Alessandro Pianese

On the robust and generalizable detection of synthetic speech tracks

Tutor: Giovanni Poggi

Cycle: XXXVIII Year: Third

Finanziato dall'Unione europea
NextGenerationEU

Candidate's information

- **MSc degree in** Computing Science – Intelligent system and visual computing – University of Groningen
- **Research group:** GRIP (Image Processing Research Group)
- **PhD start date:** 01/01/2025
- **PhD end date:** 31/12/2025
- **Scholarship type:** funded by the NextGenerationEU initiative under the PNRR agreement
- **Periods abroad**
  - 01/12/2024 – 28/02/2025 at Fraunhofer Institute for Digital Media Technology, Germany
  - 01/05/2025 – 28/07/205 at Recod.ai Lab, University of Campinas, Brazil

Summary of study activities

PhD year	Courses	Seminars	Research	Tutorship
1st	23	7.2	33.3	0.00
2nd	13	5	35.8	0.54
3rd	5.4	0.4	58.6	0.00
Total	41.4	12.6	127.7	0.54

- **PhD Schools**:
  - IEEE-SPS Summer School 2023 "Summer School on Metaverse Technologies" – Cagliari (CA), Italia
  - IEEE-SPS Summer School 2024 "Understanding and modeling the world around us" – Capri (NA), Italia
  - IEEE-SPS Summer School 2025 "From Foundational Models to Multimedia Signal Processing: A deep dive in Multimodal AI" – San Vincenzo (LI), Italia
- **PhD courses**:
  - Using Deep Learning Properly – Dr. Andrea Apicella
  - *How to boost your PhD* - Prof. Antigone Marino
  - *Strategic Orientation for STEM Research & Writing* - Dr. Chie Shin Fraser
  - Innovation and Entrepeneurship – Prof. Pierluigi Rippa
- **MSc courses**:
  - *Visione per sistemi robotici* - Prof. Davide Cozzolino
- **Conferences**:
  - IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), Vancouver, Canada, Jun. 18-22, 2023
  - IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP**), Seoul, South Korea, Apr. 14-19, 2024
  - ACM Workshop on Information Hiding and Multimedia Security (**IHMM&Sec**), Spain, Baiona, Jun. 24-26, 2024

Research field of interest

- **Multimedia Forensics**:
  - Analysis of media forensic clues
- **Multimodal Deepfake Detection**:
  - Has the video or the audio been tampered with?
- **Audio Synthesis detection**:
  - Has this audio been generated? Is the speaker who it claims to be?

Is this real or fake?

Fake Fake Fake Real

Research results

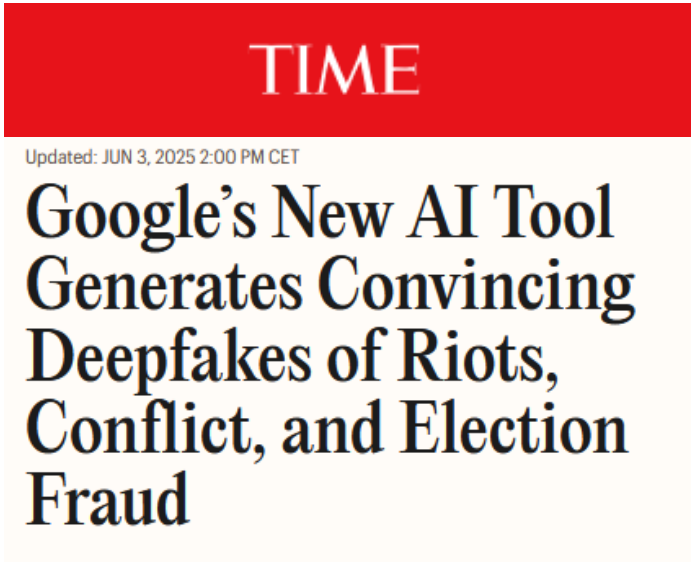
- **POI-Forensics**
  - Development of a audio-video deepfake detector based on the POI framework
- **Training-Free**
  - Focus on the audio branch by including a large language model as the feature extractor
- **X-POI-Audio**
  - Pointing toward future explainable solutions by leveraging only speaker nuances

Research results

[P1]	D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, *Audio-visual person-of-interest deepfake detection*, **The Multimedia Forensics Workshop (WMF) at IEEE/CVF conference on computer vision and pattern recognition (CVPR)**, Vancouver, Canada, June 2023 pp. 943-952, IEEE/CVF
[P2]	A. Pianese, D. Cozzolino, G. Poggi, and L. Verdoliva, *Training-free deepfake voice recognition by leveraging large-scale pre-trained models*, **ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)**, Baiona, Spain, June 2024, pp. 289-294
[P3]	A. Pianese, L. Cuccovillo, G. Poggi, T. L. Roux and P. Aichroth, *Towards Explainable Person-of-Interest-based Audio Synthesis Detection*, **VERIMEDIA workshop, International Joint Conference on Neural Networks (IJCNN)**, Rome, Italy, 2025, pp. 1-8

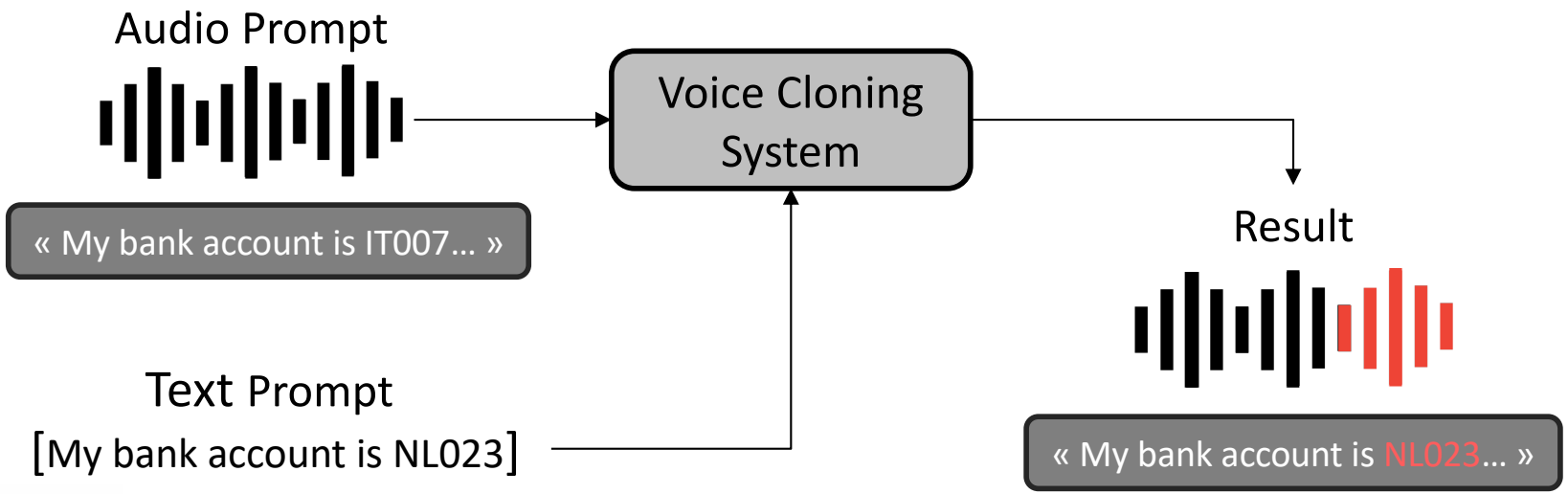
Why is this field important?

- Generated content has been spreading online at an astonishing pace



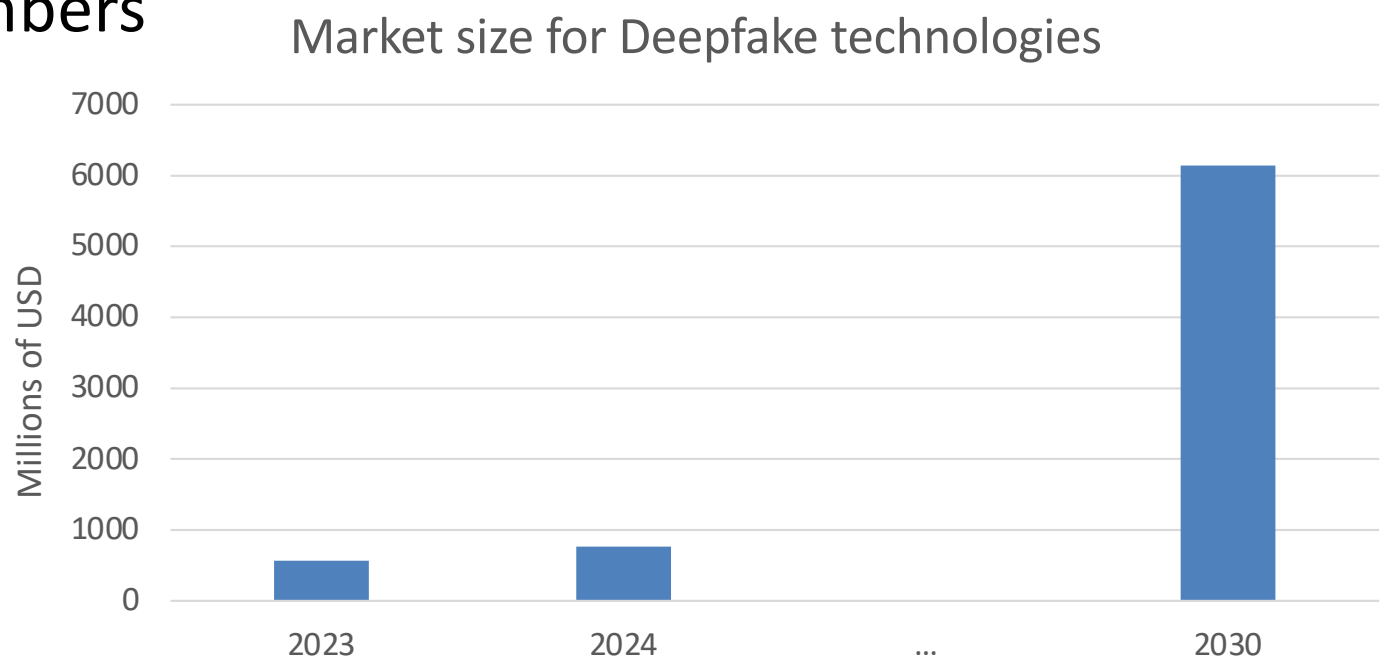
Why is this field important?

