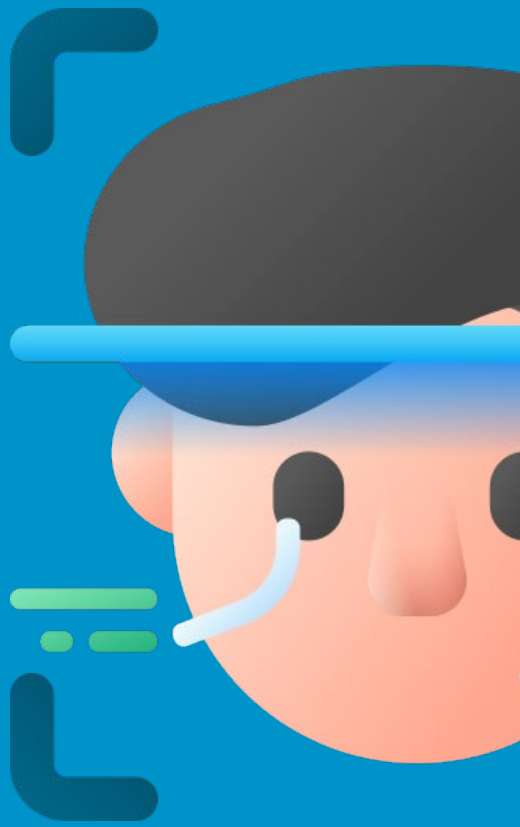
ROMCIR 2024

CMDD: A novel multimodal two-stream CNN deepfakes detector

Luca Maiano, PhD



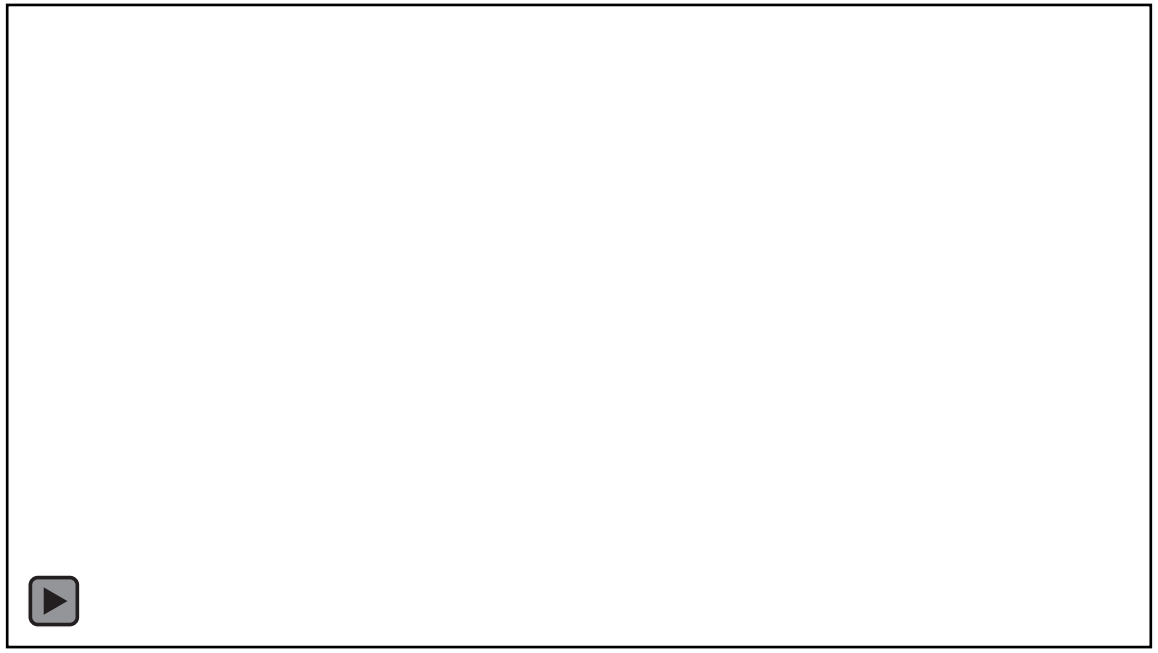
Authors: Leonardo Mongelli, Luca Maiano and Irene Amerini

Sapienza Università di Roma
ALCOR Lab
SERICS
SECURITY AND RIGHTS IN THE CYBERSPACE

