	215 × 150 × 48	–
BatchNorm_10	–	215 × 150 × 48	–
Conv_11	1 × 1 / 1 × 1	215 × 150 × 96	4.7K
MFM_12	–	215 × 150 × 48	–
BatchNorm_13	–	215 × 150 × 48	–
Conv_14	3 × 3 / 1 × 1	215 × 150 × 128	55.4K
MFM_15	–	215 × 150 × 64	–
MaxPool_16	2 × 2 / 2 × 2	107 × 75 × 64	–
Conv_17	1 × 1 / 1 × 1	107 × 75 × 128	8.3K
MFM_18	–	107 × 75 × 64	–
BatchNorm_19	–	107 × 75 × 64	–
Conv_20	3 × 3 / 1 × 1	107 × 75 × 64	36.9K
MFM_21	–	107 × 75 × 32	–
BatchNorm_22	–	107 × 75 × 32	–
Conv_23	1 × 1 / 1 × 1	107 × 75 × 64	2.1K
MFM_24	–	107 × 75 × 32	–
BatchNorm_25	–	107 × 75 × 32	–
Conv_26	3 × 3 / 1 × 1	107 × 75 × 64	18.5K
MFM_27	–	107 × 75 × 32	–
MaxPool_28	2 × 2 / 2 × 2	53 × 37 × 32	–
FC_29	–	160	10.2 MM
MFM_30	–	80	–
BatchNorm_31	–	80	–
FC_32	–	2	64
Total	–	–	371K

Classifier: ResNet-based

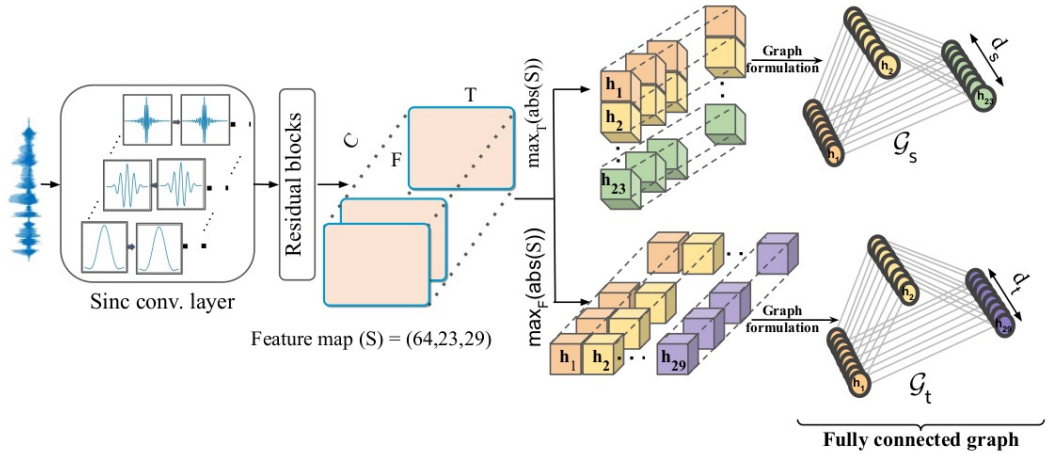
- RawNet2 [Tak+, 2021]
  - SincNet feature extractor
  - One of an official baseline in the ASVspoof challenge series.

- Other works
  - Adding SE [Lai+, 2019] or MFM [Kwak+ 2020] components to ResNet
  - Modify the bottleneck: Res2Net [Li+, 2021]

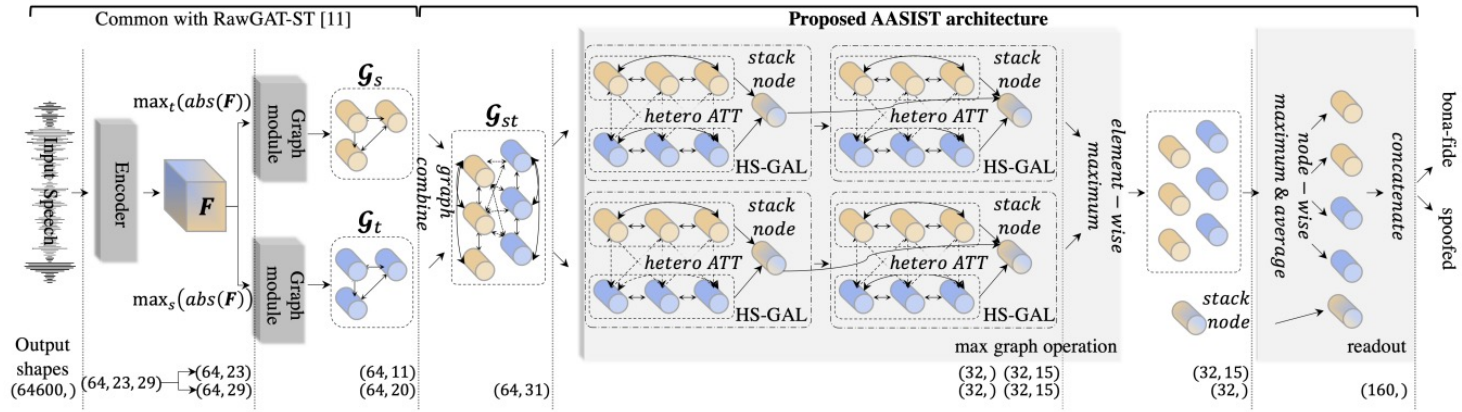
Layer	Input:**64000** samples	Output shape
Fixed Sinc filters	Conv(**129**,1,128) Maxpooling(3) BN & LeakyReLU	(**21290**,128)
Res block	{ BN & LeakyReLU Conv(3,1,128) BN & LeakyReLU Conv(3,1,128) Maxpooling(3) FMS } × 2	(**2365**,128)
Res block	{ BN & LeakyReLU Conv(3,1, **512**) BN & LeakyReLU Conv(3,1, **512**) Maxpooling(3) FMS } × 4	(**29**,512)
GRU	GRU(1024)	(1024)
FC	1024	(1024)
Output	**1024**	**2**

Classifier: Graph Neural Network (GNN)-based

- RawGAT [Tak, Jung+, 2021]
  - Form two fully-connected sub-graphs
  - Node: frequency bins and time frames
  - Graph attention mechanism



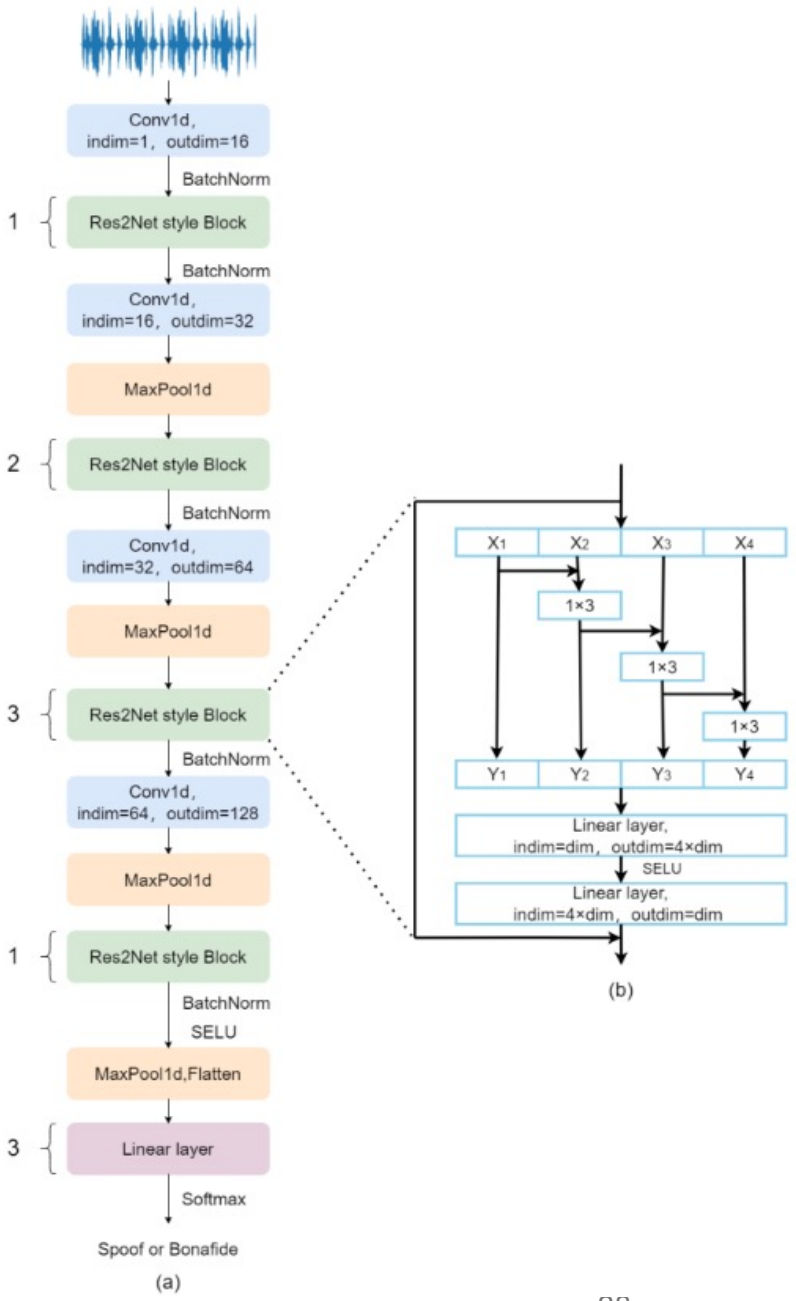
- AASIST [Jung+, 2022]
  - Heterogeneity-aware technique: integrate spectral and temporal sub-graphs
  - Aggregate information from all other spectral nodes and temporal nodes
  - Achieve 0.83% EER in ASVSpoof2019-LA dataset



End-to-End Structure

- Motivations
  - Reduce dependency on extracted features
  - Avoid information loss

- Takes raw audio as input
  - Utilize SincNet with pre-configured setting
    - RawNet2, RawGAT, ASSIST
  - Fully E2E: use 1D convolutional layer [Ma+, 2022]

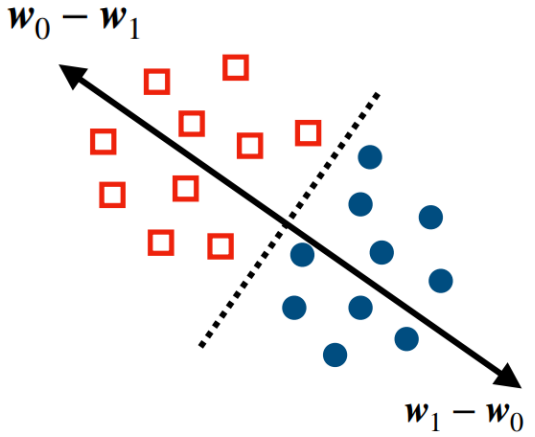


Loss Function

- Binary Cross-Entropy Loss with Softmax [Tak+, 2020]

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\boldsymbol{w}_{y_i}^{\top}\boldsymbol{x}_i}}{e^{\boldsymbol{w}_{y_i}^{\top}\boldsymbol{x}_i}+e^{\boldsymbol{w}_{1-y_i}^{\top}\boldsymbol{x}_i}}$$

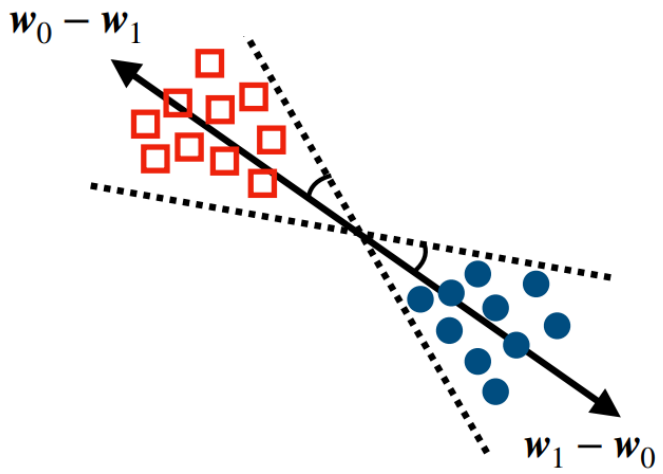
$$= -\frac{1}{N}\sum_{i=1}^{N}\log(1+e^{(\boldsymbol{w}_{1-y_i}-\boldsymbol{w}_{y_i})^{\top}\boldsymbol{x}_i}),$$



- Large Margin Cosine Loss [Chen+, 2020]

$$\mathcal{L}_{LMCL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^{\top}\hat{\boldsymbol{x}}_i-m)}}{e^{\alpha(\hat{\boldsymbol{w}}_{y_i}^{\top}\hat{\boldsymbol{x}}_i-m)}+e^{\alpha(\hat{\boldsymbol{w}}_{1-y_i}^{\top}\hat{\boldsymbol{x}}_i)}}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\log(1+e^{\alpha(m-(\hat{\boldsymbol{w}}_{y_i}-\hat{\boldsymbol{w}}_{1-y_i})^{\top}\hat{\boldsymbol{x}}_i)}).$$

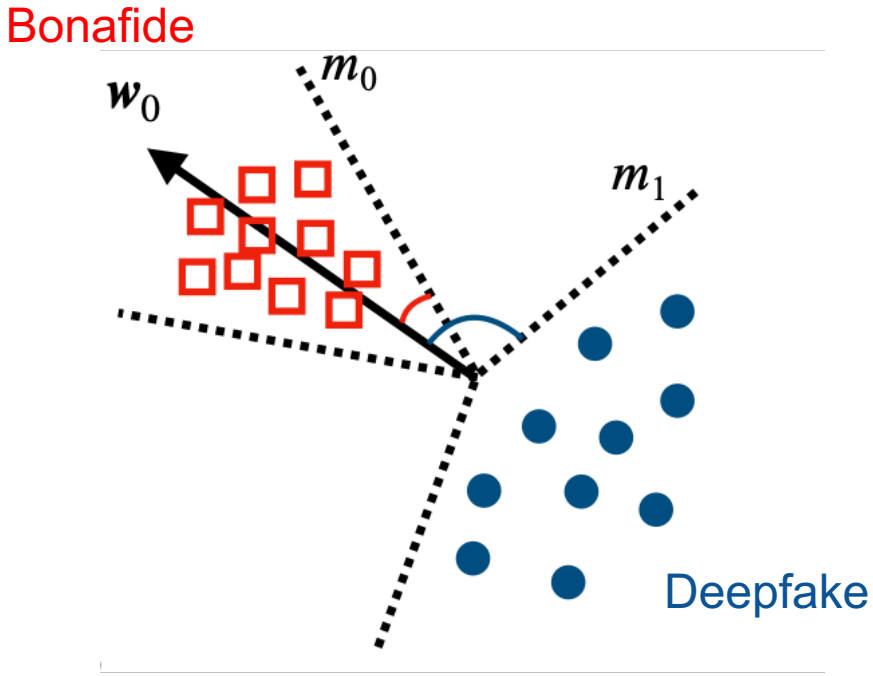


  - Compact embedding space
  - Same margin added to both classes

Loss Function

- One Class-Softmax [Zhang+, 2021]
  - Compact the Bonafide speech embeddings
  - Isolate Deepfake embeddings

$$\mathcal{L}_{OC} = -\frac{1}{N}\sum_{i=1}^{N}\log(1+e^{\alpha(m_{y_i}-\hat{w}_0\hat{x}_i(-1)^{y_i})}).$$



Advanced Topics in Audio Deepfake Detection

- Robustness in Cross-datasets
- Partially Deepfake Detection
- Integration of Deepfake countermeasures (CM) with audio speech verification systems (ASV)
- Interpretability of detection models

Advanced Topics: Cross-dataset Robustness

- Reveal a significant performance decline in cross-dataset

Publication		Data augmentation	Feature	Classifier	Loss funcion	ASVspoof 19-LA	ITW
[253]	IH&MMSec'23	w/o	Mel-Spec	Patched Transformer	CE	4.54	29.72
[218]	INTERSPEECH'23	w/o	Duration + pronunciation + wav2vec2.0-XLSR	LCNN →Bi-LSTM →MLP	CE	1.58	36.84
[248]	INTERSPEECH'23	w/o	wav2vec2.0-XLSR	LCNN →Transformer	CE, Triplet, Adversarial	0.63	24.50
[230]	ICASSP'23	w/o	wav2vec2.0-XLSR	MLP	CE	2.98	26.65
[289]	SPL'24	SpecAugment	ECAPA-TDNN	CNN→GRU →MLP	AM-Softmax	1.79	29.66
[263]	ICASSP'24	w/o	wav2vec2.0-XLSR	ResNet-18	CE	2.07	29.19
[263]	ICASSP'24	w/o	Hubert	ResNet-18	CE	6.78	27.48
[216]	ICASSP'24	w/o	Multi-scale permutation entropy	SE-ResNet	CE	20.24	29.62
[145]*	ICASSP'24	w/o	CNN →wav2vec2.0	AASIST	CE	0.39	**7.68**
[231]*	ICASSP'24	Rawboost	wav2vec2.0-XLSR-Vox	MLP	CE	**0.13**	12.50

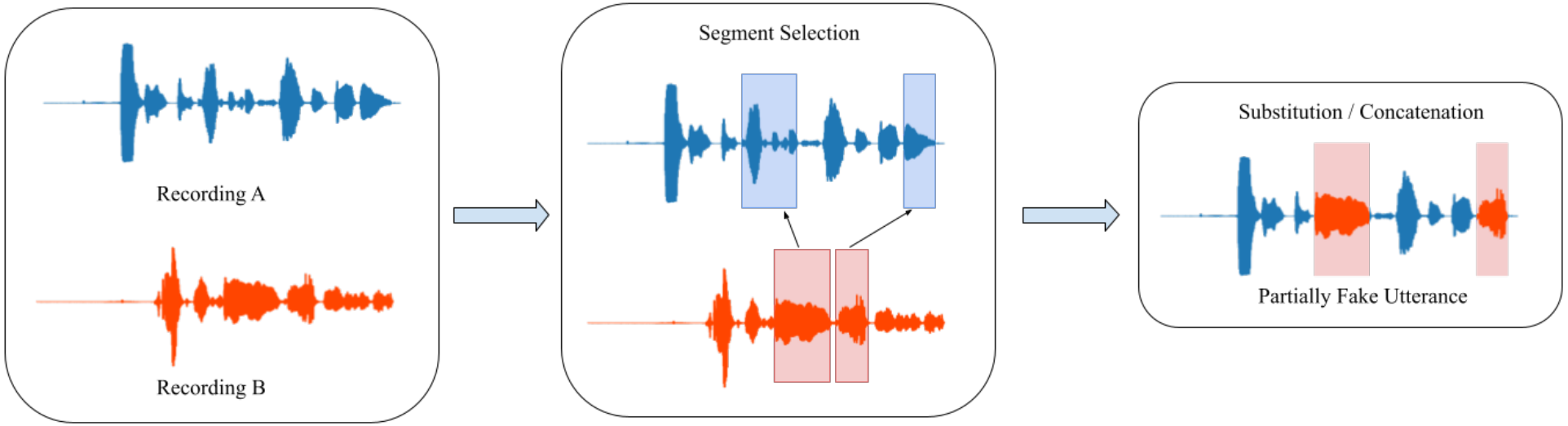
The evaluation metric is EER (%). The bold values refer to the best performance on the same dataset. "+" indicates multiple techniques processed in parallel, while "→" denotes sequential order. "w/o" means that no data augmentation techniques are applied.
* [145] and [231] utilize knowledge distillation. The reported evaluation results on both datasets are produced by the student model.