- Generated content has been spreading online at an astonishing pace
- It has become easier than ever to generate this content



Why is this field important?

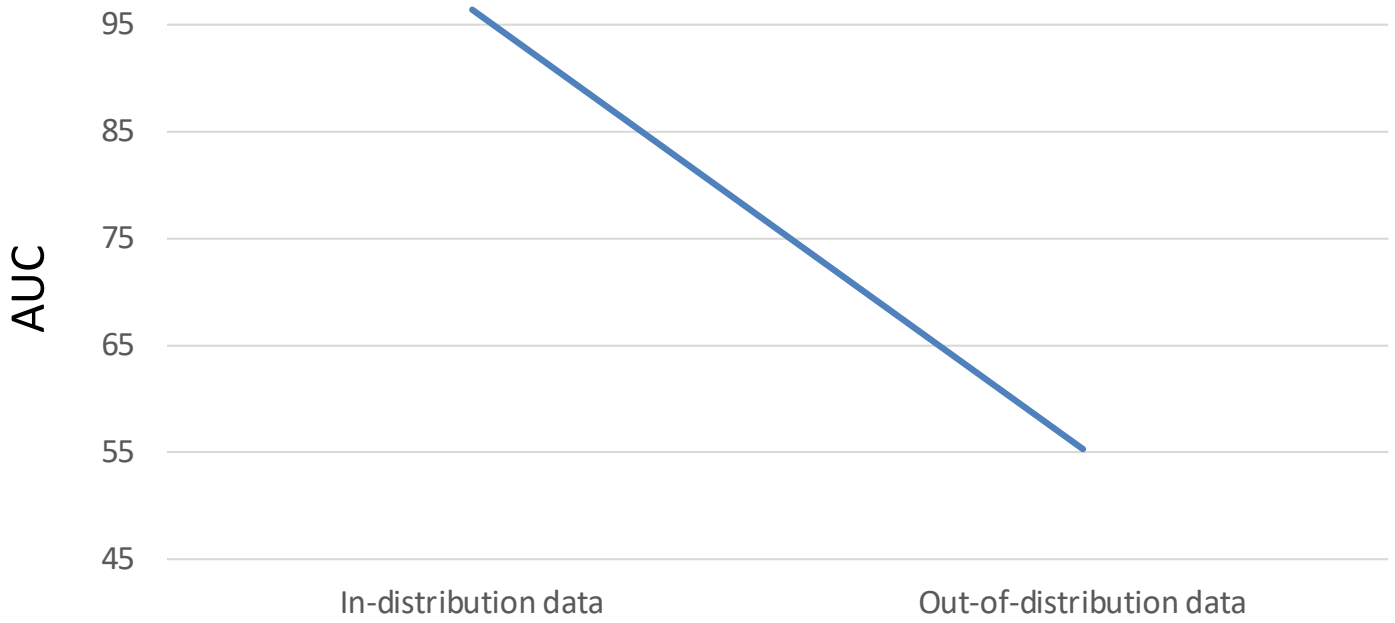
- Generated content has been spreading online at an astonishing pace
- It has become easier than ever to generate this content
- This kind of content is expected to exponentially grow in numbers



[1] https://www.grandviewresearch.com/industry-analysis/deepfake-ai-market-report

The problem with current approaches

- State of the art approaches use supervised training strategies
- Great results on in-distribution data but issues to generalize to unseen data



What we do differently

- We adopted and further developed the person of interest (POI) approach

Reference Set

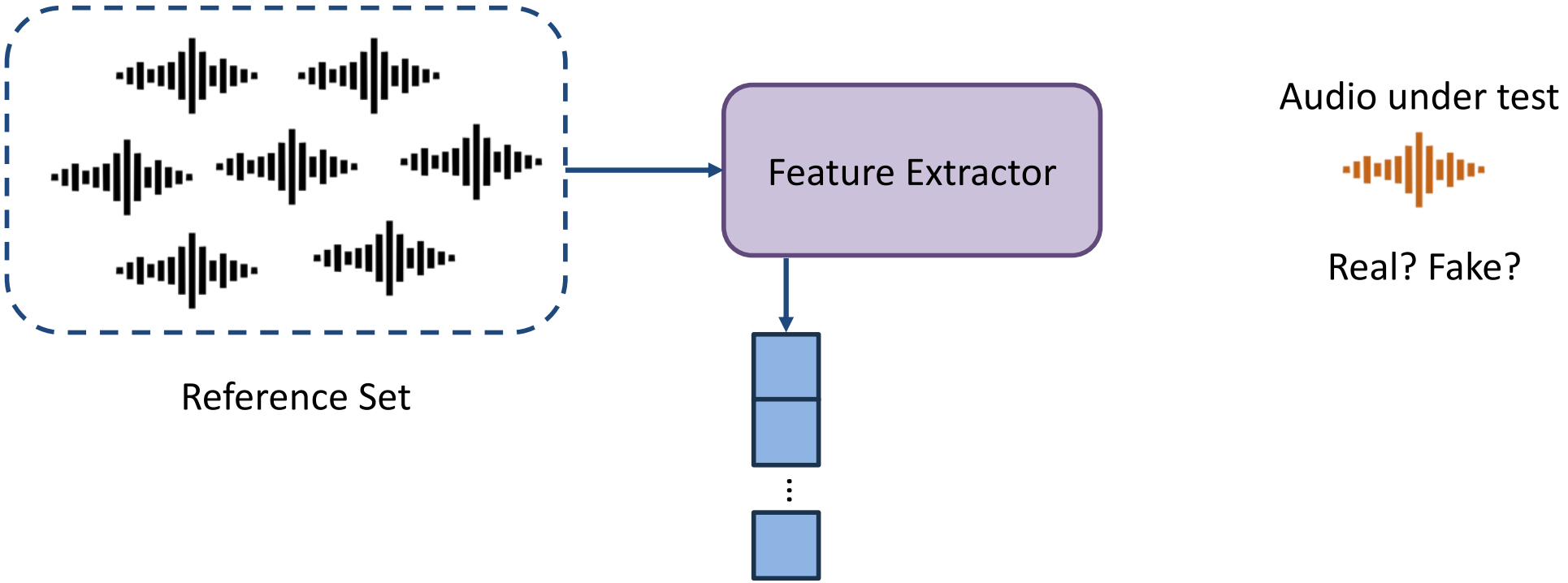
Feature Extractor

Audio under test

Real? Fake?