Outline

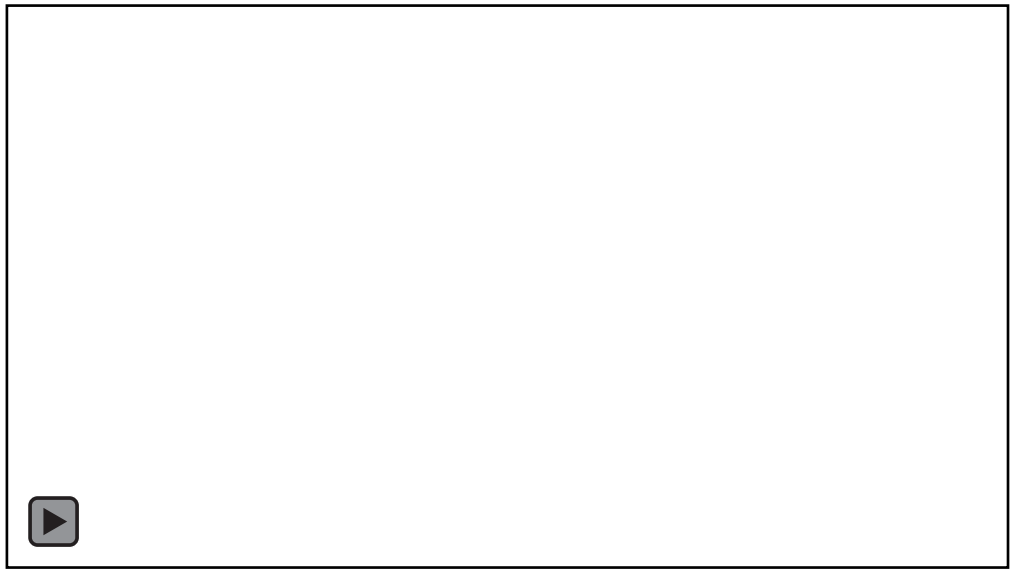
- Introduction
- Related Works
- Proposed methods
- Datasets and Metrics
- Implementation details
- Experimental results
- Conclusion and Future Works

Deepfake detection



- Generative AI allows for the generation of incredibly realistic content
- Being able to detect fake generated videos has become increasingly important (and hard!) nowadays

Detecting multimodal fake content



Visual content

Auditory content

Related works

Unimodal methods

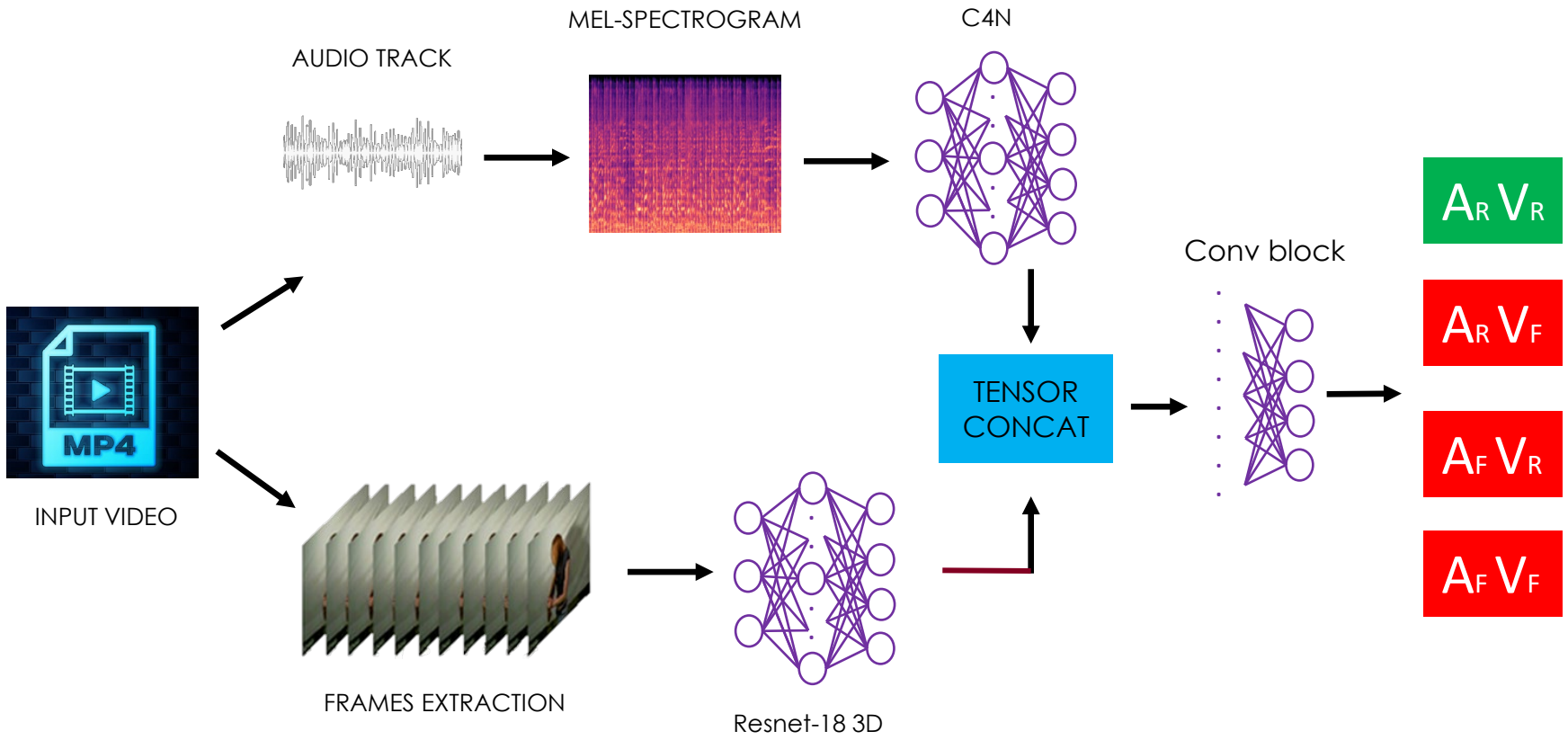
- Analyze one modality (i.e., audio or video) [7, 12]
- Limited application