- [Zhang+, 2022] suggests the performance degradation may be due to the channel effect mismatch among different datasets -> Gradient Reversal Layer
- Knowledge Distillation (KD) technique [Lu+, 2024] [Wang and Yamagishi, 2024]

Advanced Topics: Partially Deepfake Detection

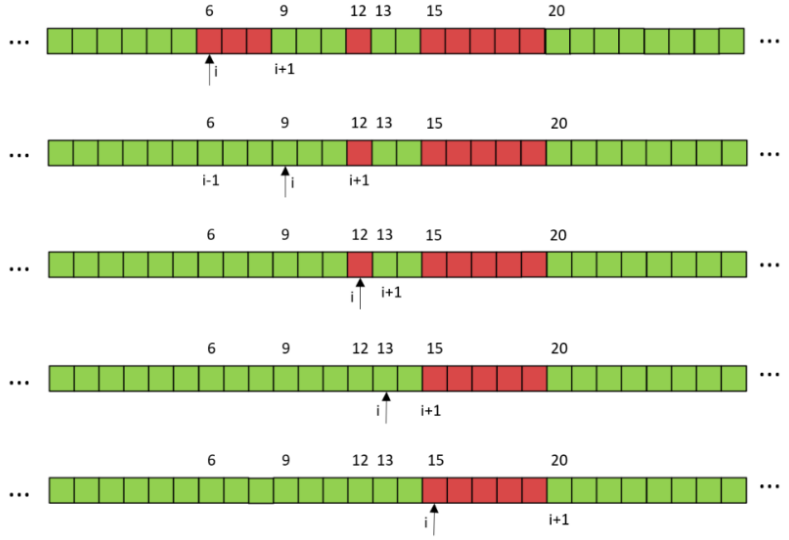
- Frame-level Detection

- Boundary Detection



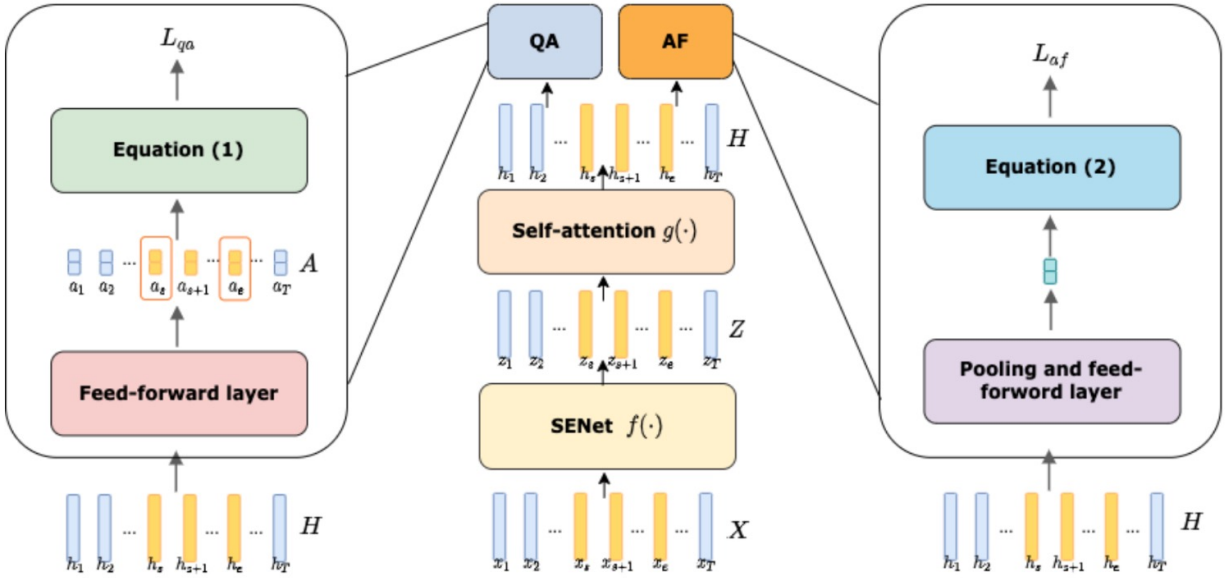
Advanced Topics: Partially Deepfake Detection

- Frame-level Detection
  - Divide speech into frames
  - Label each frames as Bonafide or Deepfake
  - Expect a Deepfake segment to be longer than the duration of a phoneme
  - Need Swapping Post-Processing [Zhang and Sim, 2022]
  - Other works
    - Isolated-frame penalty term [Liu+, 2023]



Advanced Topics: Partially Deepfake Detection

- Boundary Detection
  - Identify the transition boundaries between Bonafide and Deepfake segments
  - Eliminate the post-processing
  - One solution: add a QA layer [Wu+, 2022]

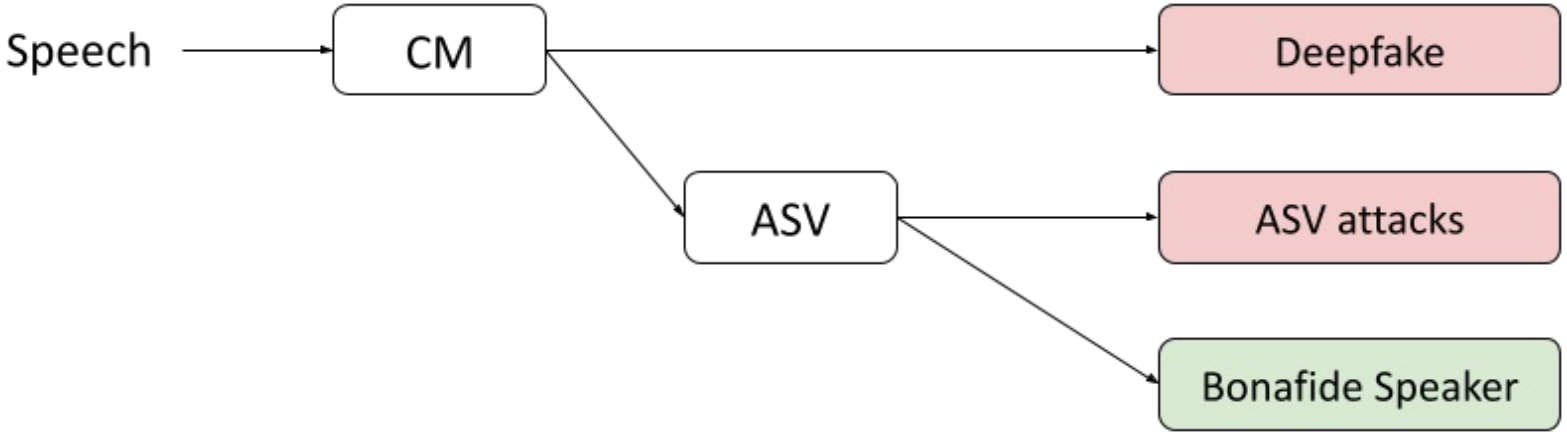


Advanced Topics: Partially Deepfake Detection

- PartialSpoof [Zhang, Wang+, 2022] - Publicly Available
- Psynd [Zhang and Sim, 2022] - Restricted
- ADD2022 [Yi+, 2022] - Restricted
- ADD2023 [Yi+, 2023] - Restricted

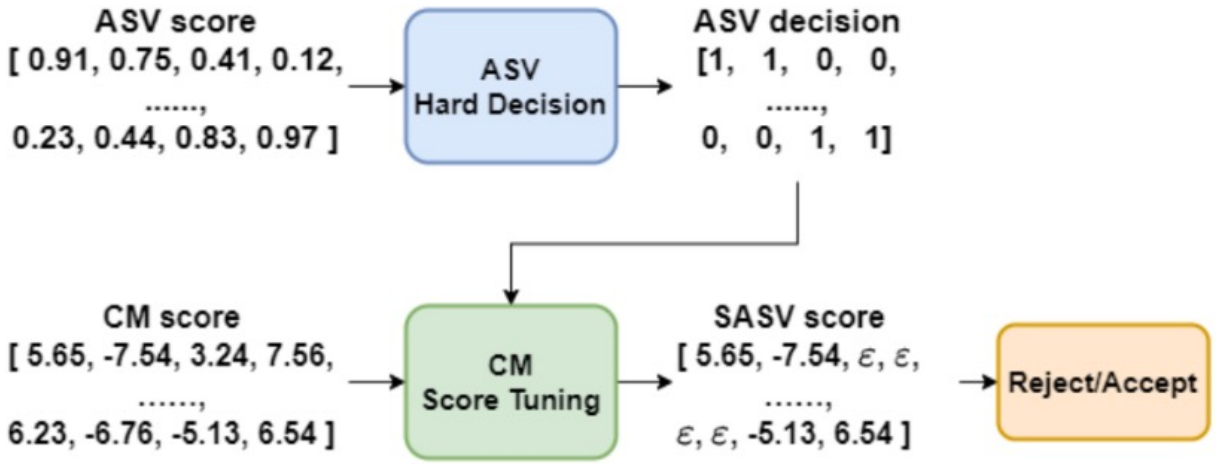
Advanced Topics: Integration of CM and ASV

- Deepfake Countermeasure (CM): detect Deepfake speech
- Audio Speaker Verification (ASV): verify the identity of speakers
- SASV: Spoofing-aware speaker verification



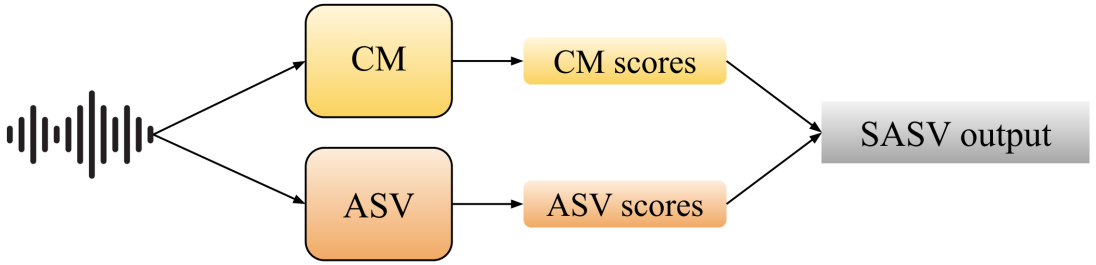
Advanced Topics: SASV

- Cascaded System [Wang+, 2022]
  - Concatenate the ASV and CM classifiers
  - Pre-train separately
  - The order of ASV and CM may affect the performance.

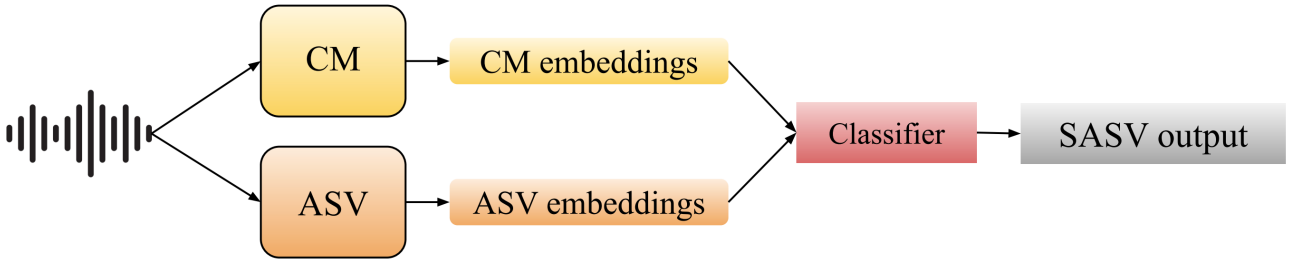


Advanced Topics: SASV

- Score-level Fusion System
  - Pre-train separately
  - May suffer from the disparity in scale ranges between ASV and CM scores.

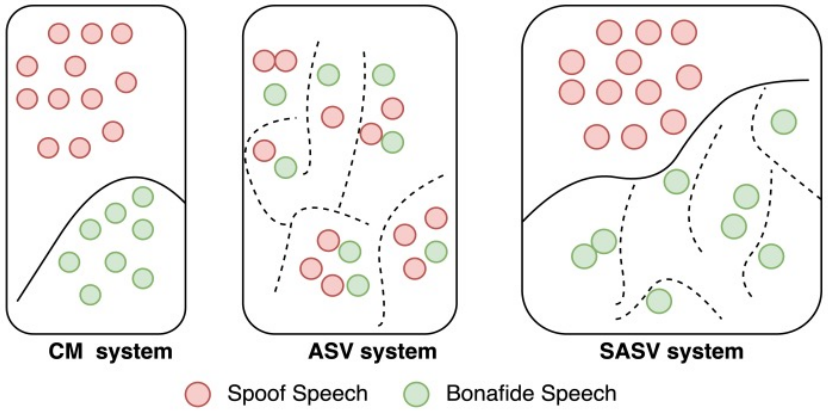


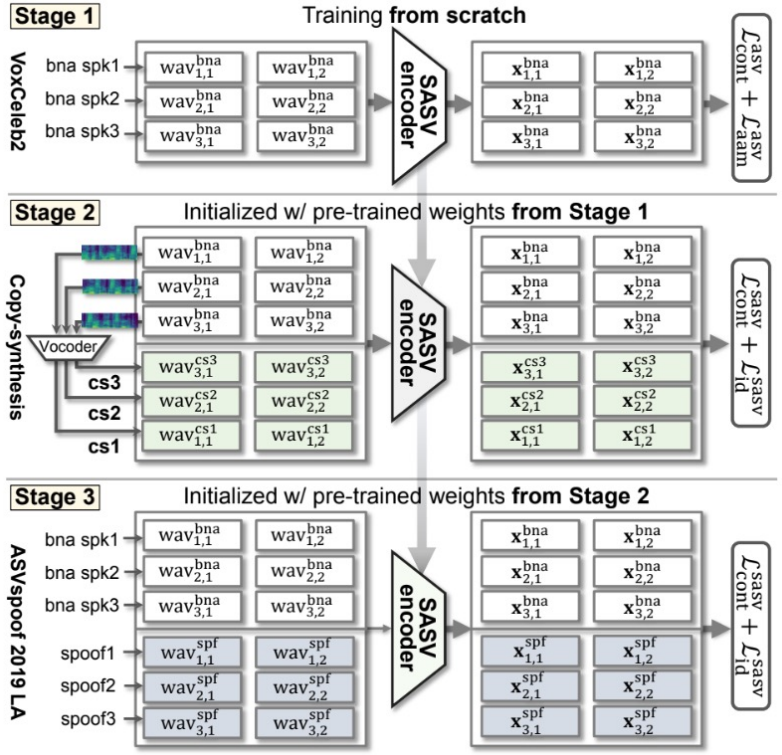
- Embedding-level Fusion System
  - Concatenate two embeddings
  - May pre-train separately



Advanced Topics: SASV

- Integrated System
  - Use one set of embedding that captures the mutual characteristics of ASV and CM [Mun+, 2023]
  - Multi-class problem [Teng+, 2022]





Advanced Topics: SASV

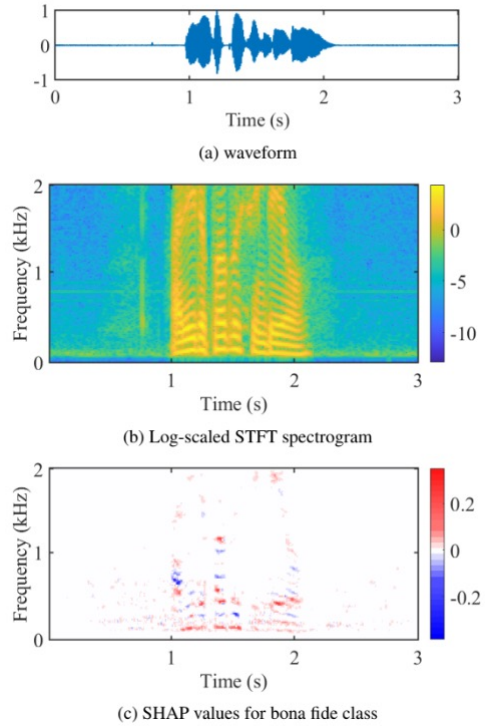
- Current stage
  - Highly rely on the capability of independent ASV and CM subsystems
  - Simple ensemble mechanisms outperform the integrated SASV systems.