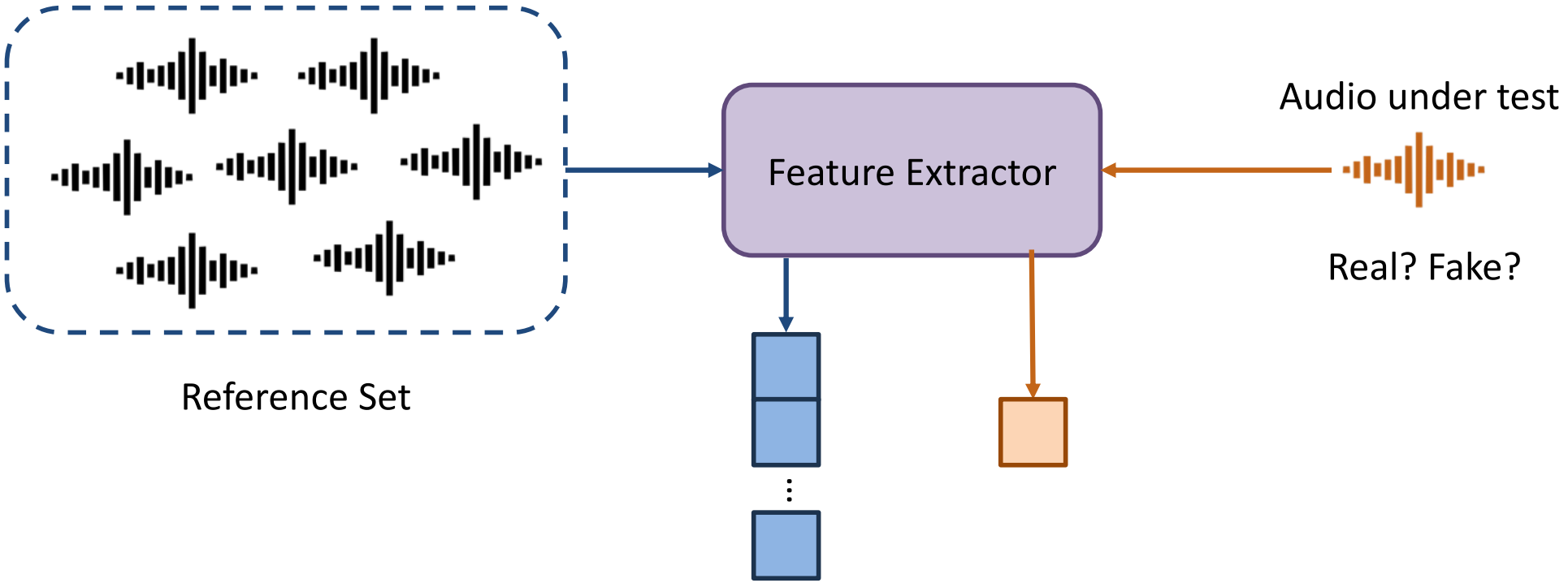
What we do differently

- We adopted and further developed the person of interest (POI) approach



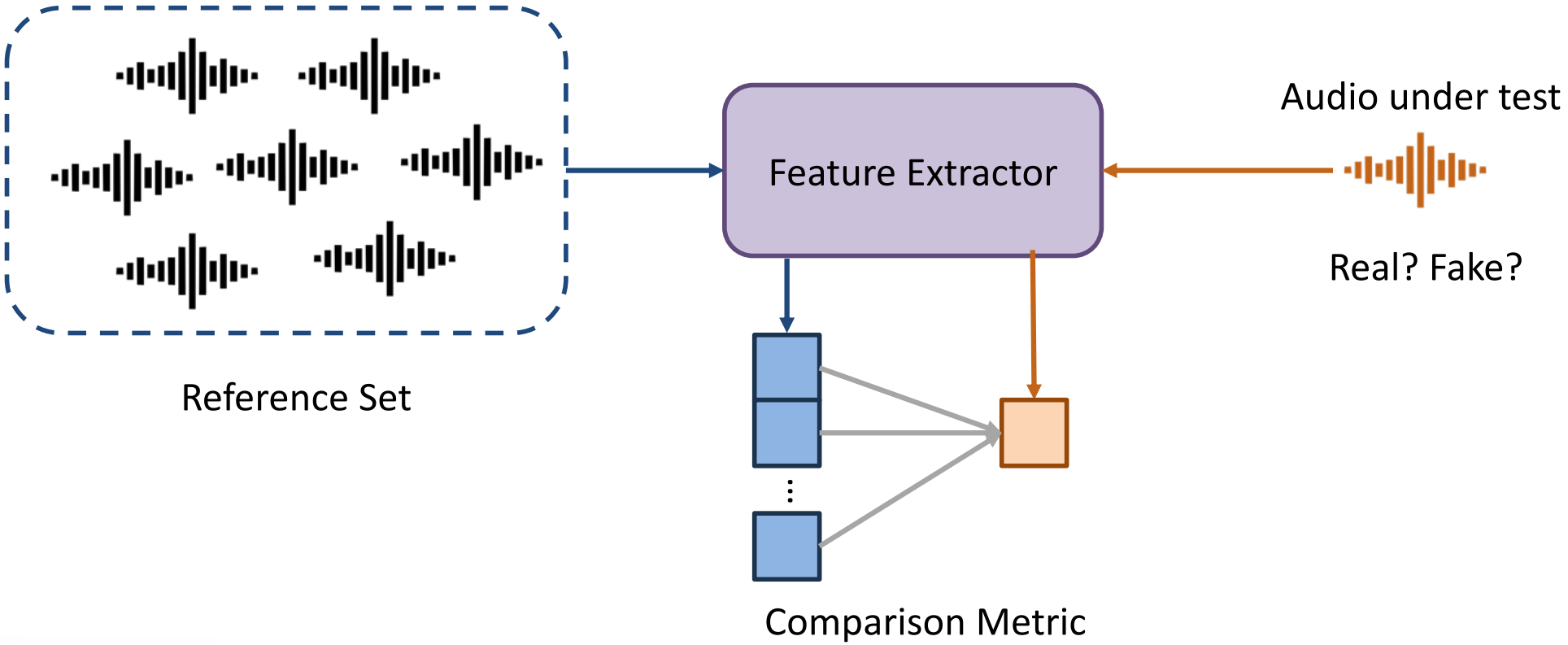
What we do differently

- We adopted and further developed the person of interest (POI) approach



What we do differently

- We adopted and further developed the person of interest (POI) approach

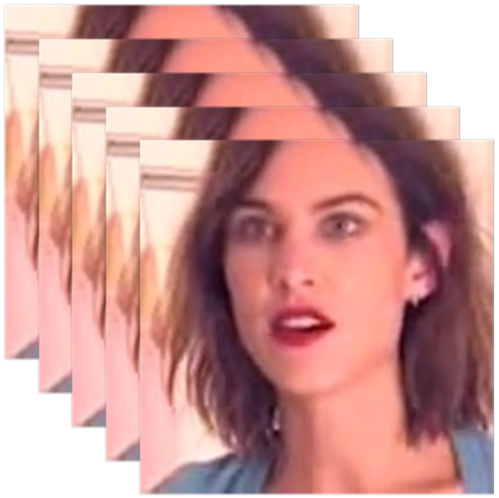


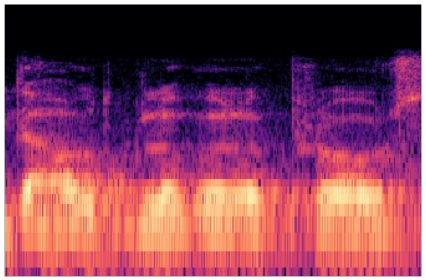
What we do differently

- We only need to train the feature extractor
- We train with a contrastive learning loss that creates cluster for each class
- We use only real files during training
- Generalization is automatically ensured
  - The extractor is fake agnostic since it never sees them during training
- Robust to perturbations
  - The loss pushes the network to look for semantic cues, ignoring compression artifacts

PhD Thesis: POI-Forensics

- How do we apply such framework in the audio-visual domain?
- We need to train a feature extractor

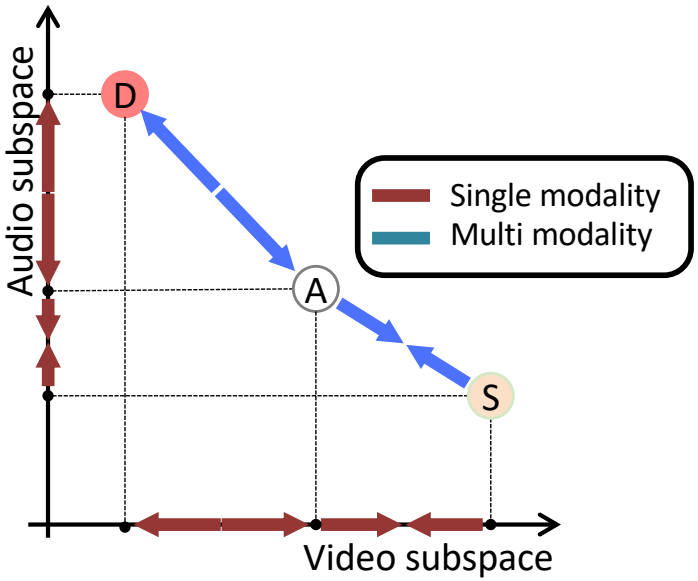




PhD Thesis: POI-Forensics

PhD Thesis: POI-Forensics

- Contrastive learning loss
  - Training forces embedded vectors of a reference video (A) to be close to vectors of the same subject (S) but far from those of different subjects (D)