Multimodal methods

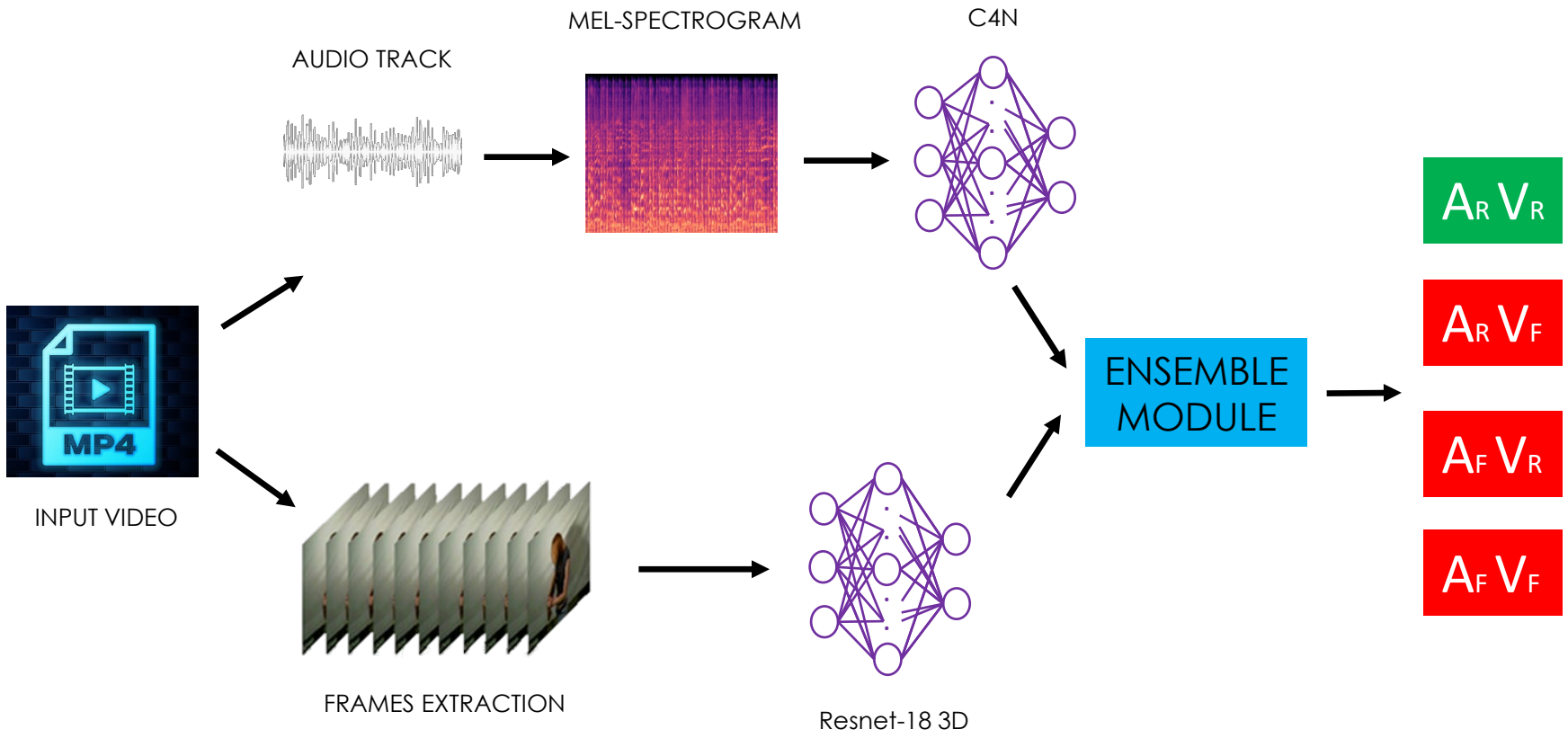
- Analyze multiple modalities at the same time (i.e., audio and video)
- Different ways to fuse modalities:
  - *Ensamble* of unimodal models [34]
  - Merging unimodal features with some *fusion mechanism* [21]