Publication		Category	Algorithms for ASV	Algorithms for CM	SV-EER ↓	SPF-EER ↓	SASV-EER ↓
[225]	INTERSPEECH'22	Cascaded	SE-ResNet-34, ECAPA-TDNN	AASIST	**0.90**	**0.26**	**0.29**
[5]	INTERSPEECH'22	Score Fusion	ResNet-48	ResNet-48	**0.19**	**0.25**	**0.22**
[38]	INTERSPEECH'22	Embedding Fusion	Res2Net	AASIST	**0.28**	**0.28**	**0.28**
[211]	INTERSPEECH'22	Integrated System	ECAPA-TDNN, AResNet		8.06	**0.50**	4.86
[167]	INTERSPEECH'23	Integrated System	MFA-Conformer		**1.83**	0.58	**1.19**

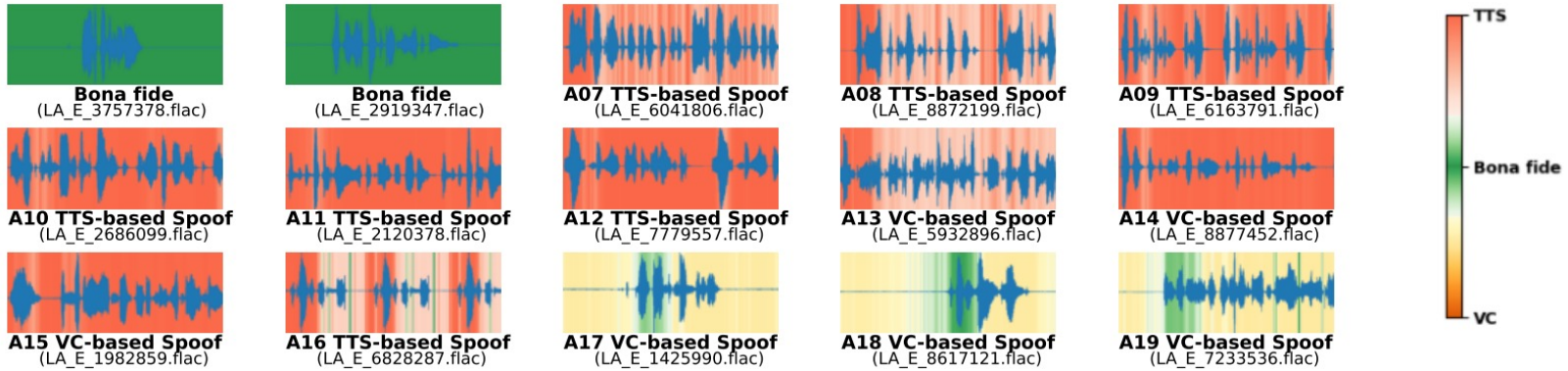
- Future direction -> Ongoing ASVSpoof5 Challenge
  - Joint optimization of ASV and CM components
  - Leverage the strengths of one subsystem to compensate for the weaknesses of the other

Advanced Topics: Interpretability

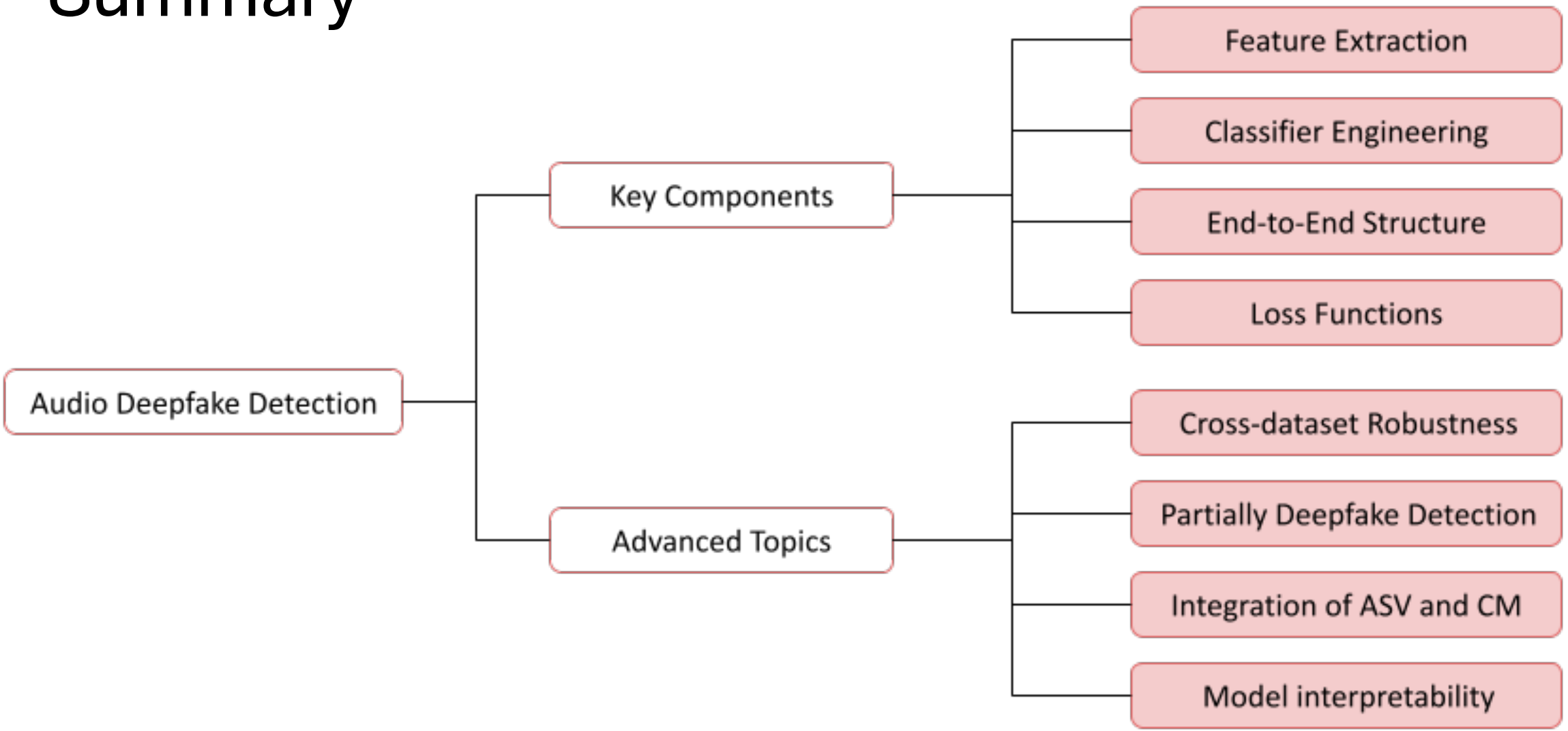
- Explainable-AI (XAI) tools
  - SHAP [Lundberg and Lee, 2017]
  - Grad-CAM [Tak+, 2020]



- Attention mechanisms
  - Temporal attention: emphasize the critical frames [Li and Zhang, 2024]



Summary


Audio Deepfake Detection
Key Components
Feature Extraction
Classifier Engineering
End-to-End Structure
Loss Functions
Advanced Topics
Cross-dataset Robustness
Partially Deepfake Detection
Integration of ASV and CM
Model interpretability

Future Directions

- Robustness
  - Additive noise, channel variation, multi-lingual, or other diverse conditions
  - Data augmentation: feature/dataset-dependent

- Diversity of training datasets
  - Substantial gap between experimental datasets and the realistic conditions
  - Multiple language, partially Deepfake, noise distortion…

- Efficiency
  - Lightweight model for real-time detection.

Survey Paper: *Audio Anti-Spoofing Detection: A Survey*
https://arxiv.org/pdf/2404.13914

Reference

[Abdzadeh and Veisi, 2023] P Abdzadeh and Hadi Veisi. 2023. A Comparison of CQT Spectrogram with STFT-based Acoustic Features in Deep Learning-based Synthetic Speech Detection. Journal of AI and Data Mining 11, 1 (2023), 119–129.

[Alegre+, 2013] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. 2013. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, 1–8

[Chen+, 2020] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury. 2020. Generalization of Audio Deepfake Detection.. In Odyssey. 132–137.

[Cheuk+, 2020] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans. 2020. nnAudio: An on-the-Fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks. IEEE Access 8 (2020), 161981–162003.

[Choi+, 2022] Sunmook Choi, Il-Youp Kwak, and Seungsang Oh. 2022. Overlapped Frequency-distributed network: Frequency-Aware Voice spoofing countermeasure. Interspeech 2022 (Sep 2022).

[Conti+, 2022] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, and Stefano Tubaro. 2022. Deepfake speech detection through emotion recognition: a semantic approach. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 8962–8966.

[Fu+, 2022] Quchen Fu, Zhongwei Teng, Jules White, Maria E Powell, and Douglas C Schmidt. 2022. Fastaudio: A learnable audio front-end for spoof speech detection. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 3693–3697

[Jung+, 2022] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 6367–6371.

[Khan+, 2024] Awais Khan, Khalid Mahmood Malik, and Shah Nawaz. 2024. Frame-to-Utterance

Convergence: A Spectra-Temporal Approach for Unified Spoofing Detection. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 10761–10765.

[Kwak+, 2020] Il-Youp Kwak, Sungsu Kwag, Junhee Lee, Jun Ho Huh, Choong-Hoon Lee, Youngbae Jeon, Jeonghwan Hwang, and Ji Won Yoon. 2021. ResMax: Detecting voice spoofing attacks with residual network and max feature map. In 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 4837–4844

[Lavrentyeva+, 2019] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. 2019. STC antispoofing systems for the ASVSPOOF2019 challenge. Interspeech 2019 (Sep 2019).

[Lai+, 2019] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak. 2019. Assert: Anti-spoofing with squeeze-excitation and residual networks. Interspeech 2019 (Sep 2019).

[Li+, 2021] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. 2021. Replay and synthetic speech detection with res2net architecture. In ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 6354–6358.

[Li+, 2023] Lanting Li, Tianliang Lu, Xingbang Ma, Mengjiao Yuan, and Da Wan. 2023. Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT. Applied Sciences 13, 14 (2023), 8488.

[Li and Zhang, 2024] Menglu Li, Xiao-Ping Zhang, "*Interpretable Temporal Class Activation Representation for Audio Spoofing Detection*", in Proc. Interspeech, 2024.

[Liu+, 2023] Jie Liu, Zhiba Su, Hui Huang, Caiyan Wan, Quanxiu Wang, Jiangli Hong, Benlai Tang, and Fengjie Zhu. 2023. TranssionADD: A multi-frame reinforcement based sequence tagging model for audio deepfake detection. Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023) (2023).

[Lu+, 2024] Jingze Lu, Yuxiang Zhang, Wenchao Wang, Zengqiang Shang, and Pengyuan Zhang. 2024. One-Class Knowledge Distillation for Spoofing Speech Detection. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 11251–11255.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30 (2017).

Reference

[Ma+, 2022] Qiaowei Ma, Jinghui Zhong, Yitao Yang, Weiheng Liu, Ying Gao, and Wing WY Ng. 2022. ConvNeXt Based Neural Network for Audio Anti-Spoofing. arXiv preprint arXiv:2209.06434 (2022).

[Mun+, 2023] Sung Hwan Mun, Hye-jin Shim, Hemlata Tak, Xin Wang, Xuechen Liu, Md Sahidullah, Myeonghun Jeong, Min Hyun Han, Massimiliano Todisco, Kong Aik Lee, and et al. 2023. Towards single integrated spoofing-aware speaker Verification Embeddings. INTERSPEECH 2023 (Aug 2023).

[Pillai+, 2022] Arun Sankar Muttathu Sivasankara Pillai, Phillip L. De Leon, and Utz Roedig. 2022. Detection of voice conversion spoofing attacks using voiced speech. Secure IT Systems (2022), 159–175.

[Ray+, 2021] Ruchira Ray, Sanka Karthik, Vinayak Mathur, Prashant Kumar, G Maragatham, Sourabh Tiwari, and Rashmi T Shankarappa. 2021. Feature genuinization based residual squeeze-and-excitation for audio anti-spoofing in sound AI. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 1–5.

[Sahidullah+, 2015] Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. 2015. A comparison of features for synthetic speech detection. In Proc. Interspeech 2015. 2087–2091.

[Shim+, 2022] Hye-jin Shim, Jungwoo Heo, Jae-Han Park, Ga-Hui Lee, and Ha-Jin Yu. 2022. Graph attentive feature aggregation for text-independent speaker verification. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 7972–7976.

[Tak+, 2020] Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas W. D. Evans, and Massimiliano Todisco. 2020. An Explainability Study of the Constant Q Cepstral Coefficient Spoofing Countermeasure for Automatic Speaker Verification. In Odyssey 2020: The Speaker and Language Recognition Workshop, 1-5 November 2020, Tokyo, Japan

[Tak+, 2021] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 6369–6373

[Tak, Jung+, 2021] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. 2021. End-toend spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. arXiv preprint arXiv:2107.12710 (2021)

[Teng+, 2022] Zhongwei Teng, Quchen Fu, Jules White, Maria Powell, and Douglas Schmidt. 2022. Sa-SASV: An end-to-end spoof-aggregated spoofing-aware speaker verification system. Interspeech 2022 (Sep 2022).

[Wang+, 2022] Ziqian Wang, Qing Wang, Jixun Yao, and Lei Xie. 2022. The NPU-ASLP System for Deepfake Algorithm Recognition in ADD 2023 Challenge. Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023) (2022).

[Wang and Yamagishi, 2022] Xin Wang and Junichi Yamagishi. 2022. Investigating self-supervised front ends for speech spoofing countermeasures. The Speaker and Language Recognition Workshop (Odyssey 2022) (Jun 2022).

[Wang+, 2023] Chenglong Wang, Jiangyan Yi, Jianhua Tao, Chu Yuan Zhang, Shuai Zhang, and Xun Chen. 2023. Detection of cross-dataset fake audio based on prosodic and pronunciation features. INTERSPEECH 2023 (Aug 2023)。

[Wang and Yamagishi, 2024] Xin Wang and Junichi Yamagishi. 2024. Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Apr 2024).

[Wu+, 2020] Zhenzong Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2020. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. Interspeech 2020 (Oct 2020)

[Wu+, 2022] Haibin Wu, Heng-Cheng Kuo, Naijun Zheng, Kuo-Hsuan Hung, Hung-Yi Lee, Yu Tsao, Hsin-Min Wang, and Helen Meng. 2022. Partially fake audio detection by self-attention-based fake span discovery. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 9236–9240.

[Yi+, 2022] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. Add 2022: the first audio deep synthesis detection challenge. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 9216–9220.

Reference

[Yi+, 2023] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, et al. 2023. ADD 2023: the Second Audio Deepfake Detection Challenge. arXiv preprint arXiv:2305.13774 (2023).

[Zeinali+, 2019] Hossein Zeinali, Themos Stafylakis, Georgia Athanasopoulou, Johan Rohdin, Ioannis Gkinis, Lukáš Burget, and Jan Černocký. 2019. Detecting spoofing attacks using VGG and SincNet: But-omilia submission to ASVSPOOF 2019 challenge. Interspeech 2019 (Sep 2019).

[Zhang+, 2021] You Zhang, Fei Jiang, and Zhiyao Duan. 2021. One-class learning towards synthetic voice spoofing detection. IEEE Signal Processing Letters 28 (2021), 937–941.

[Zhang+, 2022] You Zhang, Ge Zhu, and Zhiyao Duan. 2022. A probabilistic fusion framework for spoofing aware speaker verification. The Speaker and Language Recognition Workshop (Odyssey 2022) (Jun 2022).

[Zhang, Wang+, 2022] Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. 2022. The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022), 813–825

[Zhang and Sim, 2022] Bowen Zhang and Terence Sim. 2022. Localizing fake segments in speech. In 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 3224–3230.

[Zhang+ 2023] Yuxiang Zhang, Zhuo Li, Jingze Lu, Hua Hua, Wenchao Wang, and Pengyuan Zhang. 2023. The Impact of Silence on Speech Anti-Spoofing. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).

[Zhu+, 2023] Yi Zhu, Saurabh Powar, and Tiago H Falk. 2023. Characterizing the temporal dynamics of universal speech representations for generalizable deepfake detection. arXiv preprint arXiv:2309.08099 (2023).

Break time: 25 min

Have a coffee break and come back before 3:30 pm.

Video Deepfake Detection

Luchuan Song

Preliminaries on Forgery Detection

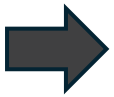
- The previous methods and datasets

Forgery-Boundary Detection (Face-Xray)

Frequency-Domain Binary Classification (F^3-Net)

Traditional binary classification from pixel analysis (LAA-Net).

Learning from the different face attributes, such as identity and Action-Unit (ID-Reveal).



FaceForensics++

Celeb-DF v1/v2

DFDC

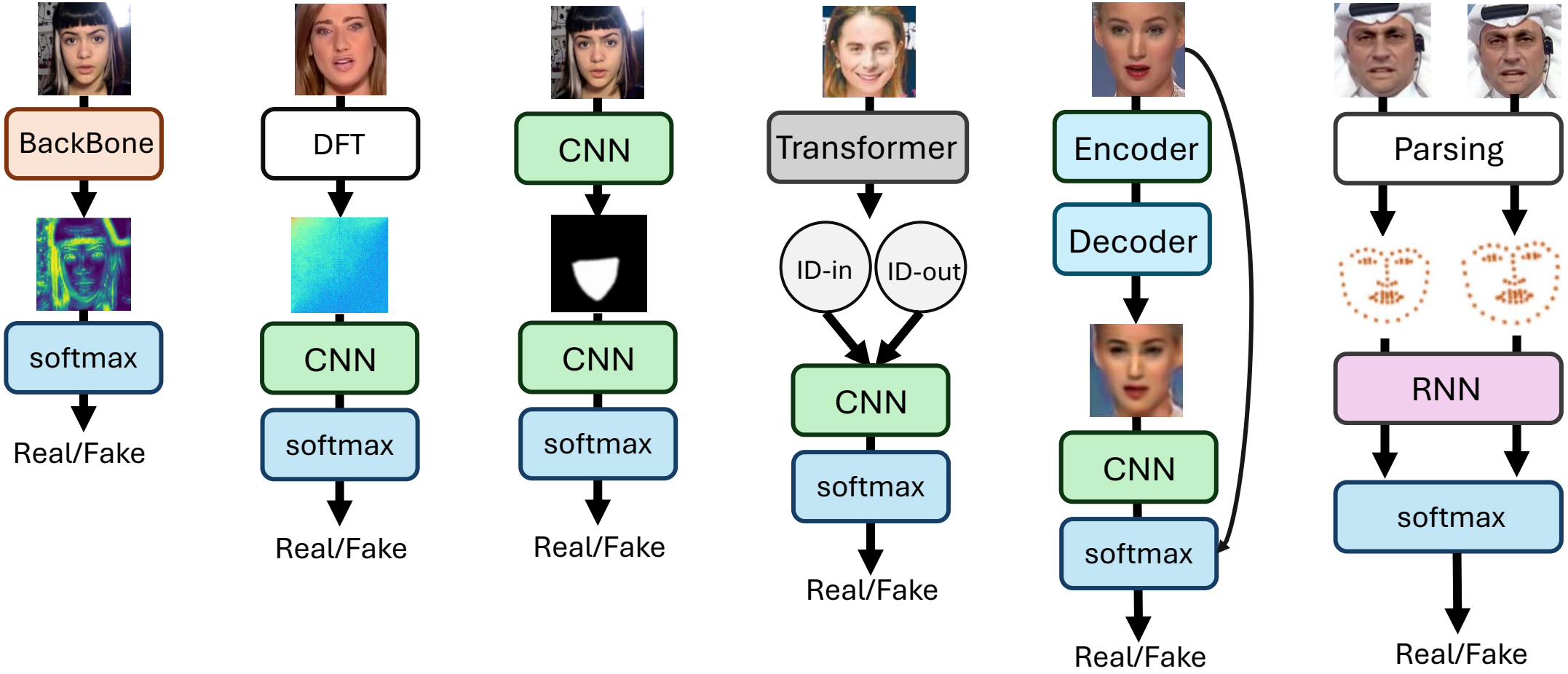
ForgeryNet

DeeperForensics-1.0

How far we away from general forgery detection?