PhD Thesis: POI-Forensics

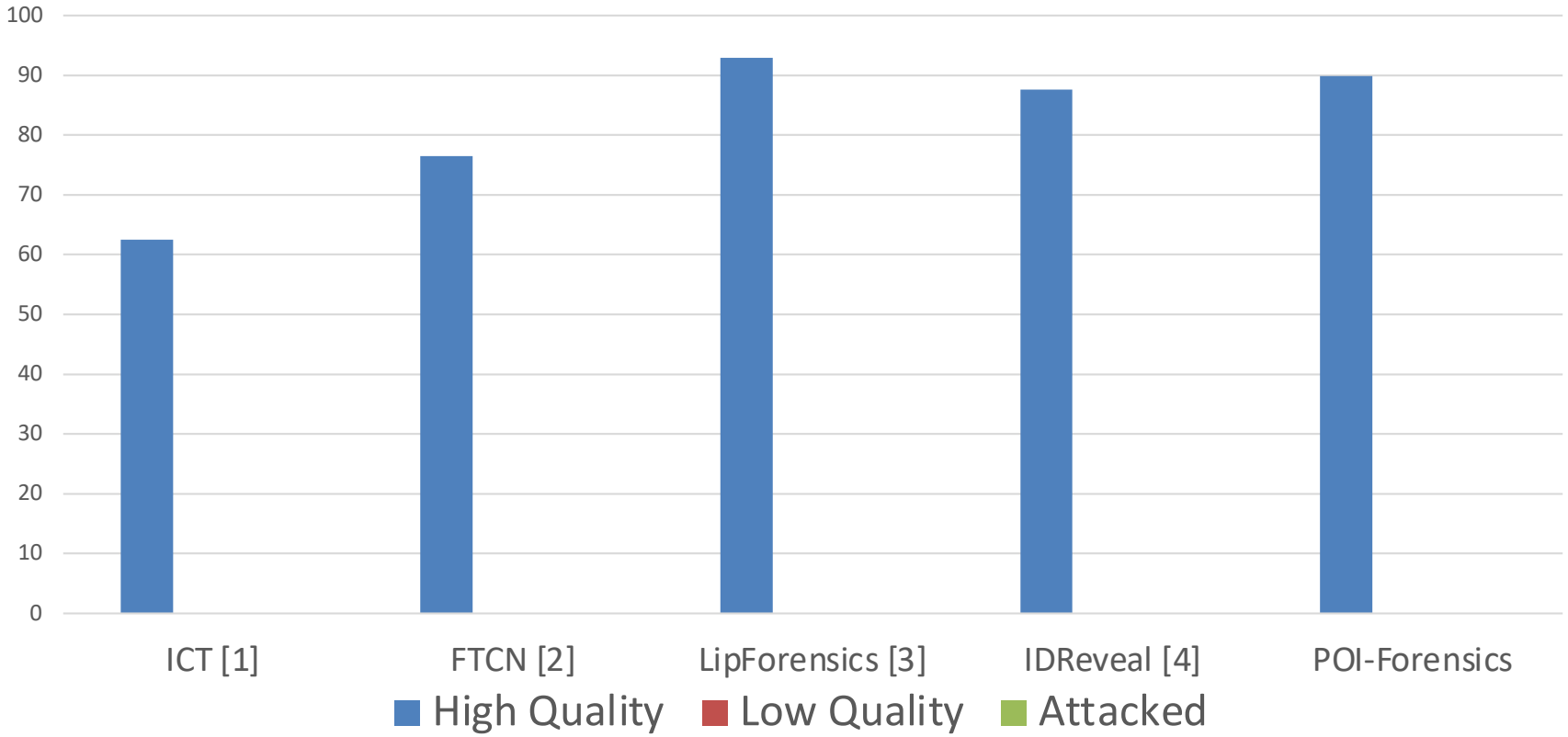
- Comparison with state-of-the-art (high quality)

AUC/ACC	pDFDC	DF-TIMIT	FakeAVCel.	KoDF	AVG
ICT [1]	77.1/70.7	87.8/77.1	68.2/63.9	62.5/58.9	73.9/67.7
FTCN [2]	72.3 / 63.9	**100.** / 87.4	84.0 / 64.9	76.5 / 63.0	83.2 / 69.8
LipForensics [3]	68.7 / 60.0	98.8 / 78.0	**97.6** / 83.3	**92.9** / 56.1	89.5 / 69.3
ID-Reveal [4]	91.3 / 80.4	99.0 / **92.8**	70.2 / 60.3	87.6 / 63.7	87.0 / 74.3
POI-Forensics	**95.2** / **86.7**	99.2 / 85.7	94.1 / **86.6**	89.9 / **81.1**	**94.6** / **85.0**

[1] Dong, X., et al. "Protecting celebrities from deepfake with identity consistency transformer." Proceedings of the IEEE/CVF CVPR. 2022.
[2] Zheng, Y., et al. "Exploring temporal coherence for more general video face forgery detection." Proceedings of the IEEE/CVF ICCV. 2021.
[3] Haliassos, A., et al. "Lips don't lie: A generalisable and robust approach to face forgery detection." Proceedings of the IEEE/CVF CVPR. 2021.
[4] Cozzolino, D., et al. "Id-reveal: Identity-aware deepfake video detection." Proceedings of the IEEE/CVF ICCV. 2021.

PhD Thesis: POI-Forensics

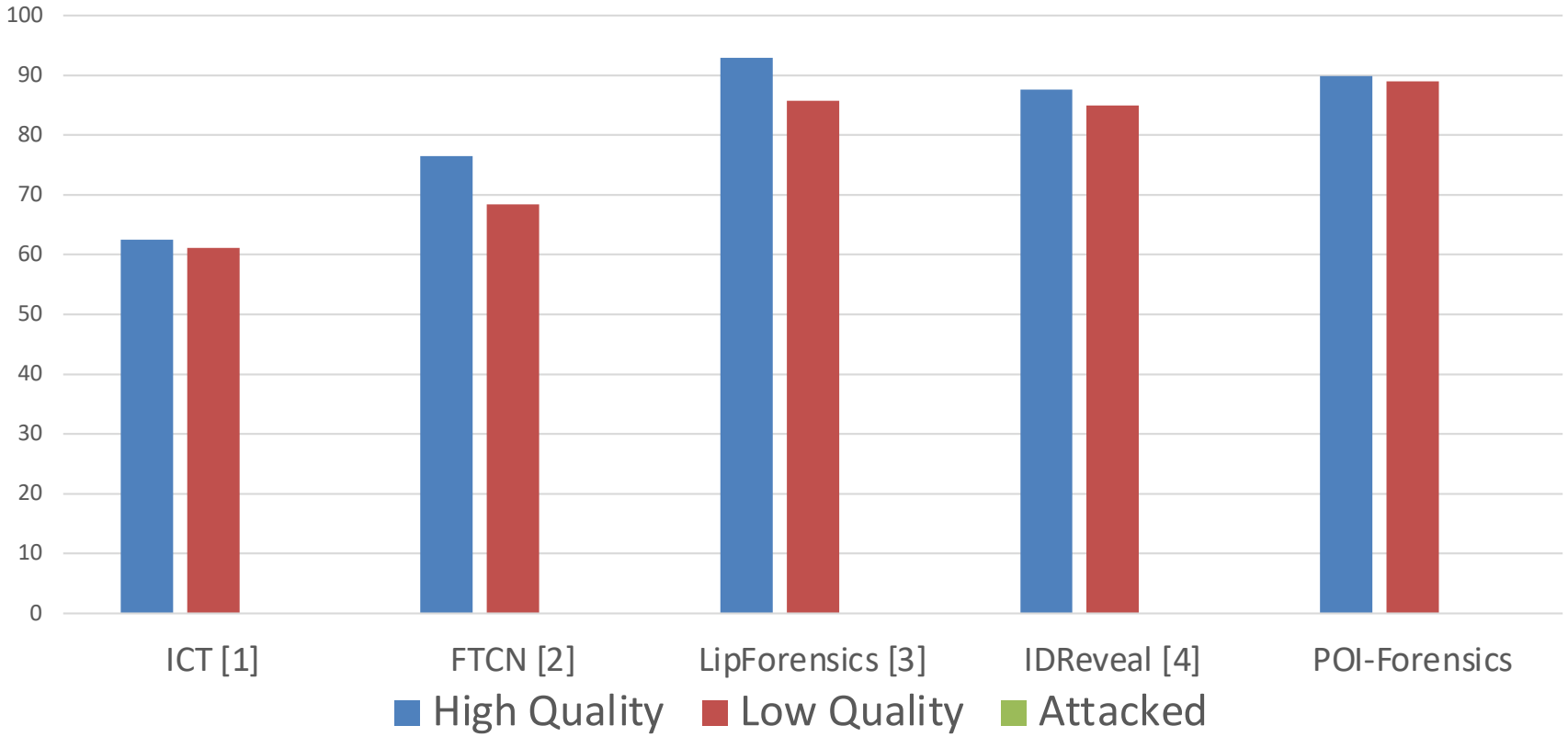
- Rosbustness analysis



[1] Dong, X., et al. "Protecting celebrities from deepfake with identity consistency transformer." Proceedings of the IEEE/CVF CVPR. 2022.
[2] Zheng, Y., et al. "Exploring temporal coherence for more general video face forgery detection." Proceedings of the IEEE/CVF ICCV. 2021.
[3] Haliassos, A., et al. "Lips don't lie: A generalisable and robust approach to face forgery detection." Proceedings of the IEEE/CVF CVPR. 2021.
[4] Cozzolino, D., et al. "Id-reveal: Identity-aware deepfake video detection." Proceedings of the IEEE/CVF ICCV. 2021.