[7] Wani, et al. "*Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks*." International Conference on Image Analysis and Processing. Cham: Springer Nature Switzerland, 2023
[12] Maiano, et al. "*Depthfake: a depth-based strategy for detecting deepfake videos*." International Conference on Pattern Recognition. Cham: Springer Nature Switzerland, 2022
[21] Shahzad, et al. "*Lip sync matters: A novel multimodal forgery detector*." 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2022
[34] Hashmi, et al. "*Multimodal forgery detection using ensemble learning*." 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2022

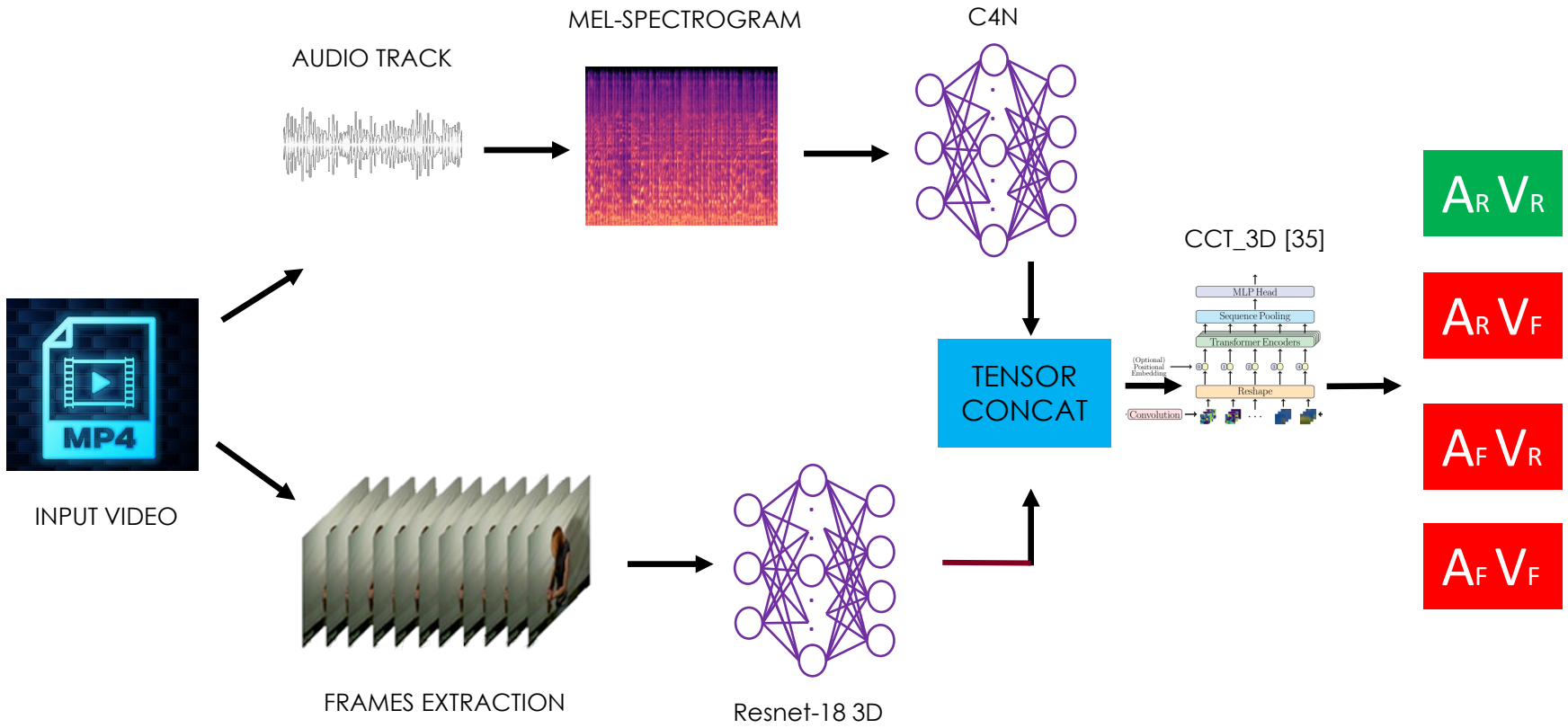
Proposed method: CMDD



Baselines (1/2): DeepMerge



Baselines (2/2): DeepFakeCVT



[35] Hassani, et al. "Escaping the big data paradigm with compact transformers." arXiv preprint arXiv:2104.05704 (2021)

Dataset FakeAVCeleb

- Multimodal dataset containing 400 real videos and 19,500 deepfake ones
- Balanced with respect to ethnicities and sex
- Various generative techniques