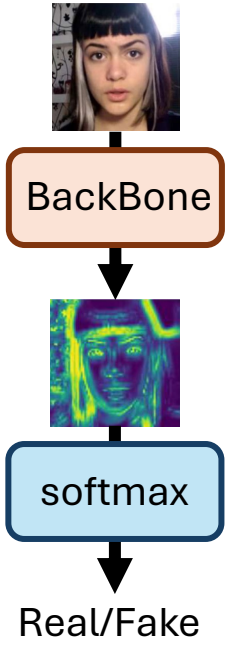
Preliminaries on Forgery Detection

- The overview of previous forgery detection methods

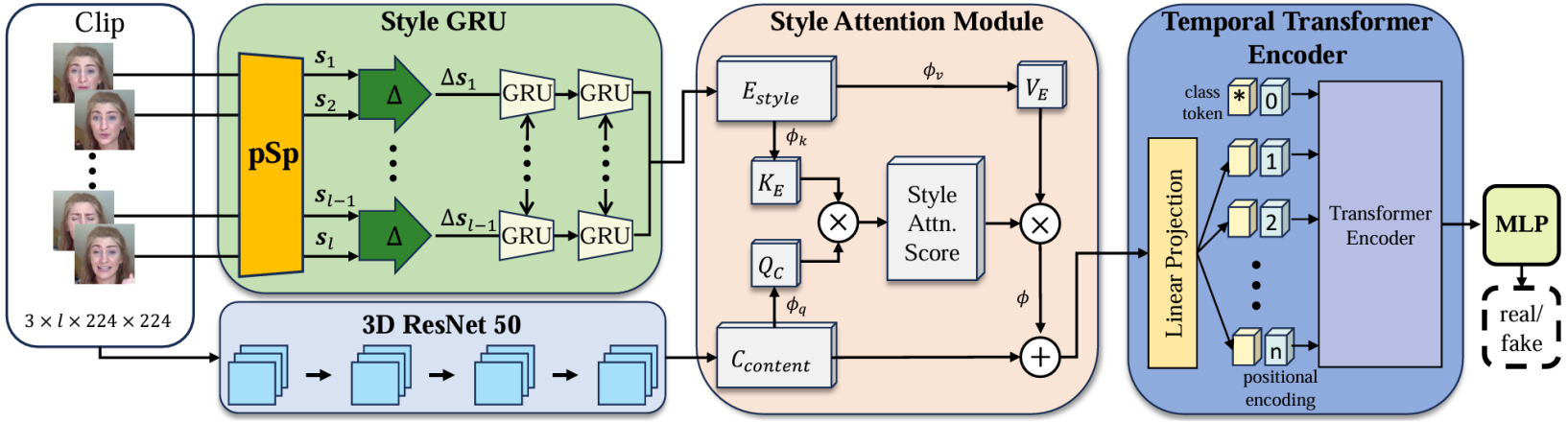


Preliminaries on Forgery Detection

- Learning from the image features



Exploiting Style Latent Flows for Generalizing Deepfake Video Detection [CVPR 2024]

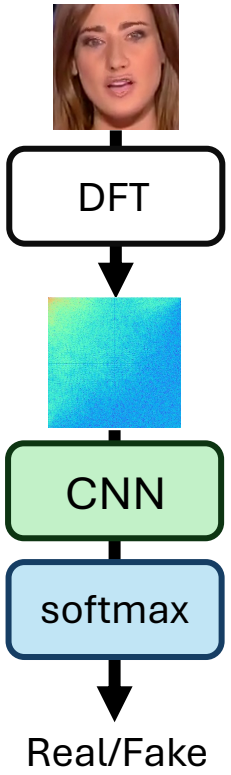


From StyleGAN space to learn the true or fake from video- / pixel- level.

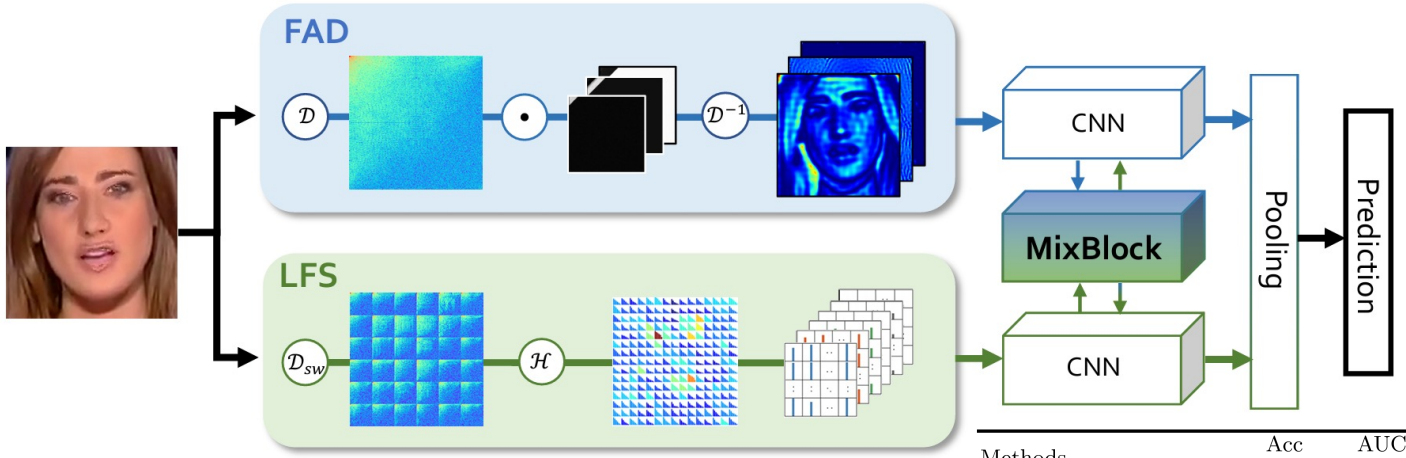
Method	Train on remaining three				
	DF	FS	F2F	NT	Avg
Xception [8]	93.9	51.2	86.8	79.7	77.9
CNN-aug [53]	87.5	56.3	80.1	67.8	72.9
PatchForensics [4]	94.0	60.5	87.3	84.8	81.6
Face X-ray [30]	99.5	93.2	94.5	92.5	94.9
CNN-GRU [44]	97.6	47.6	85.8	86.6	79.4
LipForensics [18]	99.7	90.1	98.8	98.3	97.1
FTCN* [60]	99.8	99.3	95.9	95.3	97.5
AltFreeze [55]	99.8	99.7	98.6	96.2	98.6
StyleGRU	94.0	68.5	88.8	80.6	83.0
Ours	99.7	98.8	98.6	96.4	98.4

Preliminaries on Forgery Detection

- Learning from the frequency features



Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues [ECCV 2020]

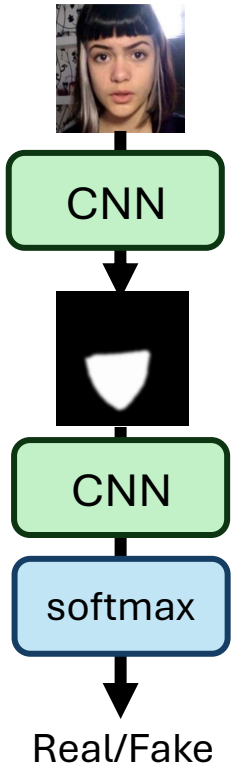


Apply the DFT on the image and obtain the frequence map, then, use the CNN for feature.

Methods	Acc (LQ)	AUC (LQ)	Acc (HQ)	AUC (HQ)	Acc (RAW)	AUC (RAW)
Steg.Features [24]	55.98%	-	70.97%	-	97.63%	-
LD-CNN [14]	58.69%	-	78.45%	-	98.57%	-
Constrained Conv [6]	66.84%	-	82.97%	-	98.74%	-
CustomPooling CNN [49]	61.18%	-	79.08%	-	97.03%	-
MesoNet [3]	70.47%	-	83.10%	-	95.23%	-
Face X-ray [40]	-	0.616	-	0.874	-	-
Xception [12]	86.86%	0.893	95.73%	0.963	99.26%	0.992
Xception-ELA [27]	79.63%	0.829	93.86%	0.948	98.57%	0.984
Xception-PAFilters [10]	87.16%	0.902	-	-	-	-
F^3-Net (Xception)	**90.43%**	**0.933**	**97.52%**	**0.981**	**99.95%**	**0.998**
Optical Flow [5]	81.60%	-	-	-	-	-
Slowfast [20]	90.53%	0.936	97.09%	0.982	99.53%	0.994
F^3-Net(Slowfast)	**93.02%**	**0.958**	**98.95%**	**0.993**	**99.99%**	**0.999**

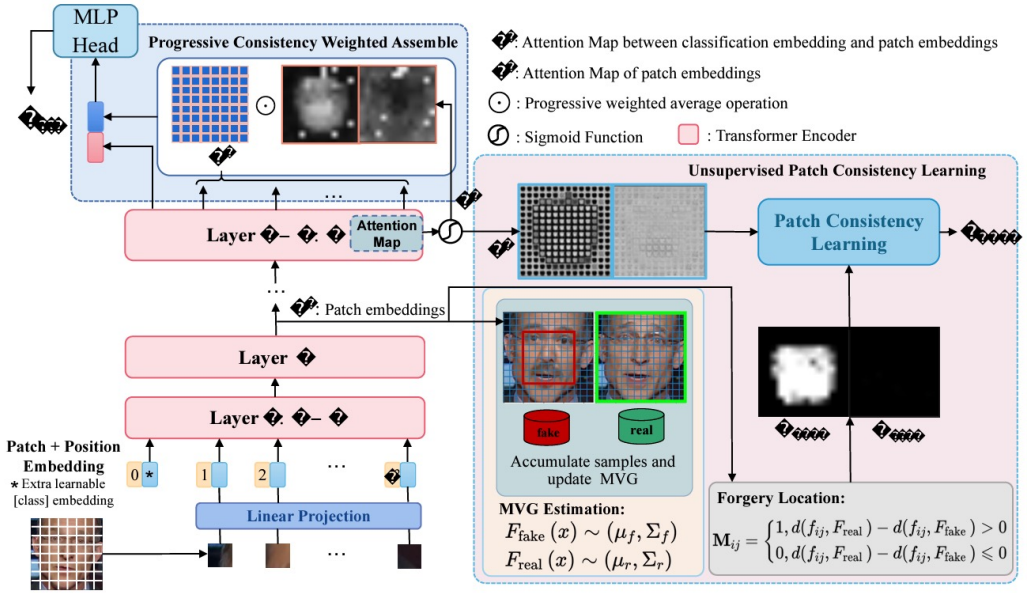
Preliminaries on Forgery Detection

- Learning from the face boundary



UIA-ViT: Unsupervised Inconsistency-Aware Method based on Vision Transformer for Face Forgery Detection[ECCV 2022]

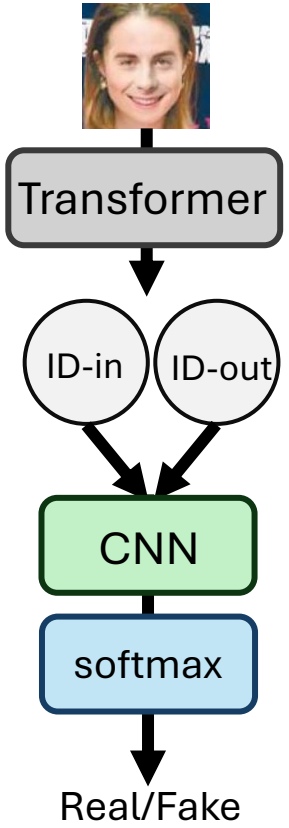
Methods	FF++.HQ	DFD	Celeb-DF-v2	Celeb-DF-v1	DFDC-P
Xception[24]	96.30	70.47	65.50	62.33	72.20
Capsule[20]	96.46	62.75	57.50	60.49	65.95
Multi-Attention[33]	99.29	75.53	67.44	54.01	66.28
FRLM[19]	99.50	68.17	70.58	76.52	69.81
Face X-ray[12]	87.40	85.60	74.20	80.58	70.00
LTW[26]	99.17	88.56	77.14	——	74.58
PCL+I2G[34]	99.11	——	81.80	——	——
Local-relation[1]	99.46	89.24	78.26	——	76.53
DCL[27]	99.30	91.66	82.30	——	**76.71**
UIA-ViT	99.33	**94.68**	**82.41**	**86.59**	75.80



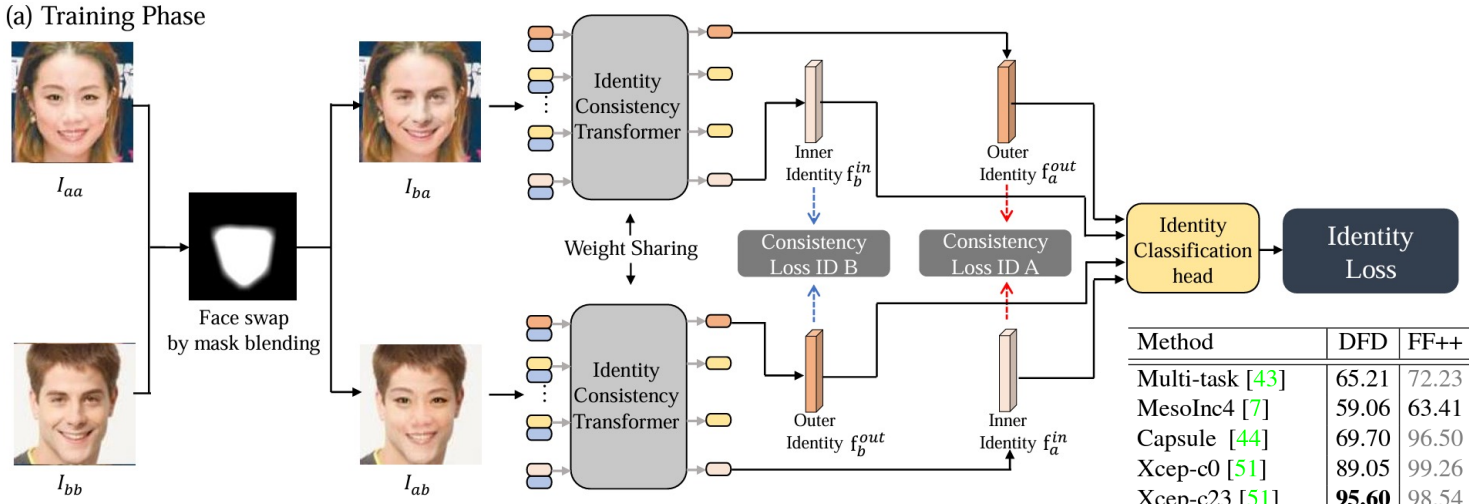
Some computer graphics based methods take the 3D template model to generate the mask.

Preliminaries on Forgery Detection

- Learning from the facial identity score



Protecting Celebrities from DeepFake with Identity Consistency Transformer [CVPR 2022]

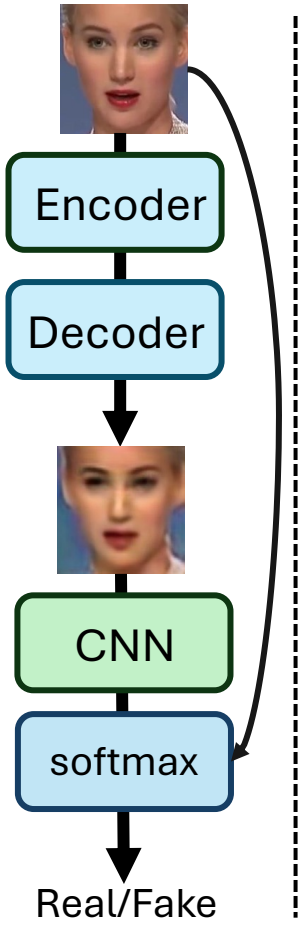


Learning the identity from the datasets, and calculate the identity score for the real or fake.

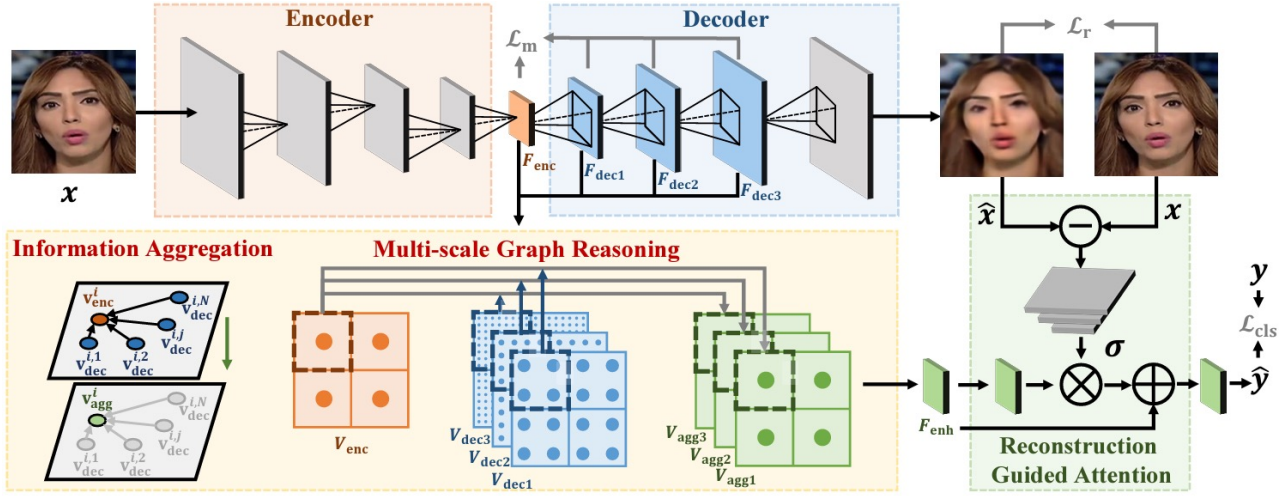
Method	DFD	FF++	Deeper	CD1	CD2	Avg
Multi-task [43]	65.21	72.23	65.32	72.28	61.06	65.96
MesoInc4 [7]	59.06	63.41	51.41	42.26	53.60	53.95
Capsule [44]	69.70	96.50	68.44	69.98	63.65	67.94
Xcep-c0 [51]	89.05	99.26	57.76	48.08	50.37	61.32
Xcep-c23 [51]	**95.60**	98.54	69.85	74.97	77.82	79.56
FWA [34]	80.59	74.82	45.46	72.88	64.87	67.72
DSP-FWA [34]	90.99	81.90	60.00	78.51	81.41	78.56
CNNDetect [60]	60.12	71.08	57.16	56.12	57.17	60.33
Patch-Foren [13]	49.91	73.75	55.35	59.66	57.16	55.52
FFD [20]	76.61	**92.32**	46.64	74.15	77.80	73.50
Face X-ray [32]	94.14	98.44	72.35	74.76	75.39	79.16
Two-Branch [38]*	-	-	-	-	73.41	-
PCL+I2G [63]*	-	-	-	-	81.80	-
Nirkin.*et al.* [47]*	-	99.70	-	-	66.00	-
ICT (Ours)	84.13	90.22	**93.57**	**81.43**	**85.71**	**87.01**
ICT-Ref (Ours)	93.17	98.56	99.25	96.41	94.43	96.34

Preliminaries on Forgery Detection

- Learning from self-reconstruction



End-to-End Reconstruction-Classification Learning for Face Forgery Detection [CVPR 2022]

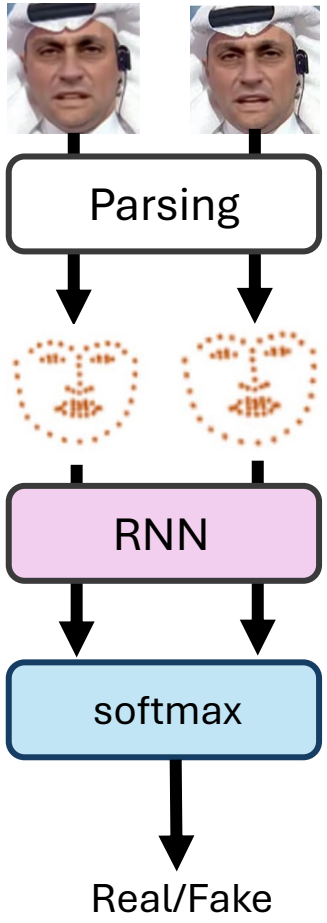


It uses the VAE to reconstruction the face image within the forgery detection training.