PhD Thesis: POI-Forensics

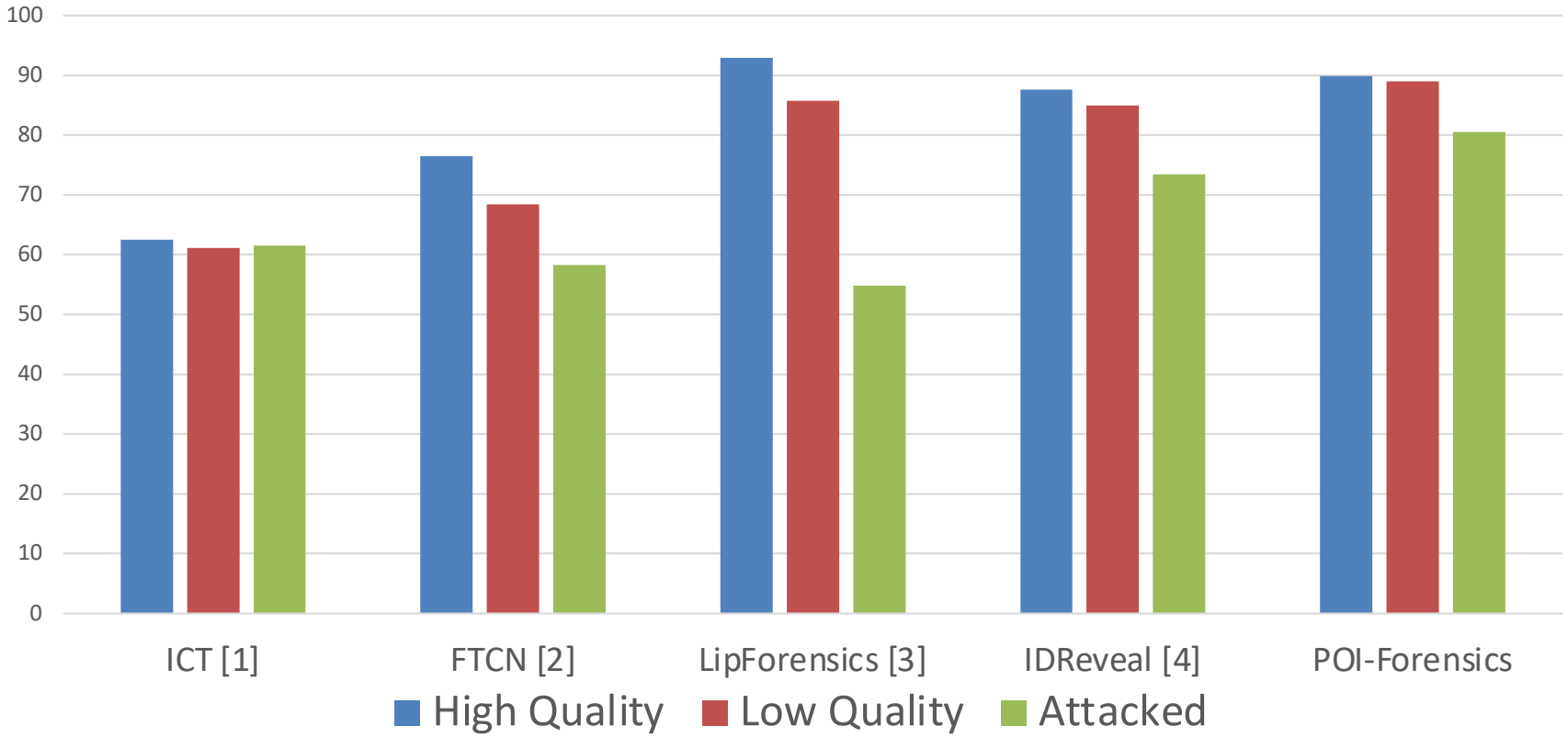
- Rosbustness analysis



[1] Dong, X., et al. "Protecting celebrities from deepfake with identity consistency transformer." Proceedings of the IEEE/CVF CVPR. 2022.
[2] Zheng, Y., et al. "Exploring temporal coherence for more general video face forgery detection." Proceedings of the IEEE/CVF ICCV. 2021.
[3] Haliassos, A., et al. "Lips don't lie: A generalisable and robust approach to face forgery detection." Proceedings of the IEEE/CVF CVPR. 2021.
[4] Cozzolino, D., et al. "Id-reveal: Identity-aware deepfake video detection." Proceedings of the IEEE/CVF ICCV. 2021.

PhD Thesis: POI-Forensics

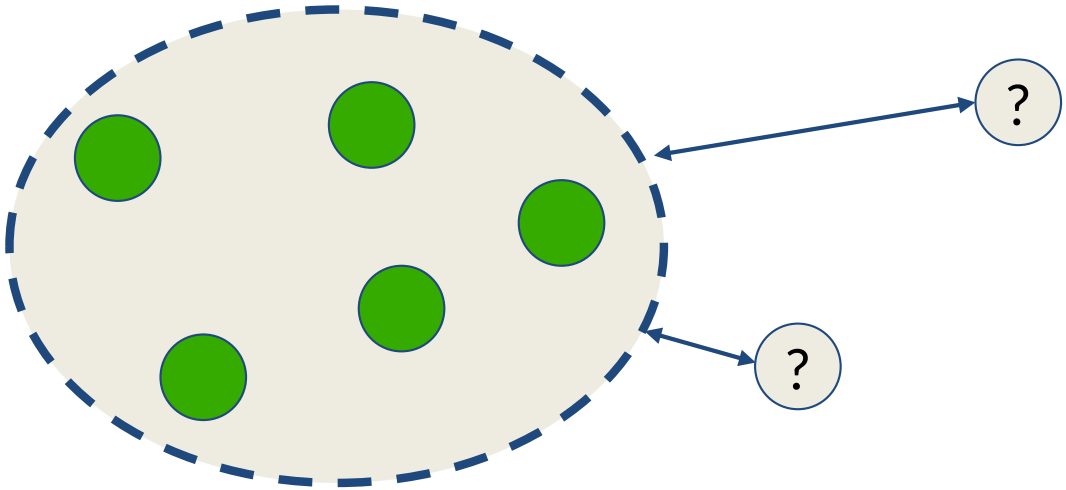
- Rosbustness analysis



[1] Dong, X., et al. "Protecting celebrities from deepfake with identity consistency transformer." Proceedings of the IEEE/CVF CVPR. 2022.
[2] Zheng, Y., et al. "Exploring temporal coherence for more general video face forgery detection." Proceedings of the IEEE/CVF ICCV. 2021.
[3] Haliassos, A., et al. "Lips don't lie: A generalisable and robust approach to face forgery detection." Proceedings of the IEEE/CVF CVPR. 2021.
[4] Cozzolino, D., et al. "Id-reveal: Identity-aware deepfake video detection." Proceedings of the IEEE/CVF ICCV. 2021.

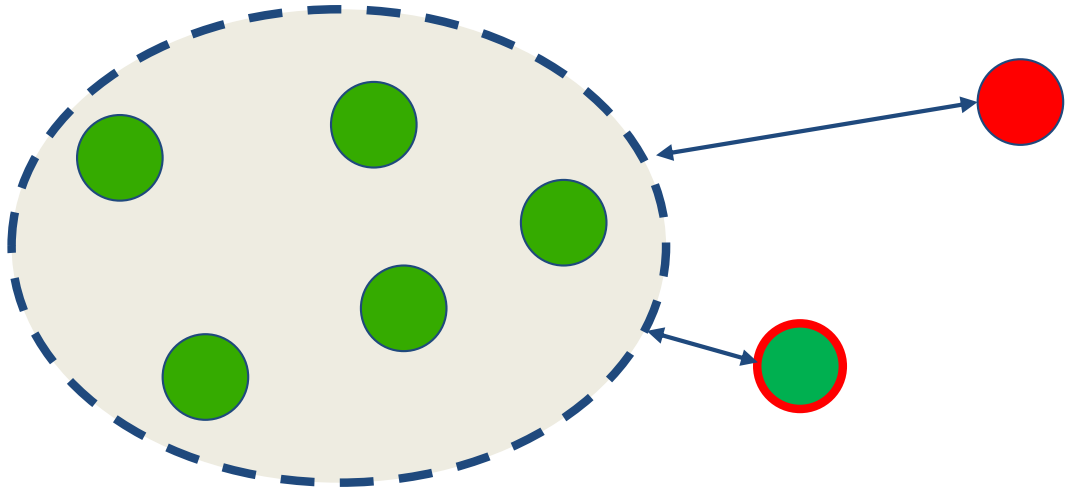
PhD Thesis: Training-free

- With one class approaches, generalization is guaranteed as models are trained only on real data