Methods	FF++ (c23)		FF++ (c40)		Celeb-DF		WildDeepfake	
	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc(%)	AUC (%)
MesoNet [1]	83.10	–	70.47	–	–	–	64.47	–
Multi-task [30]	85.65	85.43	81.30	75.59	–	–	–	–
Xception [35]	95.73	96.30	86.86	89.30	97.90	99.73	77.25	86.76
Face X-ray [23]	–	87.40	–	61.60	–	–	–	–
Two-branch [29]	96.43	98.70	86.34	86.59	–	–	–	–
SPSL [25]	91.50	95.32	81.57	82.82	–	–	–	–
RFM [45]	95.69	98.79	87.06	89.83	97.96	**99.94**	77.38	83.92
Freq-SCL [22]	96.69	99.28	89.00	92.39	–	–	–	–
Add-Net [60]	96.78	97.74	87.50	91.01	96.93	99.55	76.25	86.17
F^3-Net [33]	97.52	98.10	90.43	93.30	95.95	98.93	80.66	87.53
MultiAtt [55]	**97.60**	99.29	88.69	90.40	97.92	**99.94**	82.86	90.71
RECCE (Ours)	97.06	**99.32**	**91.03**	**95.02**	**98.59**	**99.94**	**83.25**	**92.02**

Preliminaries on Forgery Detection

- Learning from face motion

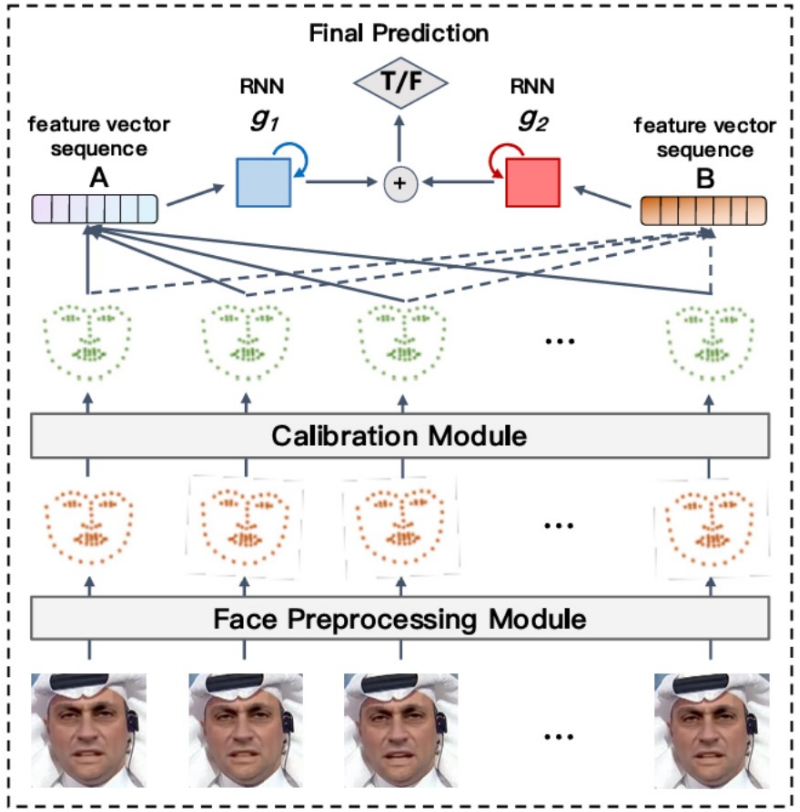


Improving the Efficiency and Robustness of Deepfakes Detection Through Precise Geometric Features [CVPR 2021]

Methods	FF++			Decline
	raw	c23	c40	
Xception [25]	99.7	93.3	86.5	6.4/13.2
X-Ray [16]	99.1	87.3	61.6	11.8/37.5
LRNet (ours)	**99.9**	**97.3**	**95.7**	**2.6/4.2**

Methods	Celeb-DF			Decline
	raw	c23	c40	
Xception-c23 [25]	**65.3**	**65.5**	52.5	**-0.2**/12.8
FWA [18]	56.9	54.6	52.2	2.3/4.7
DSP-FWA [18]	64.6	57.7	47.2	6.9/17.4
LRNet (ours)	57.4	56.3	**55.4**	1.1/**2.0**

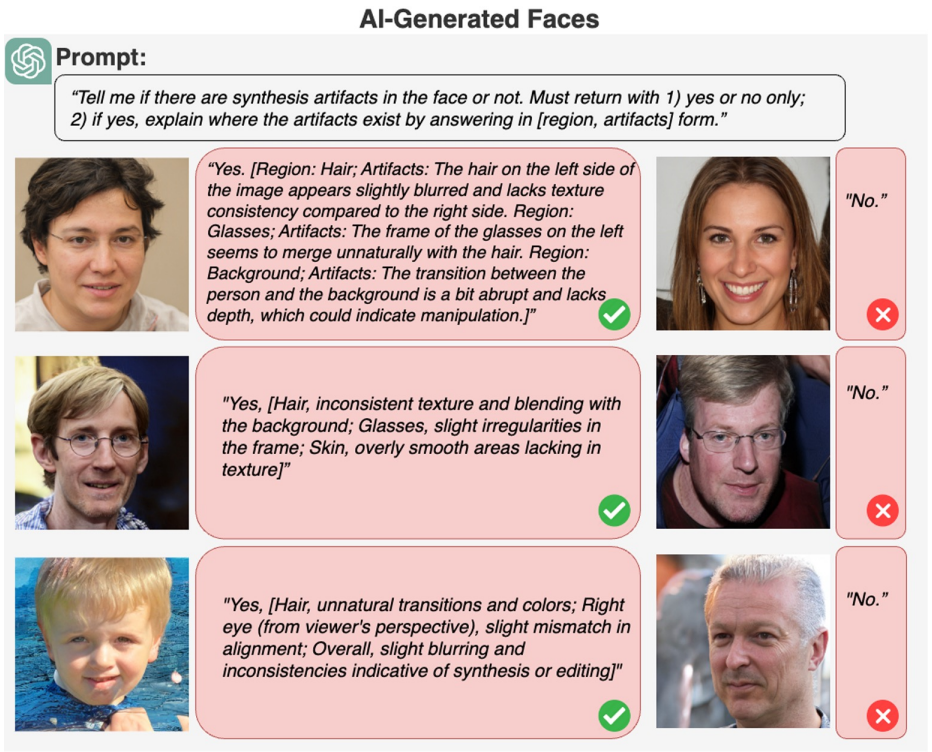
Learning the temporal consistency from the detected facial landmarks in the image sequence.

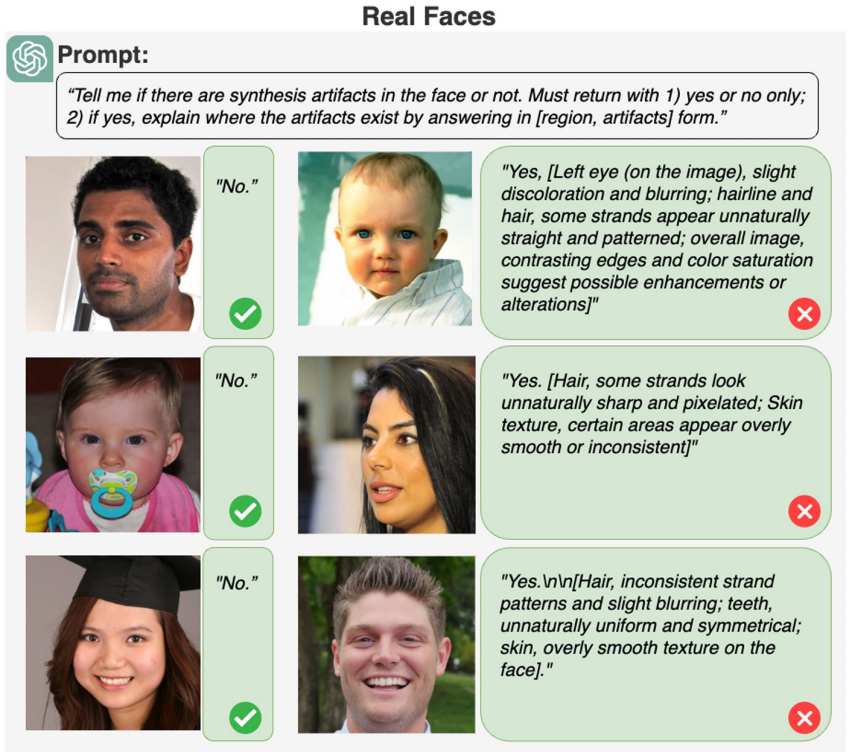


Preliminaries on Forgery Detection

- Learn the forgery images from ChatGPT! (Another insight)

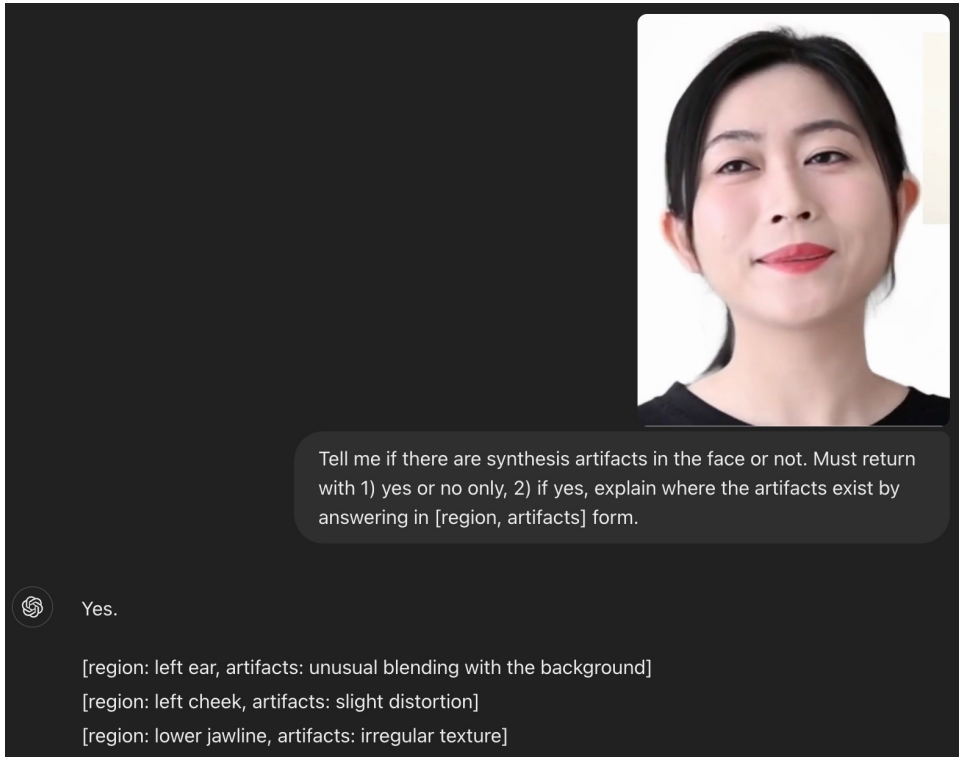
  Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics [CVPRW 2024]

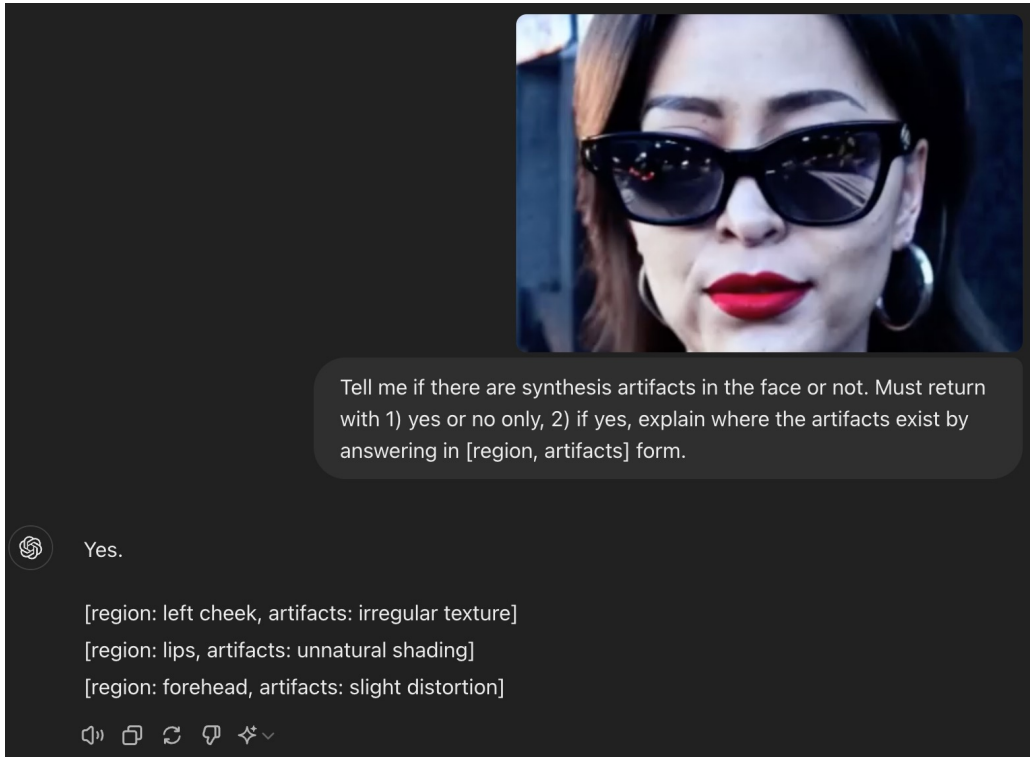




Preliminaries on Forgery Detection

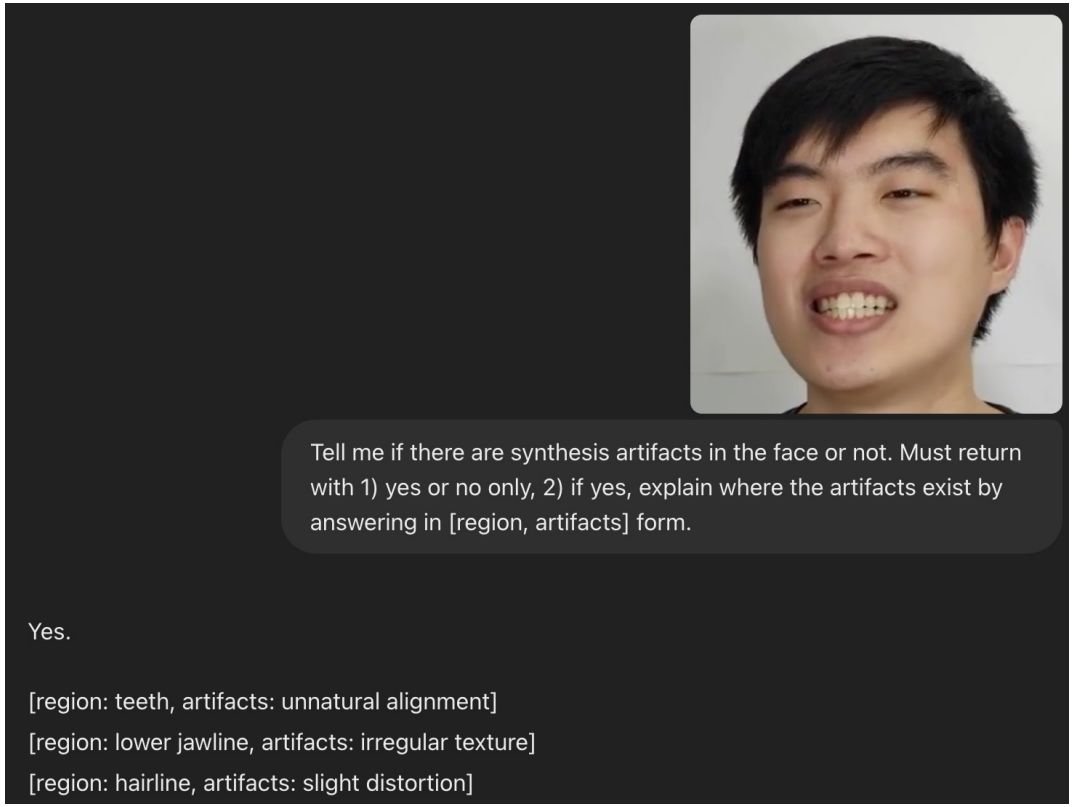
- Learn the forgery images from ChatGPT! (Another insight)

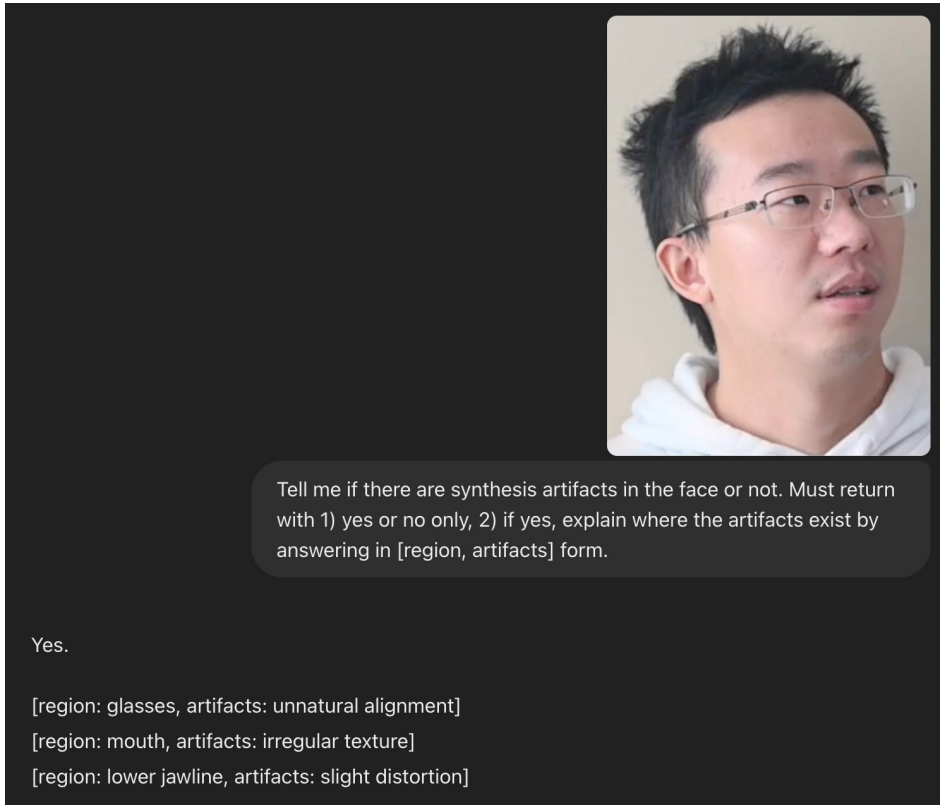




Preliminaries on Forgery Detection

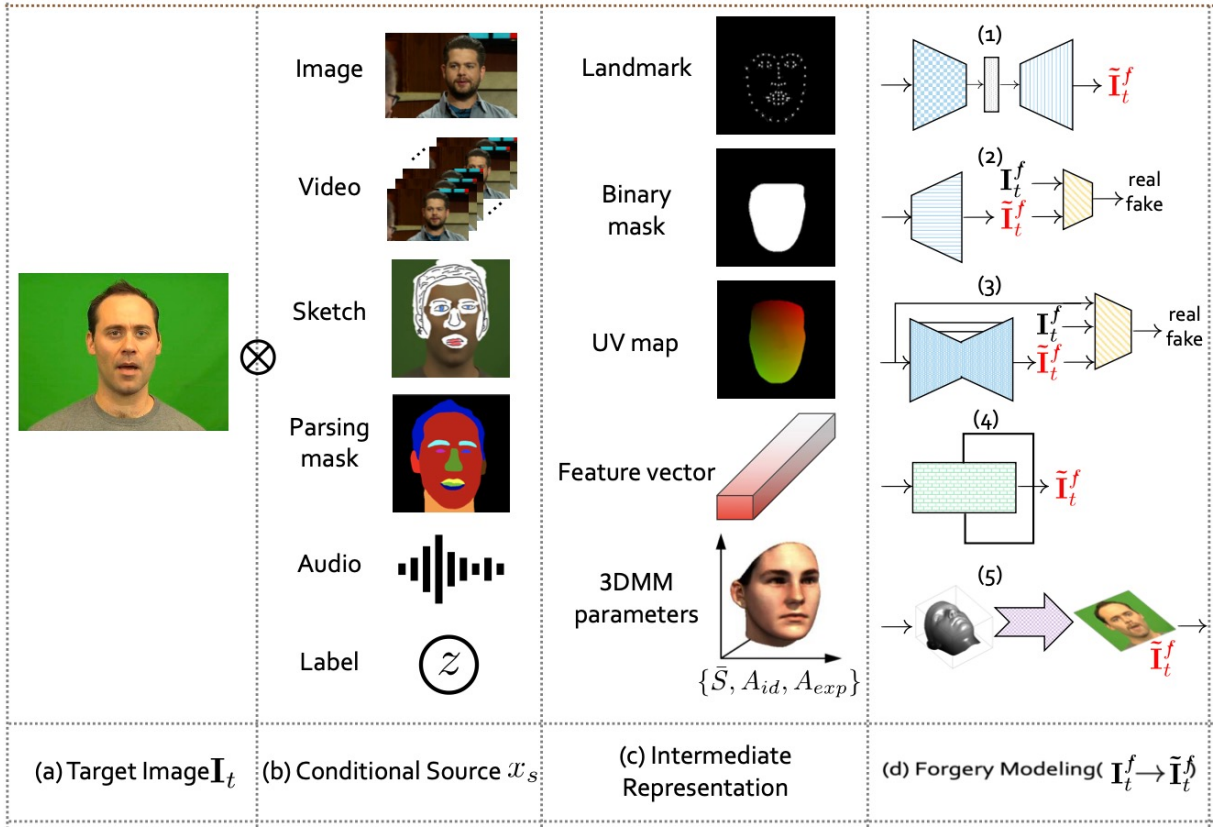
- Learn the forgery images from ChatGPT! (Another insight)