PhD Thesis: Training-free

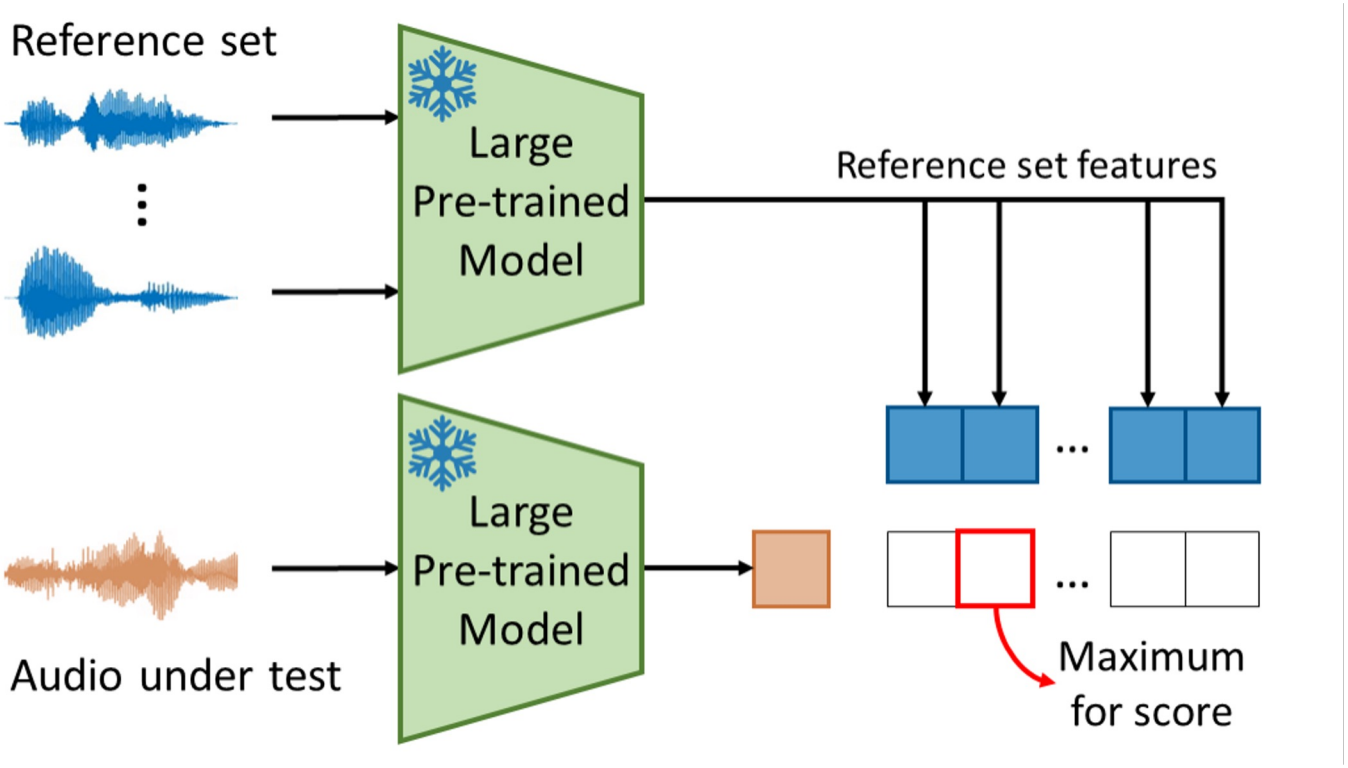
- With one class approaches, generalization is guaranteed as models are trained only on real data



- Good modeling of the real class is required!
- How can we improve on this aspect?

PhD Thesis: Training-free

- What if we leverage the power of large language models?
- Our approach does not require knowledge of fake data, generalization is automatically ensured



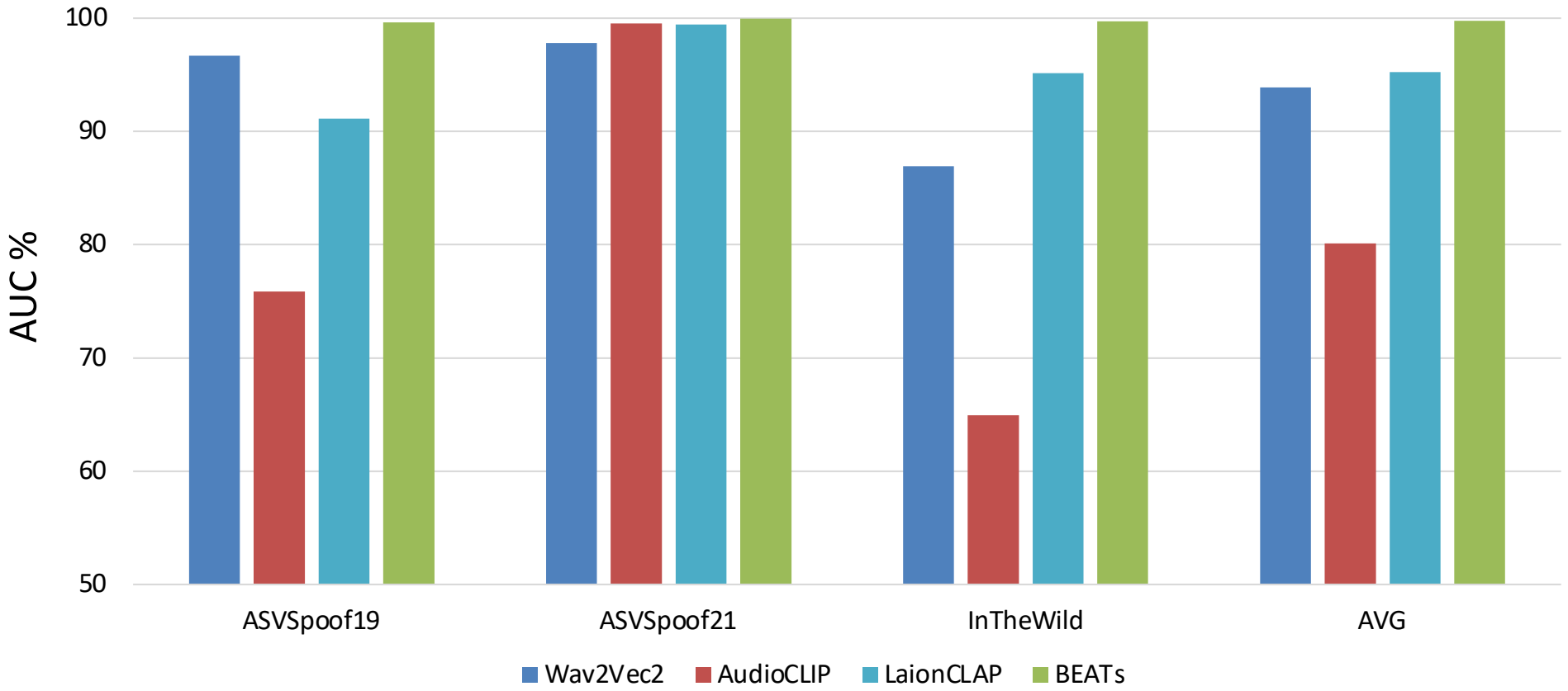
PhD Thesis: Training-free

- The models we compared all possess a very broad pre-training objective

Name	Network	# Params	Training size (h)	Training Task
Wav2Vec2	CNN+Transformer	2B	436K	Speech Recognition
AudioCLIP	CNN+Transformer	134M	5K	Latent Representation Learning
LaionCLAP	Transformer	158M	9K	Latent Representation Learning
BEATs	Transformer	90M	5K	Latent Representation Learning

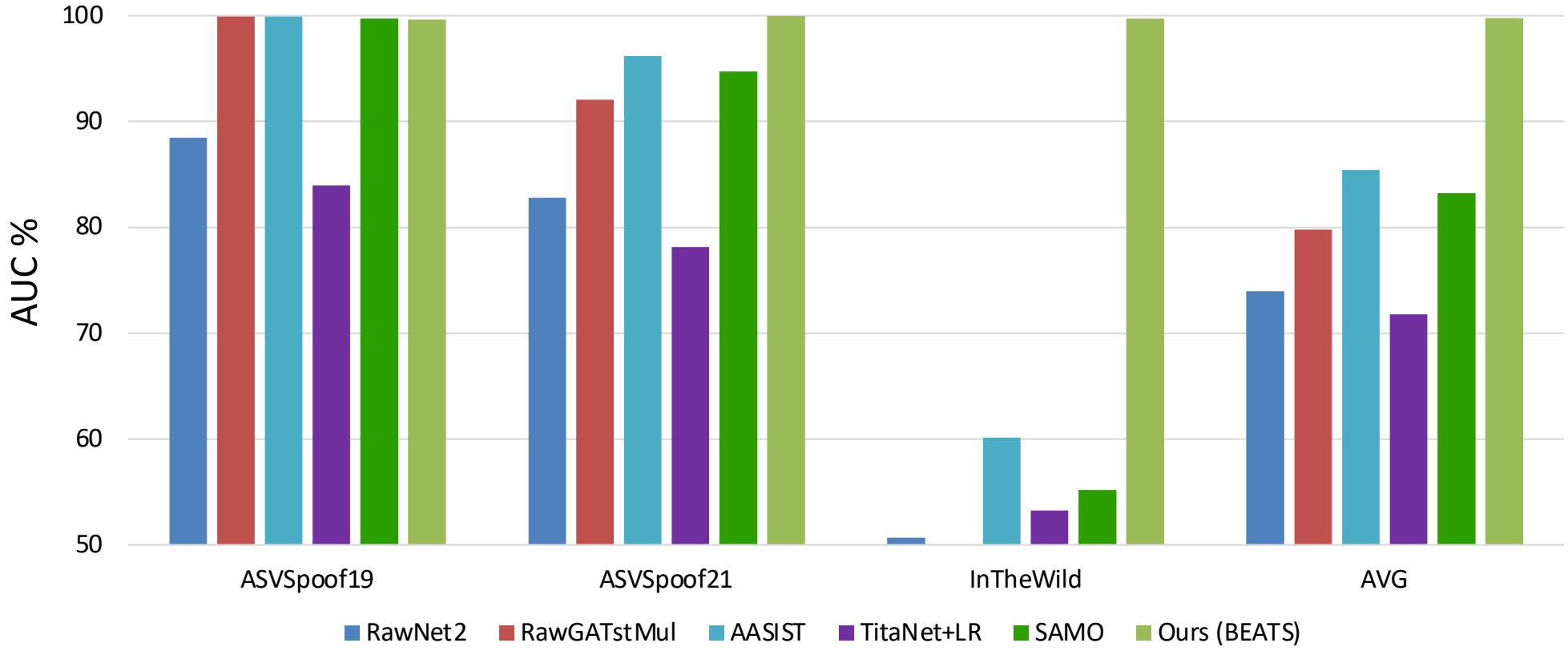
PhD Thesis: Training-free

- Comparison of four state of the art audio pre-trained models
- They all possess a very broad pre-training objective



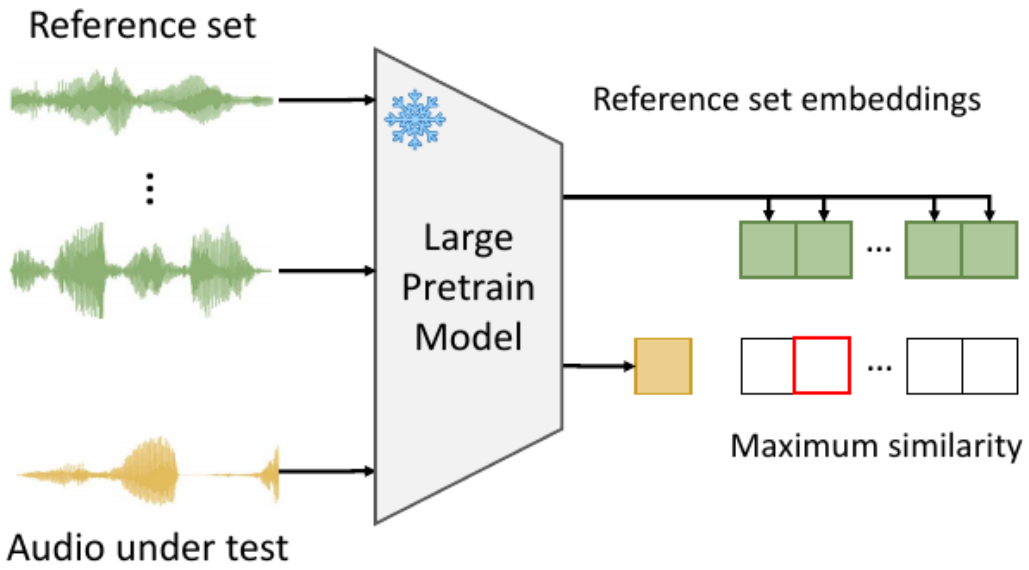
PhD Thesis: Training-free

- Comparison of our method and SOTA supervised detectors
- Our proposal can generalize on unseen data



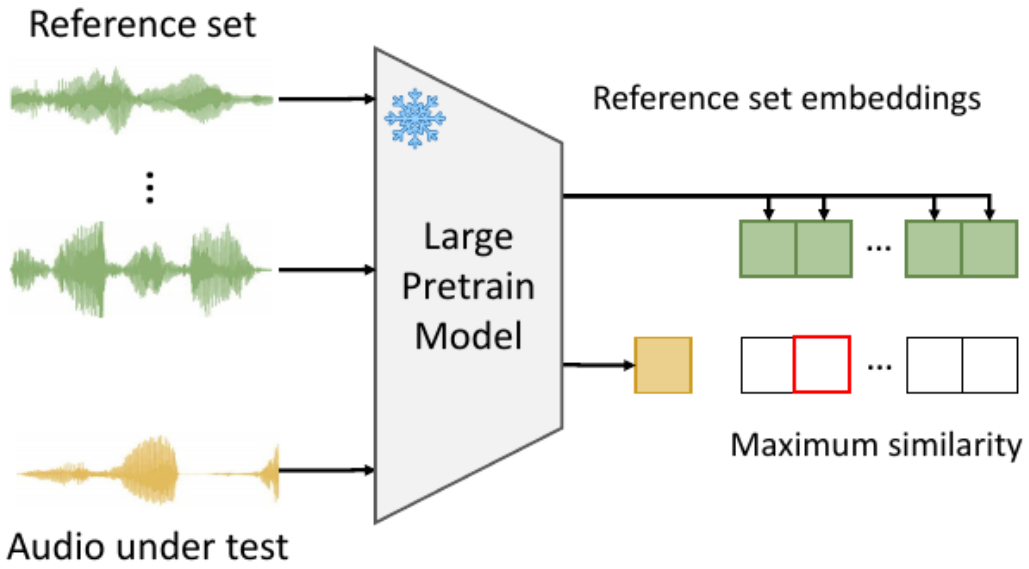
PhD Thesis: X-POI-Audio

Can we improve upon this while retaining the positive aspects?



PhD Thesis: X-POI-Audio

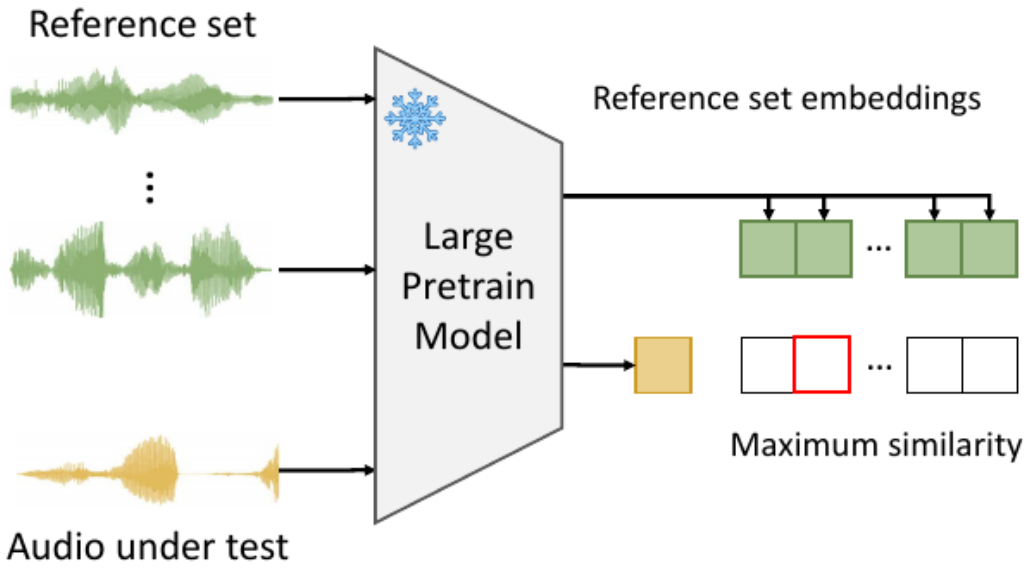
Can we improve upon this while retaining the positive aspects?



- It works better !
- Generalizes well

PhD Thesis: X-POI-Audio

What can we do differently?

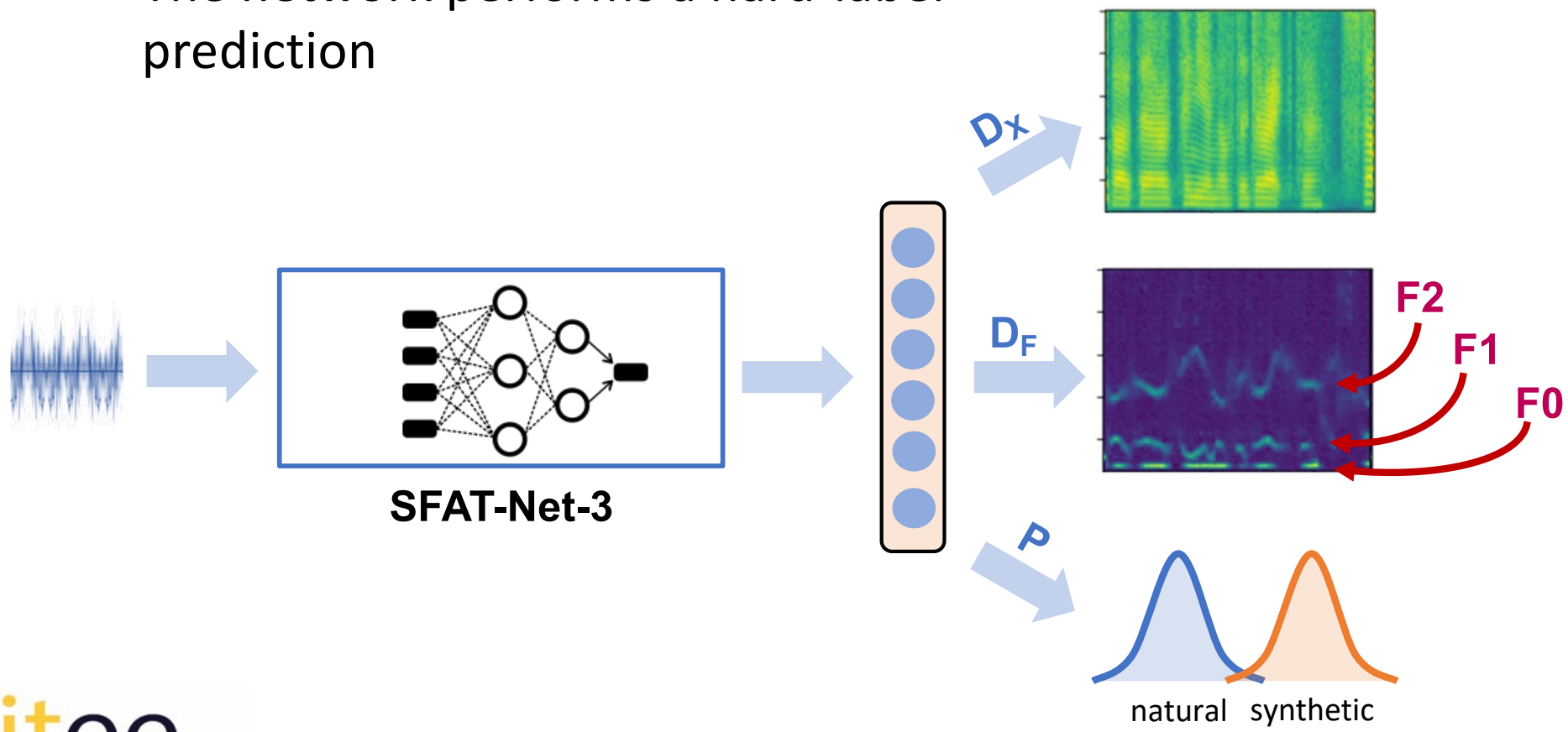


- It works better !
- Generalizes well

- Sensitive to references
- Person-specific threshold
- Explainability ?