Preliminaries on Forgery Detection

Preliminaries
Methods on
Face Forgery



Preliminaries on Forgery Detection

- The overview of our previous datasets – ForgeryNet

Table 1: Comparison of various face forgery datasets. ForgeryNet surpasses any other dataset both in scale and diversity. It provides both video- and image-level data. The forgery data are constructed by 15 manipulation approaches within 4 categories. We also employ 36 types of perturbations from 4 kinds of distortions for post-processing.

Dataset	Video Clips		Still images		Approaches	Subjects	Uniq. Perturb.	Mix Perturb.	Annotations
	Real	Fake	Real	Fake					
UADFV [60]	49	49	241	252	1	49	-	×	591
DF-TIMIT [36]	320	640	-	-	2	43	-	×	1,600
Deep Fake Detection [4]	363	3,068	-	-	5	28	-	×	3,431
Celeb-DF [39]	590	5,639	-	-	1	59	-	×	6,229
SwapMe and FaceSwap [64]	-	-	4,600	2,010	2	-	-	×	6,610
DFFD [14]	1,000	3,000	58,703	240,336	7	-	-	×	8,000
FaceForensics++ [52]	1,000	5,000	-	-	5	-	2	×	11,000
DeeperForensics-1.0 [33]	50,000	10,000	-	-	1	100	7	✓	60,000
DFDC [18]	23,564	104,500	-	-	8	960	19	×	128,064
ForgeryNet (Ours)	**99,630**	**121,617**	**1,438,201**	**1,457,861**	**15**	**5400+**	**36**	✓	**9,393,574**

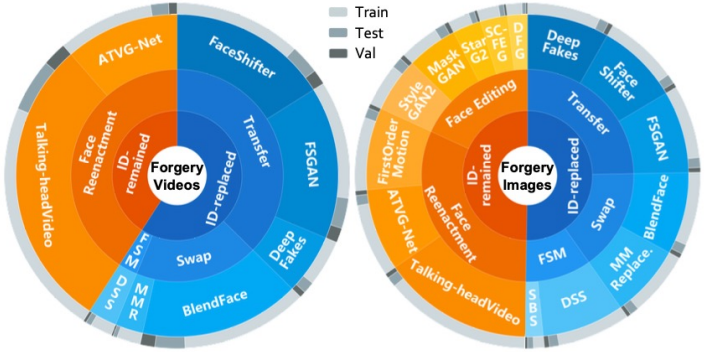


Figure 6: Illustration of image- and video-level sets. From the inside to the outside are categories of *Identity-remained* and *Identity-replaced*, corresponding sub-types, specific forgery approaches and the situation of data split.

Preliminaries on Forgery Detection

- The generation methods for ForgeryNet

Table 10: **Summary of the four types of forgery approaches.** In this table, the input, output, architecture, resolution, modification ability, and whether to retrain in inference of each forgery approach are presented. S/T represents the modality of x_s and x_t. v:=video, i:=image, a:=audio, m:= mask, s:=sketch, l:= noise, S:=single identity, M:=multiple identity

	Method	S/T	CG/GAN	Input	Modification	Resolution	Retraining
Face Reenactment	FirstOrderMotion [56]	v/i	GAN	M/M	pose,expression	256*256	No need
	ATVG-Net [9]	v/i	GAN	M/M	pose,expression	128*128	No need
	Talking-head Video [23]	a/v	CG+GAN	M/S	mouth	256*256	1∼3 portraits
Face Editing	StarGAN2 [11]	i/i	GAN	M/M	attribute transfer	256*256	portraits
	StyleGAN2 [35]	l/i	GAN	M/M	rebuild from latent	1024*1024	portraits
	MaskGAN [37]	m,i/i	GAN	M/M	editing record	512*512	portraits,mask
	SC-FEGAN [34]	s,i/i	GAN	M/M	sketch record	512*512	portraits,sketch
	DiscoFaceGAN [17]	i/i	CG+GAN	M/M	3dmm attributes	1024*1024	portraits
Face Transfer	BlendFace	v/v	CG	M/M	identity, expression	Any	No need
	MMReplacement	i/i	CG	M/M	identity, expression	Any	at least 1 protrait
Face Swap	FSGAN [47]	v/v	GAN	M/M	identity	256*256	No need
	DeepFakes [49]	v/v	GAN	S/S	identity	192*192	2k∼5k portraits
	FaceShifter [38]	i/i	GAN	M/M	identity	256*256	No need

Preliminaries on Forgery Detection

- How will the backbone performance on the ForgeryNet?

Table 2: **Image Forgery Classification (Protocol 1):** binary classification. We report accuracy and AUC scores of the compared forensics methods.

Method	Param.	Acc	AUC
MobileNetV3 Small [29]	1.7M	76.24	85.51
MobileNetV3 Large [29]	4.2M	78.30	87.56
EfficientNet-B0 [58]	4.0M	79.86	89.31
ResNet-18 [28]	11.2M	78.31	87.75
Xception [12]	20.8M	80.78	90.12
ResNeSt-101 [62]	46.2M	82.06	91.02
SAN19-patchwise [63]	18.5M	80.08	89.38
ELA-Xception [27]	20.8M	73.77	82.69
SNRFilters-Xception [10]	20.8M	81.09	90.52
GramNet [44]	22.1M	80.89	90.20
F^3-Net [50]	57.3M	80.86	90.15

Table 9: **Temporal Forgery Localization.** We show AP, AR and mAP scores of all compared methods.

	AR		AP			avg.
	2	5	0.5	0.75	0.9	AP
Xception [12]	25.83	73.95	68.29	62.84	58.30	62.83
X3D-M+BSN [42]	81.33	86.88	80.46	77.24	55.09	70.29
X3D-M+BMN [41]	88.44	91.99	90.65	88.12	74.95	83.47
SlowFast+BSN [42]	83.63	88.78	82.25	80.11	60.66	73.42
SlowFast+BMN [41]	90.64	93.49	92.76	91.00	80.02	86.85

Video Forgery Detection

- Recent years, we have witnessed a branch of generative methods

Diffusion Models

NeRF rendering

3D Gaussian-Splatting rendering

StyleGAN-Series Models

→ More Potential Risk than Face-Swapping

→ Out-of-distribution in previous datasets

Preliminaries on Forgery Detection

- Could these methods achieve general forgery detection?

  1. For the methods do not have the boundary (diffusion-based) ?
  2. The self-reconstruction model have large distribution with wildforgery ?
  3. The extracted features are robust in frequency domain ?
  4. How about the forgery methods are built from the landmarks?

Find it on the state-of-the-art forgery methods

Video Forgery Detection

OOD Forgery Datasets

Datasets	FF++ [2019]	DFDC [2019]	WildDF [2021]	ForgeryNet [2021]	Celeb-DF^2 [2019]	DeeperF-1.0 [2020]
One-Shot Reconstruction						
Face-Vid2Vid	–	–	–	–	–	–
FOMM	–	–	–	✓	–	–
NOFA	–	–	–	–	–	–
Next3D	–	–	–	–	–	–
DiffTalk	–	–	–	–	–	–
AdaSR-TH	–	–	–	–	–	–
PIRender	–	–	–	–	–	–
StyleGAN	–	✓	–	✓	✓	–
Personality Reconstruction						
Tri^2-plane	–	–	–	–	–	–
NeuralTexture	✓	–	–	✓	–	✓
DVP	–	–	–	–	–	–
3D Gaussian	–	–	–	–	–	–
INSTA	–	–	–	–	–	–
PointAvatar	–	–	–	–	–	–
StyleGAN*	–	–	–	–	–	–

Table 1. The evaluation of forgery methodologies in preceding datasets is presented, where a checkmark (✓) denotes the inclusion of generation methods within the dataset. It is noteworthy that almost all cutting-edge generation techniques are absent from these collections.

Kaleidoscope Forgery Methods in AIGC (bbox)

- We list the forgery methods in recent years, called AIGC forgery methods

- **FOMM**: The first-order-motion-model [Siarohin et al. 2019] represents a one-shot 2D motion retargeting approach, which is trained on extensive facial datasets and necessitates no additional computational overhead during the inference phase.
- **LIA**: The Latent Image Animator [Wang et al. 2022] is also the one-shot animation method. Different from the FOMM, the LIA is wrapped by the optical flow instead of adaptive Jacobin matrix.
- **Face-Vi2Vid**: The Face-Vid2Vid [Wang et al. 2021] is the 3D expanded version of FOMM. And it does not include inpainting methods based on neural networks for the generation backbone.
- **AdaSR-TH**: The AdaSR Talking-Head [Song et al. 2024] is the high-resolution extension of Face-Vid2Vid, it uses super-resolution modules to improve video quality during encoding and decoding.

- **StyleHEAT**: The one-shot facial generation method with the backbone of StyleGAN. It is trained on large-scale face datasets, and in the evaluation phase, only one source face needs to be inverted.
- **StyleAvatar**: The StyleAvatar is the StyleGAN-based personality method, it finetunes the StyleGAN on the short identity-specific video.

- **EMO-Portrait**: The EMO-Portrait applies the diffusion model for one-shot image to head generation. It takes audio as input, but can still be modified to adapt pose-driven facial animation
- **VASA-1**: The VASA-1 is a kind of DiT structure, it takes the audio as input, we are try to reproduce it and make it run with pose driven.

- **Deep Video Portrait**: The DVP [Kim et al. 2018] is the UNET-based personality generative model. As a representative work of generative adversarial networks, it has achieved state-of-the-art quality.
- **Next3D***: We finetune the one-shot-based next3d method on the short specific identity videos for better performance.
- **Tri^2-plane**: The Tri^2-plane is a method that adopts multiple triplane structures to fine-tune on a person-specific video set. It achieves good performance on video quality.
- **3D Gaussian**: We adopt the 3D gaussian splatting in the monocular avatar generation. It is the state-of-the-art graphics technology that can achieve the breakthrough between efficiency and synthesis quality.

One-Shot Generation Methods (GAN + Diffusion + StyleGAN)

3D-based Methods

Kaleidoscope Forgery Methods in AIGC

Kaleidoscope Forgery Methods

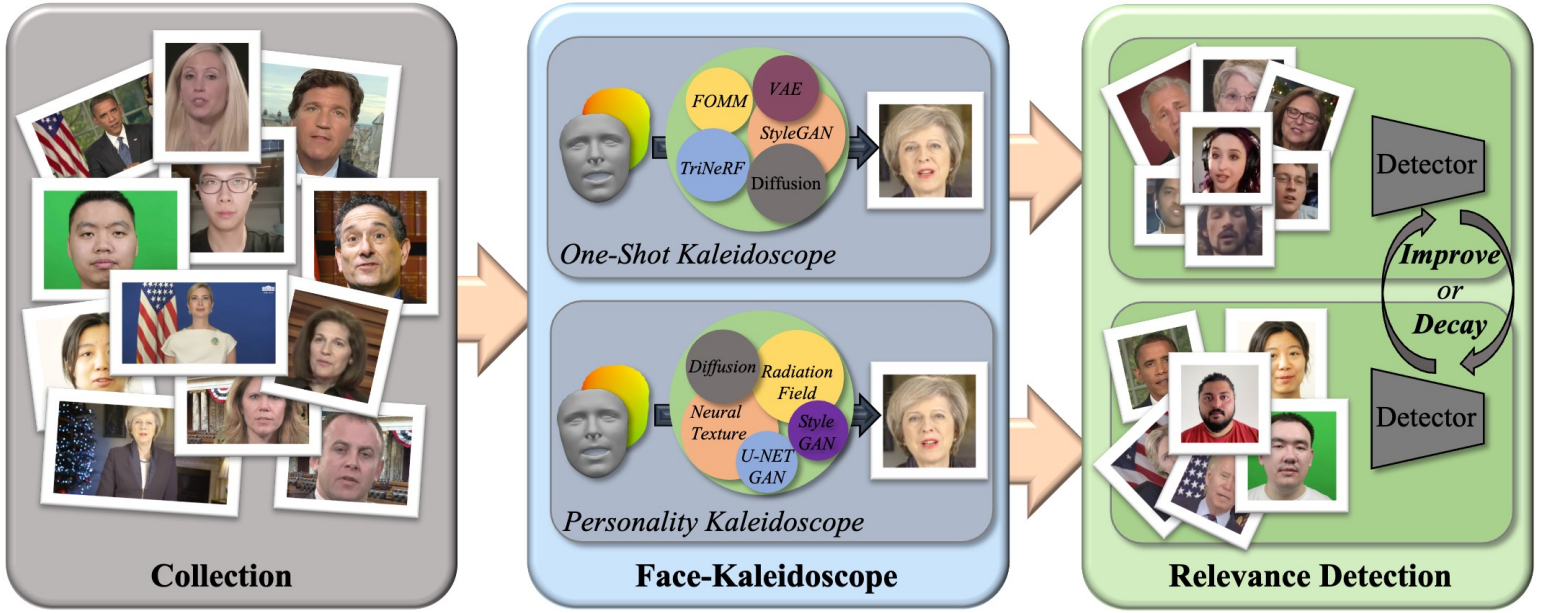
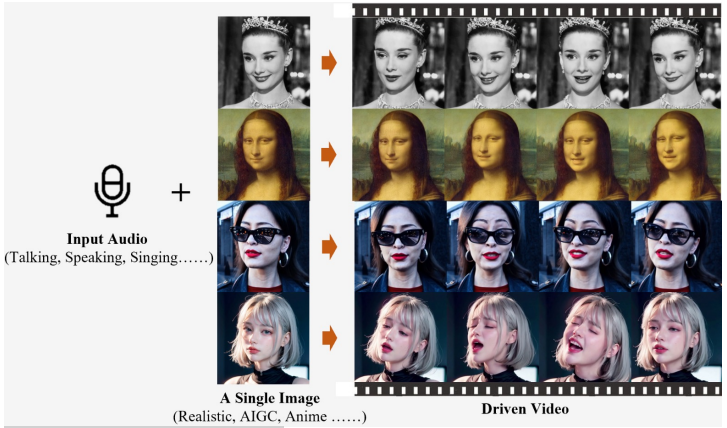


Fig. 1. The Face-Kaleidoscope is a mega-scale datatset for facial forgery detection. In the context of the recent exponential surge in AI-generated content, the imperative for robust mechanisms to identify face forgeries has been underscored by the significant security challenges introduced by advanced face reconstruction methodologies. The previous face forgery datasets suffer from the state-of-the-art face forgery methods, such as UNET-based rendering (*e.g.* Deep Video Portraits [Kim et al. 2018]), StyleGAN-based rendering (*e.g.* StyleAvatar [Wang et al. 2023]) and RadiationField-based rendering (*e.g.* NeRF [Mildenhall et al. 2020], Gaussian Splatting [Kerbl et al. 2023], Point Cloud [Zheng et al. 2023]) *e.t.c.*. Moreover, we provide source code as toolchain to these methods for researchers to personalize self-data. The Face-Kaleidoscope includes two different sets, One-Shot Kaleidoscope and Personality Face-Kaleidoscope. The One-Shot Kaleidoscope is a resource-free facial reconstruction method that can produce data faster but has obvious artifacts. The Personality Face-Kaleidoscope is the identity-specially rendering, which relies more on computing resources but has better quality. We discuss the deepfake detector adaptation to these two methods, which can be inspiring for general deepfake detection.

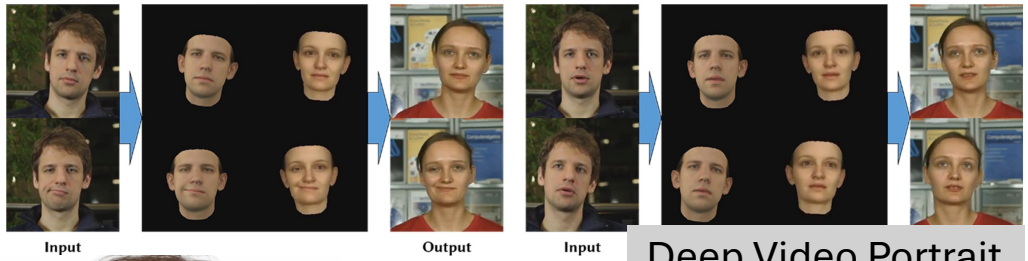
Kaleidoscope Forgery Methods in AIGC

LivePortrait



EMO-Portrait

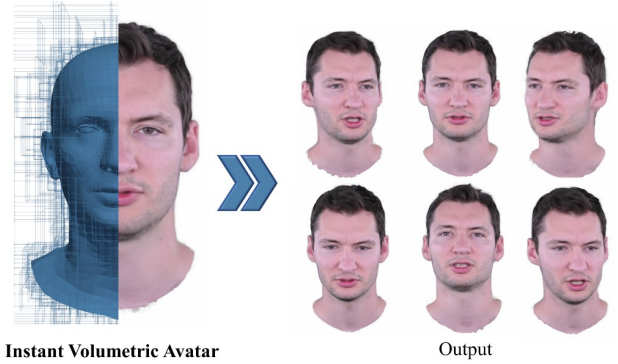
Face-Vid2Vid



Deep Video Portrait

VASA-1

NPGA



INSTA

Overview of some New Forgery Methods

Kaleidoscope Forgery Methods in AIGC

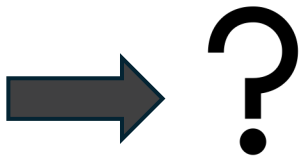
- Here we show some video results on these new forgery methods. Please pay attention to our new work on ECCV 2024.

https://songluchuan.github.io/Tri2Plane.github.io/

Out-of-distribution Analysis of Datasets

- We take each method synthesized about 200 videos, and we evaluated the toy dataset on several popular DeepFake detection methods. We list the video DeepFake detection methods in below:

  - **SlowFast**: The SlowFast [Feichtenhofer et al. 2019] is a classic video understanding backbone. It is also the baseline for video-level deepfake classification.
  - **IIL**: The Implicit Identity Leakage [Dong et al. 2023] is the state-of-the-art deepfake detection method, which claims to have achieved general deepfake detection. We also evaluate the generalizability within our dataset.
  - **ID-DFD**: The Identity Driven Deepfake [Huang et al. 2023] is the state-of-the-art deepfake detection method. It performs binary classification based on the identity embedding.
  - **F^3-Net**: The F^3-Net [Qian et al. 2020] is a frequency-based forgery classification method, which is a kind of classic two-branch deepfake detection method.
  - **UIA-ViT**: The UIA-ViT [Zhuang et al. 2022] is the transformer based method for binary forgery classification. It uses the attention map between classification embedding and patch embedding.



How will the previous methods work on new forgery data?

Out-of-distribution Analysis of Datasets

- We take each method synthesized about 200 videos, and we evaluated the toy dataset on several popular DeepFake detection methods. We list the video DeepFake detection methods in below:

Method	Acc.(%) [Our]	AUC(%) [Our]	Acc.(%) [FF++]	AUC(%) [FF++]
SlowFast [Feichtenhofer et al. 2019]	63.72	67.10	90.53	93.60
IIL [Dong et al. 2023]	51.95	60.07	98.51	99.8
ID-DFD[Huang et al. 2023]	55.08	57.41	97.00	99.46
F^3-Net [Qian et al. 2020]	61.90	64.27	90.43	93.30
UIA-ViT [Zhuang et al. 2022]	69.47	71.22	90.40	99.33

Table 2. The binary classification evaluation results of previous deepfake detection methods on the FF++ dataset and our toy dataset. We report the accuracy and AUC scores of the compared forensics methods. Each method are trained and tested on our toy dataset and FF++ dataset respectively.

Those methods on our toy dataset seem not perform as well as the FF++?

Out-of-distribution Analysis of Datasets

- The previous methods seem to be overfitted on FF++ dataset, but on our dataset, the performance is not very good.
- Some methods specially designed for FF++, such as Face X-ray are not suitable for the GAN/StyleGAN-based methods, they are work on the forgery boundary, but some methods do not have such boundary.
- There are some obvious drawbacks in the previous DeepFake (Face Forgery) datasets, such as the color inconsistency produced by the forced combination of ATVGNet on background in FogeryNet.

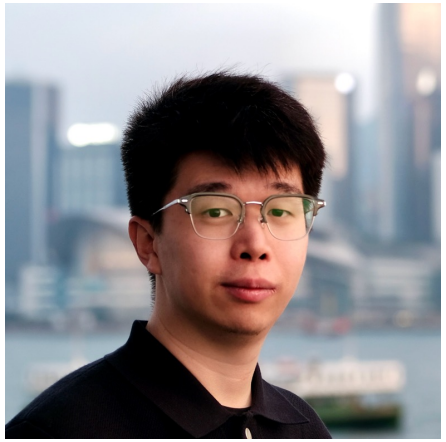
A new benchmark is need for the video DeepFake Detection topic!

Include more new methods, not only play with the four methods on FF++

Some new forgery methods on the state-of-the-art methods should be explored

Audio-Visual Deepfake Detection (AVDD)

You (Neil) Zhang

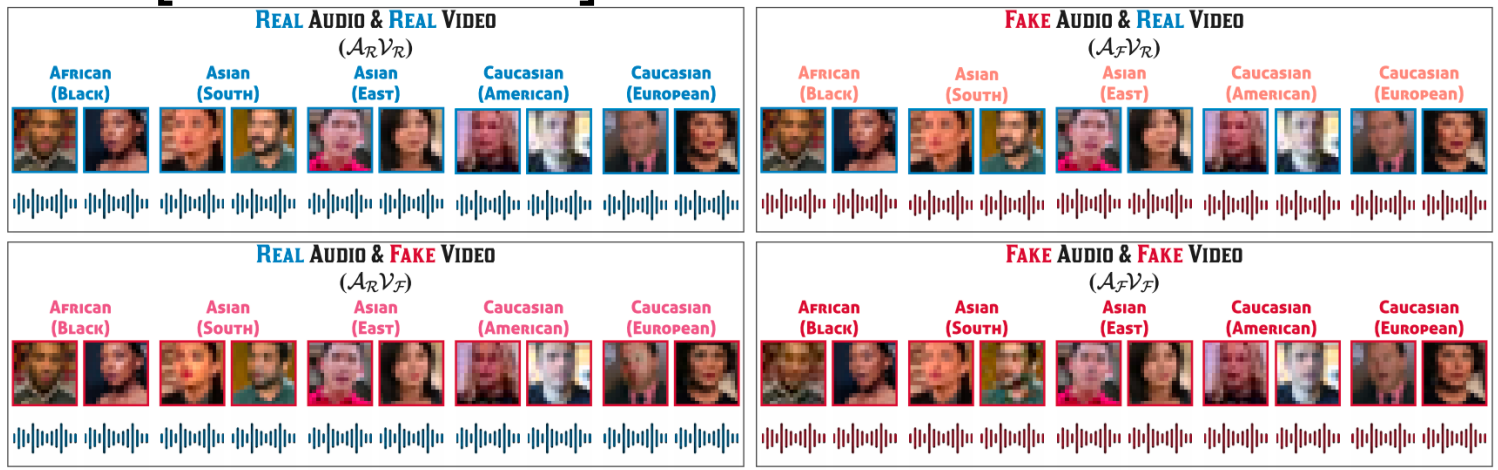


Audio-Visual Deepfake Detection

- Detecting talking face deepfakes
- Temporal forgery localization
- Detecting general video deepfakes
- Emerging & future directions

Audio-Visual Deepfake Datasets: Talking Faces

- DFDC [Dolhansky+2020]: 8 facial modification algorithms + 1 TTS
- FakeAVCeleb [Khalid+2021]

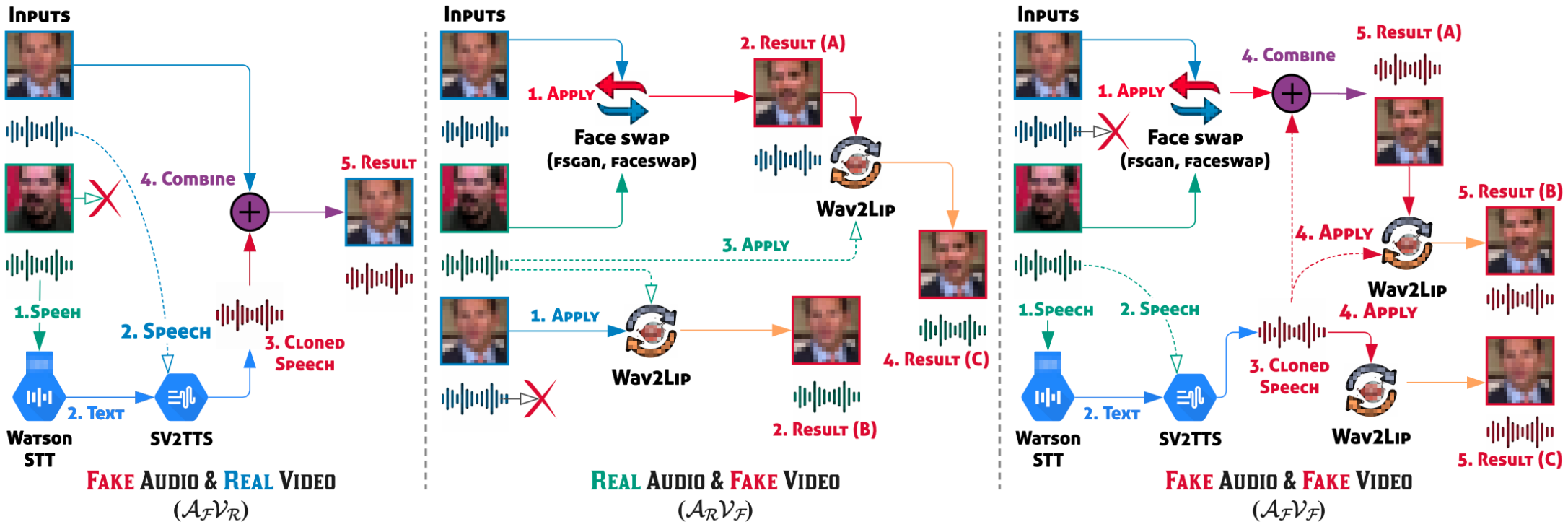


- SWAN-DF [Korshunov+2023]
- PloyGlotFake [Hou+2024]

Dolhansky, Brian, et al. "The deepfake detection challenge (DFDC) dataset." *arXiv* 2020.
Khalid, Hasam, et al. “FakeAVCeleb: A novel audio-video multimodal deepfake dataset.” *NeurIPS Datasets Track* 2021.
Korshunov, Pavel, et al. “Vulnerability of Automatic Identity Recognition to Audio-Visual Deepfakes.” *IJCB* 2023.
Hou, Yang, et al. "PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset." *arXiv* 2024.

FakeAVCeleb

- FakeAVCeleb generation pipeline [Khalid+2021]



Khalid, Hasam, et al. "FakeAVCeleb: A novel audio-video multimodal deepfake dataset." *NeurIPS Datasets Track* 2021.

PloyGlotFake

- Comparison with other datasets

DataSet	Release Data	Manipulated Modality	Mutilingual	Real video	Fake video	Total video	Manipulation Methods	Techniques labeling	attribute labeling
UADFV [43]	2018	V	No	49	49	98	1	No	No
TIMI [19]	2018	V	No	320	640	960	2	No	No
FF++ [38]	2019	V	No	1,000	4,000	5,000	4	No	No
DFD [38]	2019	V	No	360	3,068	3,431	5	No	No
DFDC [11]	2020	A/V	No	23,654	104,500	128,154	8	No	No
DeeperForensics [16]	2020	V	No	50,000	10,000	60,000	1	No	No
Celeb-DF [23]	2020	V	No	590	5,639	6,229	1	No	No
FFIW [44]	2020	V	No	10,000	10,000	20,000	1	No	No
KoDF [20]	2021	V	No	62,166	175,776	237,942	5	No	No
FakeAVCeleb [18]	2021	A/V	No	500	19,500	20,000	4	No	Yes
DF-Platter [30]	2023	V	No	133,260	132,496	265,756	3	No	Yes
PolyGlotFake	2023	A/V	Yes	766	14,472	15,238	10	Yes	Yes

Hou, Yang, et al. "PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset." *arXiv* 2024.

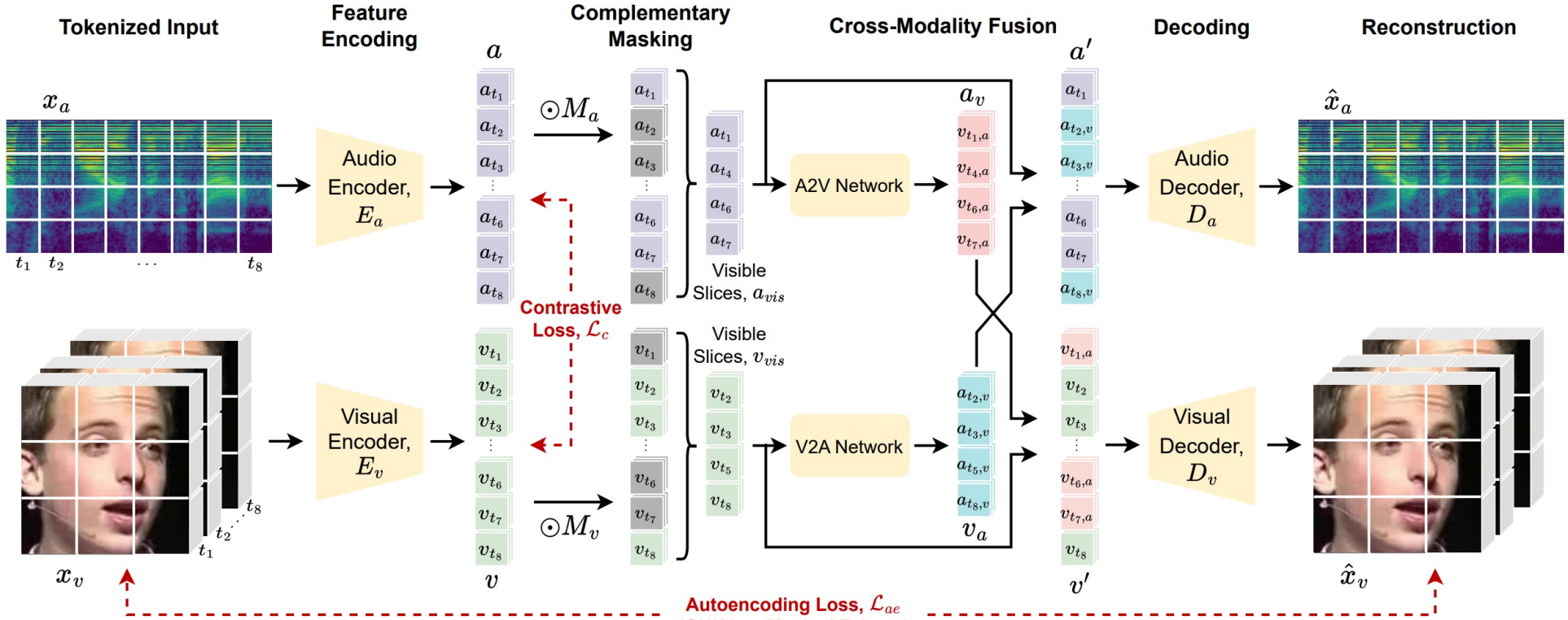
Fusion Methods

- Cross-Attention: Joint AV [Zhou&Lim2021], AVoiD-DF [Yang+2023]
- + Regularization: Cross- and within-modality regularization [Zou+2024]
- + Multi-task: Correlation distillation [Yu+2024], contrastive learning: AVA-CL [Zhang+2024], reconstruction+ contrastive learning: AVFF [Oorloff+2024]

Zhou, Yipin, and Ser-Nam Lim. "Joint audio-visual deepfake detection." *ICCV* 2021.
Yang, Wenyuan, et al. “Avoid-df: Audio-visual joint learning for detecting deepfake.” *TIFS* 2023.
Zou, Heqing, et al. "Cross-Modality and Within-Modality Regularization for Audio-Visual Deepfake Detection." *ICASSP* 2024.
Zhang, Yibo, Weiguo Lin, and Junfeng Xu. “Joint audio-visual attention with contrastive learning for more general deepfake detection.” *TOMM* 2024.
Yu, Cai, et al. “Explicit Correlation Learning for Generalizable Cross-Modal Deepfake Detection.” *ICME* 2024.
Oorloff, Trevine, et al. "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection." *CVPR* 2024.

SoTA Fusion Method: AVFF [Oorloff+2024]

- Contrastive learning and autoencoding objectives on real videos + supervised deepfake classification on real and fake videos



Oorloff, Trevine, et al. "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection." *CVPR* 2024.

SoTA Fusion Method: AVFF [Oorloff+2024]

- Results on FakeAVCeleb

Method	Modality	RVFA		FVRA-WL		FVFA-FS		FVFA-GAN		FVFA-WL		AVG-FV	
		AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
Xception [52]	V	-	-	88.2	88.3	92.3	93.5	67.6	68.5	91.0	91.0	84.8	85.3
LipForensics [21]	V	-	-	**97.8**	97.7	99.9	99.9	61.5	68.1	98.6	98.7	89.4	91.1
FTCN [70]	V	-	-	96.2	97.4	**100.**	**100.**	77.4	78.3	95.6	96.5	92.3	93.1
RealForensics [22]	V	-	-	88.8	93.0	99.3	99.1	99.8	99.8	93.4	96.7	95.3	97.1
AV-DFD [71]	AV	74.9	73.3	97.0	97.4	99.6	99.7	58.4	55.4	**100.**	**100.**	88.8	88.1
AVAD (LRS2) [16]	AV	62.4	71.6	93.6	93.7	95.3	95.8	94.1	94.3	93.8	94.1	94.2	94.5
AVAD (LRS3) [16]	AV	70.7	80.5	91.1	93.0	91.0	92.3	91.6	92.7	91.4	93.1	91.3	92.8
AVFF (Ours)	AV	**93.3**	**92.4**	94.8	**98.2**	**100.**	**100.**	**99.9**	**100.**	99.4	99.8	**98.5**	**99.5**

Oorloff, Trevine, et al. "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection." *CVPR* 2024.

Beyond Fusion: Audio-Video Mismatch

- Emotions don’t lie [Mittal+2020]
- Matching-based learning
  - Person-of-Interest [Cozzolino+2023]
  - Voice-face homogeneity [Cheng+2023]
- Synchronization
  - Temporal synchronization [Feng+2023]
  - AV-Lip-Sync+ [Shahzad+2023]
- Transcription for lip-sync deepfake [Bohacek&Farid2024]

Mittal, Trisha, et al. “Emotions don‘t lie: An audio-visual deepfake detection method using affective cues.” *ACM MM* 2020.
Cozzolino, Davide, et al. "Audio-visual person-of-interest deepfake detection." *CVPRW* 2023.
Cheng, Harry, et al. “Voice-face homogeneity tells deepfake.” *TOMM* 2023.
Feng, Chao, Ziyang Chen, and Andrew Owens. "Self-supervised video forensics by audio-visual anomaly detection." *CVPR* 2023.
Shahzad, Sahibzada Adil, et al. "AV-Lip-Sync+: Leveraging AV-HuBERT to exploit multimodal inconsistency for video deepfake detection." *arXiv* 2023.
Bohacek, Matyas, and Hany Farid. "Lost in Translation: Lip-Sync Deepfake Detection from Audio-Video Mismatch." *CVPRW* 2024.