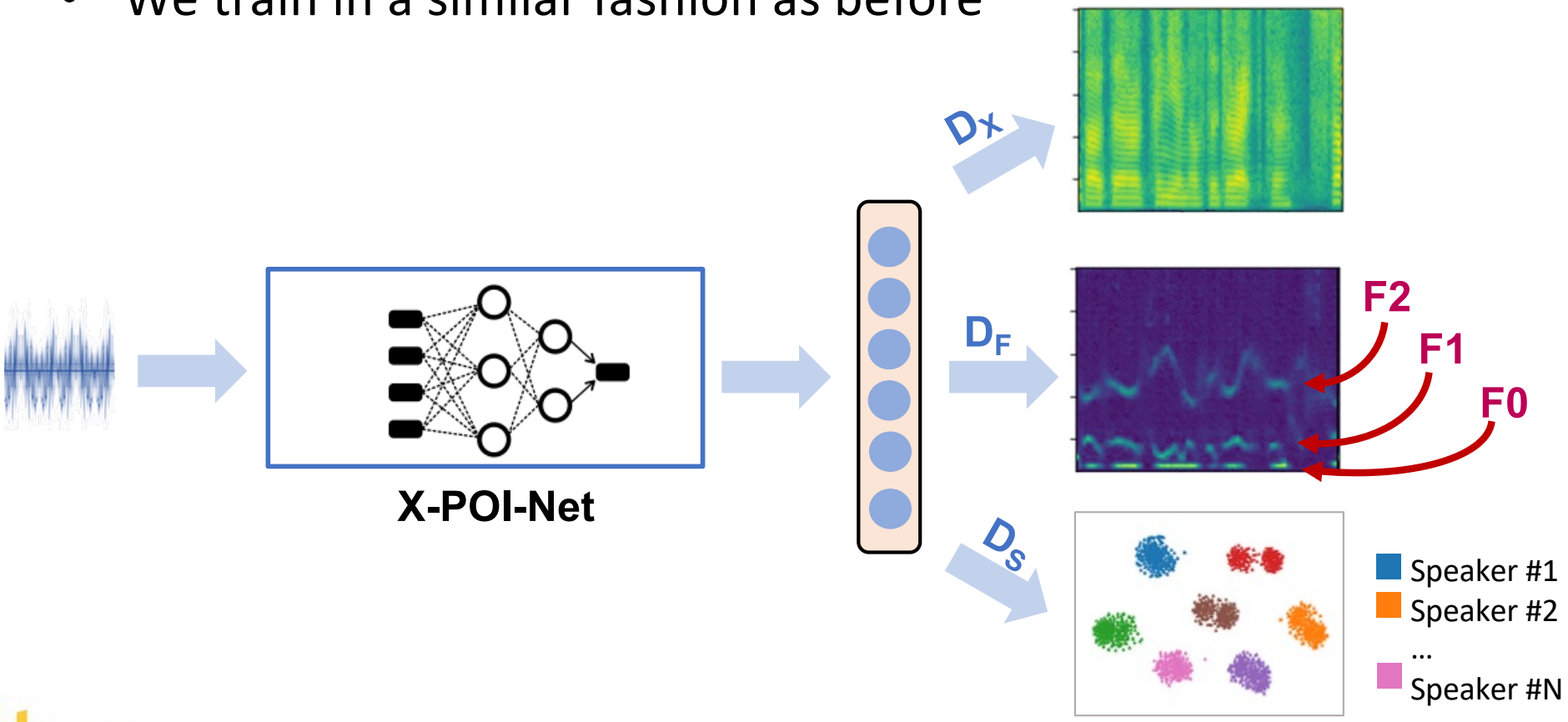
PhD Thesis: X-POI-Audio

- We took inspiration from SFAT-Net-3 multiple training objective
- The network performs a hard-label prediction



PhD Thesis: X-POI-Audio

- We swap the classification branch with a prediction one
- It returns a speaker embedding
- We train in a similar fashion as before



PhD Thesis: X-POI-Audio

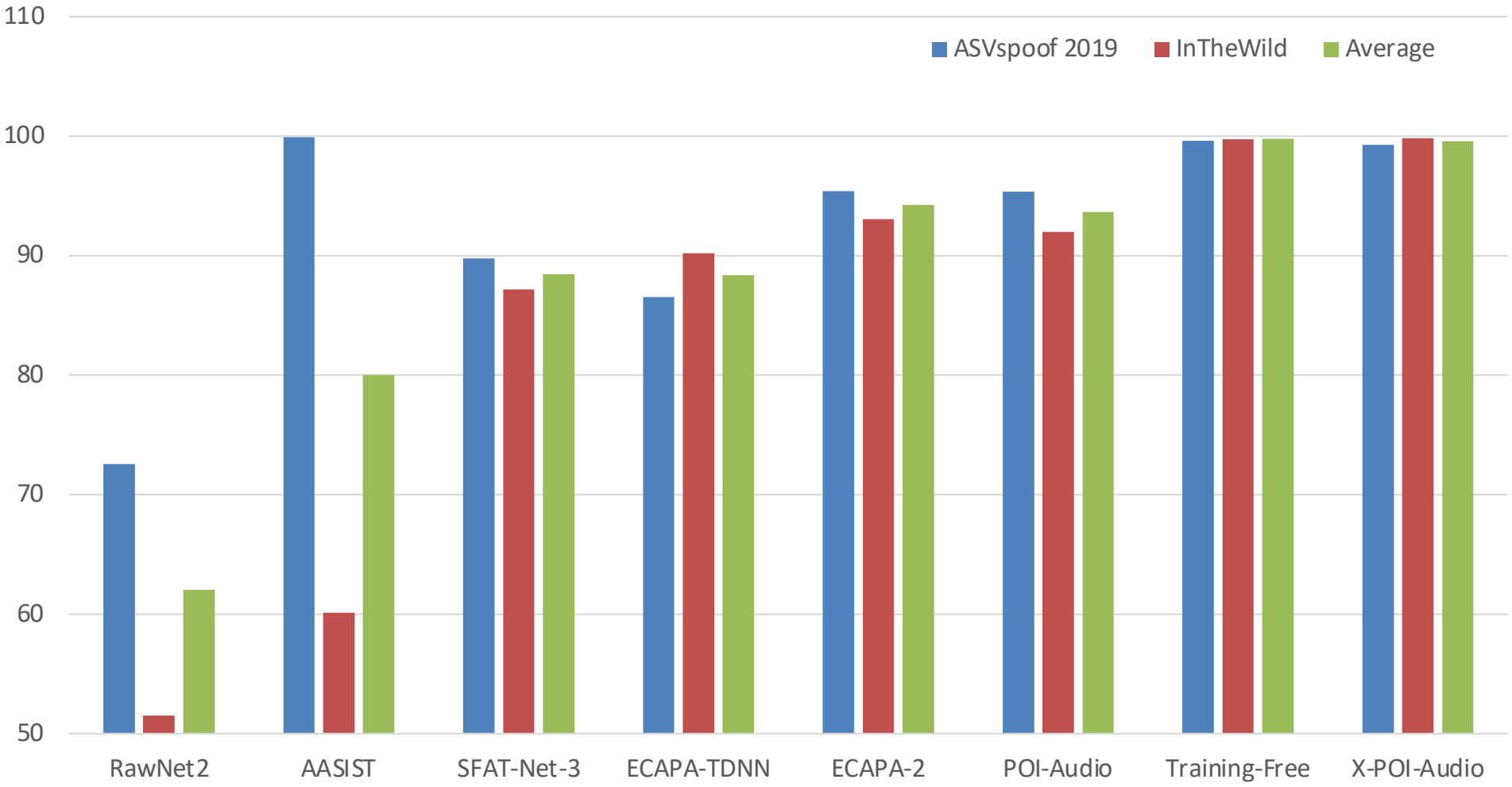
Positive aspects

- Explainable by design
- Leverages Phonetics Know-how

Negative aspects

- Language specific – e.g., French vs English
- Lower performance than SotA speaker verification

PhD Thesis: X-POI-Audio



Conclusions

- We developed an audio-visual deepfake detector that is able to generalize to unseen forgeries by leveraging a contrastive learning training.

- We then focused on the audio branch, employing large pre-trained language models as feature extractors, removing the training step and all the issues that come with it.

- We then wanted to start moving towards explainable solution. We designed a feature extractor that thanks to an auto-encoder architecture is able to focus only on speaker vocal cues.

Thank you for the attention!