Example Methods for AV Mismatch

- [Feng+2023]

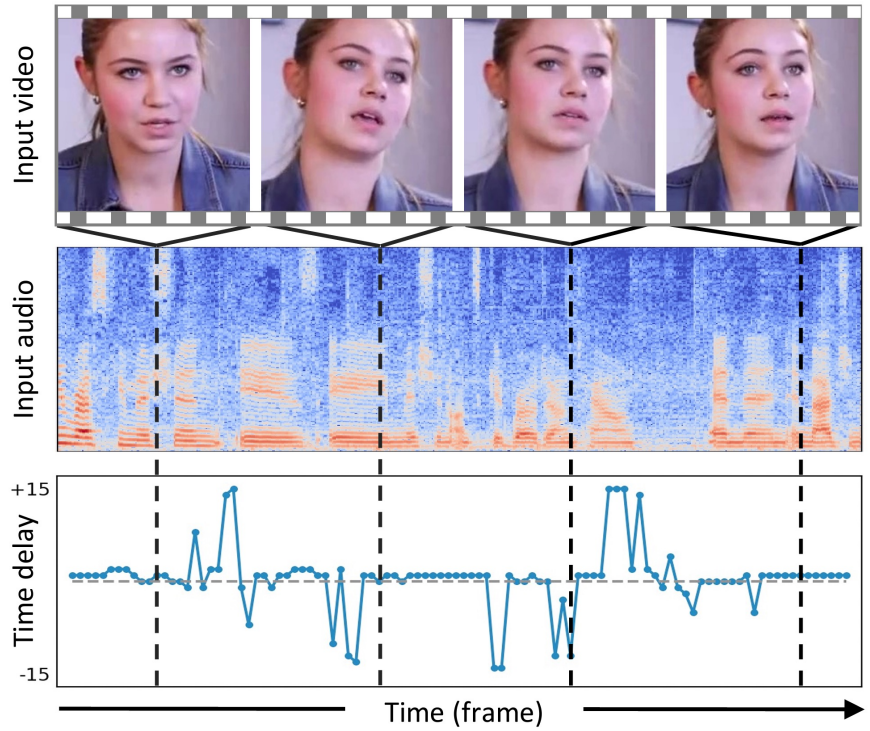
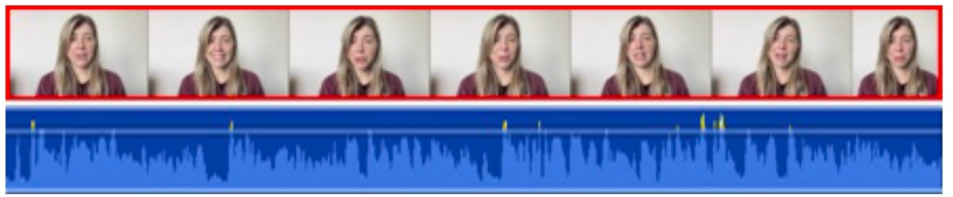


Figure 1. **Audio-visual anomaly detection.** We identify fake videos by finding anomalies in their audio-visual features, using generative models trained entirely on *real* videos. In one variation

- [Bohacek&Farid2024]



video transcription: I just had its bread roll it's your presence about the media in a way

audio transcription: I just think it's really feel good and excellent piece of cinema

manual transcription: I just think it's really feel-good and an excellent piece of cinema

Figure 1. An audio/video clip from a lip-sync deepfake in which the participant responds to the question "what is your favorite movie and why?" The mismatch between the video (lip reading) and audio transcriptions reveals evidence of a lip-sync deepfake.

Temporal Forgery Localization

- LAV-DF [Cai+2022], AV-Deepfake1M [Cai+2023]

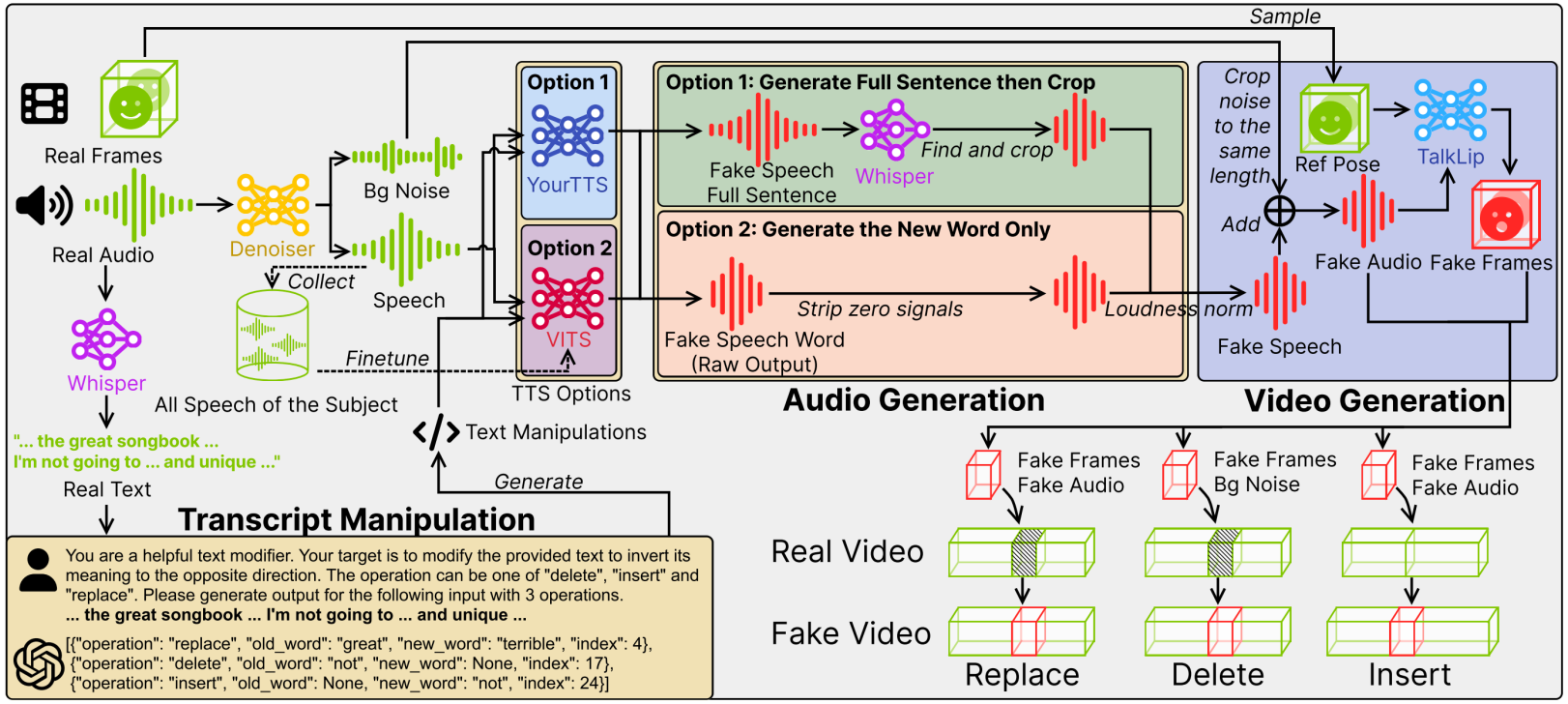
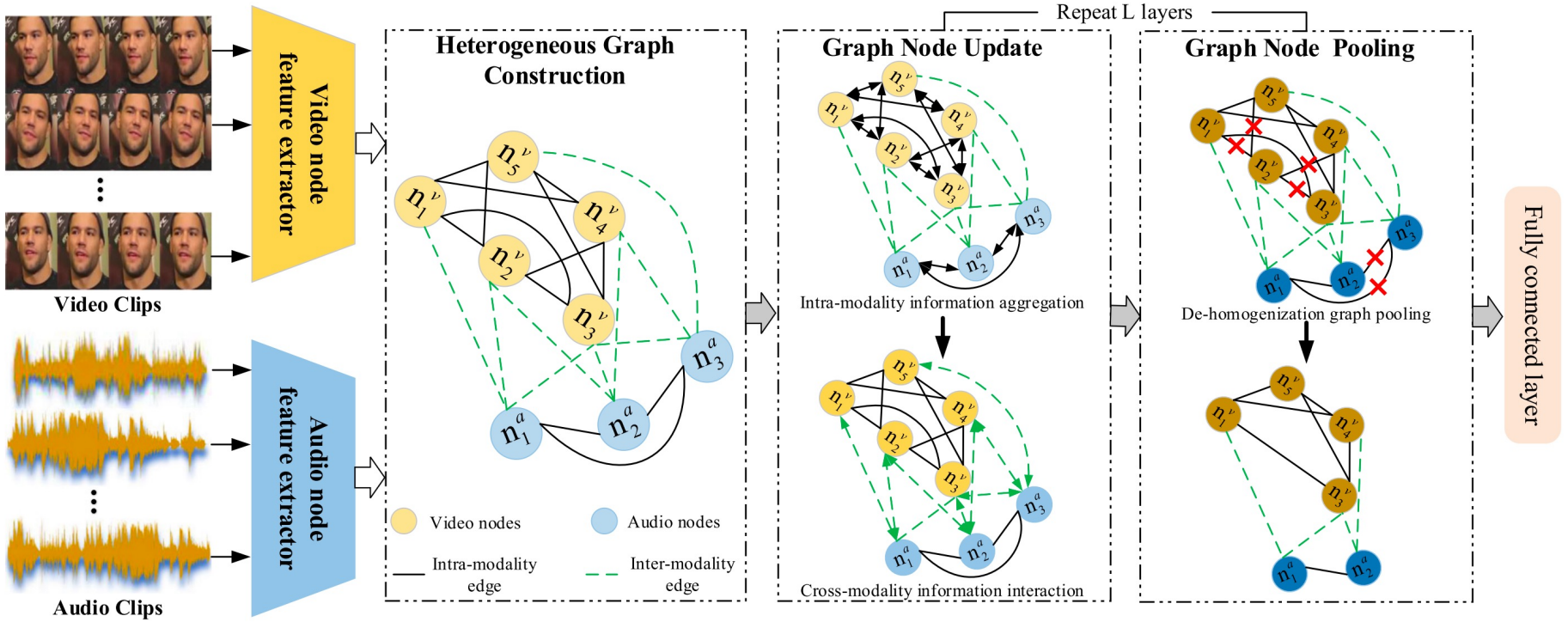


Figure 2. **Data manipulation and generation pipeline.** Overview of the proposed three-stage pipeline. Given a real video, the pre-

Cai, Zhixi, et al. “Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization.” *DICTA* 2022.
Cai, Zhixi, et al. "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset." *arXiv* 2023.

SoTA Methods for Temporal Localization

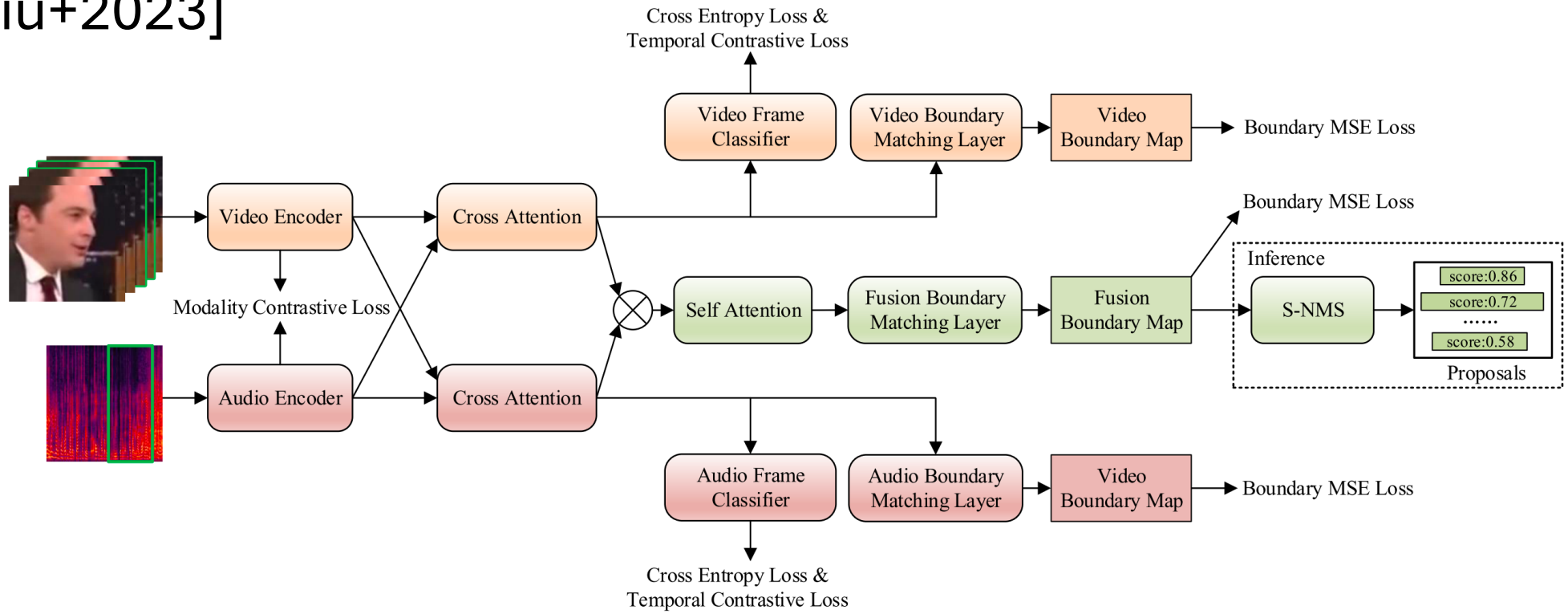
- Heterogeneous graph attention network for intra- and inter-modal relationships both at spatial and temporal scales. [Yin+2024]



Yin, Qilin, et al. "Fine-Grained Multimodal DeepFake Classification via Heterogeneous Graphs." I*JCV* 2024.

SoTA Methods for Temporal Localization

- Embedding-level fusion + multi-dimensional contrastive loss [Liu+2023]



Liu, Miao, et al. "Audio-visual temporal forgery detection using embedding-level fusion and multi-dimensional contrastive loss." *TCSVT* 2023.

General Video AVDD

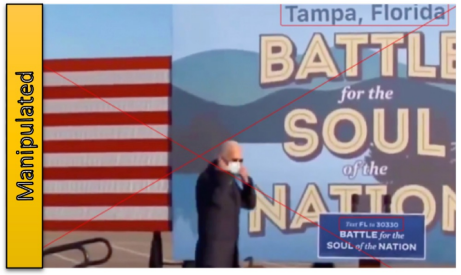
- VideoSham [Mittal+2023]

(a1) The original photo, from Getty Images shows an armed man parked in front of a car.

(b1) This is an original clip of a presidential candidate addressing public in the US state, Minnesota.

(c1) An original image shows three missiles being launched by Iran's government.

(a2) The photo above was altered by digitally placing the armed man in front of a peaceful protest, insinuating violence.



(b2) The clip above is altered by changing the location and the signs on the podium to a different US state, Florida.

(c2) In an altered image released on Iran's Revolutionary Guards website, claimed that 4 missiles were launched simultaneously.

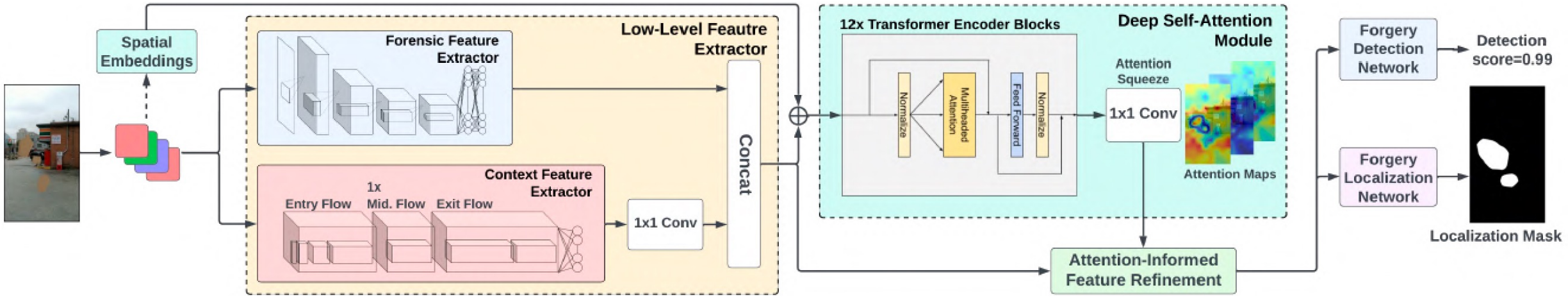
Figure 1: **Spatial manipulations: (a)** [10], **(b)** [42], and **(c)** [44] are examples of videos on social media spatially manipulated with the intent to mislead audiences.

Figure 2: **VIDEOSHAM:** *(top)* VIDEOSHAM consists of diverse, context-rich, and human-centric manipulated videos by professional video editors via 6 spatial and temporal attacks (*e.g.* jersey color change and person removal). *(bottom)* In contrast, deepfake datasets (DF-TIMIT and DFDC) only consist of facial manipulations individual subjects from a close-up angle.

Mittal, Trisha, et al. "Video manipulations beyond faces: A dataset with human-machine analysis." *WACV* 2023.

SoTA Method for General Video AVDD

- VideoFACT: forensic feature embeddings (FFE) and context feature embeddings (CFE)+attention [Nguyen+2024]



Nguyen, Tai D., Shengbang Fang, and Matthew C. Stamm. "Videofact: detecting video forgeries using attention, scene context, and forensic traces." *WACV* 2024.

Summary of AVDD Datasets

- Talking face videos
  - DFDC [Dolhansky+2020]
  - FakeAVCeleb [Khalid+2021]
  - LAV-DF [Cai+2022]
  - AVDeepfake-1M [Cai+2023]
  - PloyGlotFake [Hou+2024]
- General videos
  - VideoSham [Mittal+2023]

Dolhansky, Brian, et al. "The deepfake detection challenge (DFDC) dataset." *arXiv* 2020.
Khalid, Hasam, et al. “FakeAVCeleb: A novel audio-video multimodal deepfake dataset.” *NeurIPS Datasets Track* 2021.
Hou, Yang, et al. "PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset." *arXiv* 2024.
Cai, Zhixi, et al. “Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization.” *DICTA* 2022.
Cai, Zhixi, et al. "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset." *arXiv* 2023.
Mittal, Trisha, et al. "Video manipulations beyond faces: A dataset with human-machine analysis." *WACV* 2023.

Emerging & Future Directions

- Talking faces
  - Generalization to new deepfake techniques, such as lip-to-speech synthesis
  - Joint detection of deepfake and face-voice association
  - Interpretability: which modality is fake, or which generation method is used

- General video
  - Reasoning for audio-visual mismatch with VLM
  - Generalization to more recent Sora with video-to-audio synthesis

- Proactive methods: watermarking
- More modalities for multimedia deepfake detection

Summary of the Tutorial

- Each distinct research topic has rich research questions to solve. Some common interests:
  - Generalization ability
  - More diverse datasets
  - Partial deepfake
  - Interpretability
- Techniques can be borrowed and generalized
- Fusion methods or modality-inconsistency detection for multimedia deepfake detection

Q & A

10-15min

Question 1

- What are the prospects for deepfake detection technology to keep pace with advancements in deepfake generation? What are the potential consequences if it does not?

Question 2

- What are the application scenarios of deepfake detection technology? Based on the application, can we develop something beyond just binary classification?