[24] Khalid, Hasam, et al. "FakeAVCeleb: A novel audio-video multimodal deepfake dataset." arXiv preprint arXiv:2108.05080 (2021)

Evaluation metrics

- **Accuracy**: the correct predictions over the whole predicted sample

$$\frac{TP+TN}{TP+TN+FP+FN}$$

- **Precision**: the ratio of the correct predictions over the whole correct samples

$$\frac{TP}{TP+FP}$$

- **Recall**: the ratio of correct predictions for a class to the total number of cases in which it occurs

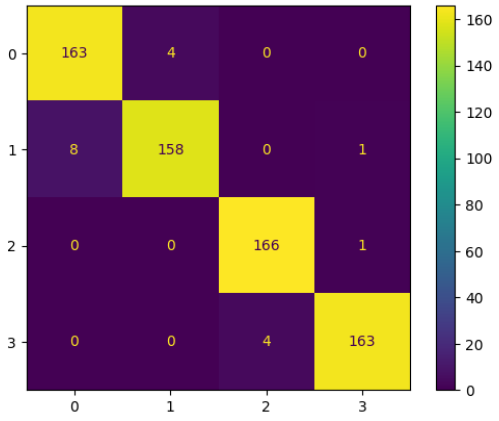
$$\frac{TP}{TP+FN}$$

- **F1-score**: the Harmonic mean between Precision and Recall

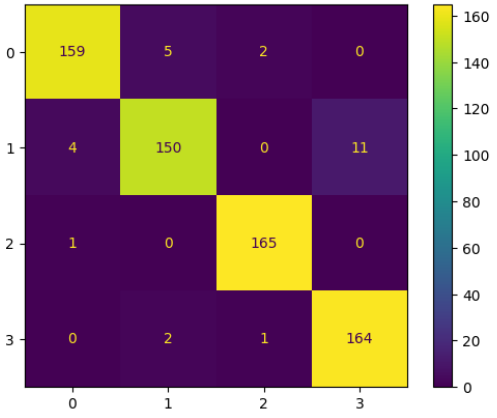
$$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Results: baseline evaluation

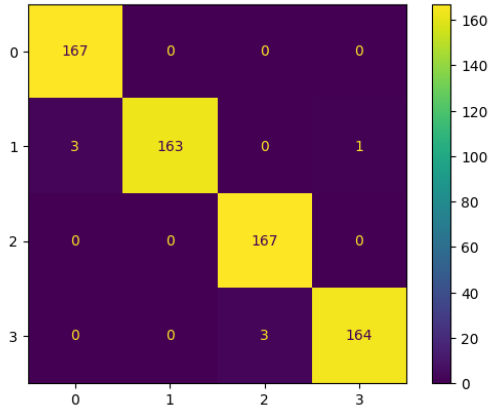
Model	Accuracy	Precision-M	Precision-m	Recall-M	Recall-m	F1-M	F1-m
DeepFakeCVT	0.9608	0.9611	0.9608	0.9607	0.9607	0.9606	0.9606
DeepMerge	0.9731	0.9732	0.9731	0.9731	0.9731	0.9730	0.9730
CMDD (proposed)	**0.9895**	**0.9897**	**0.9895**	**0.9895**	**0.9895**	**0.9895**	**0.9895**



DeepMerge



DeepFakeCVT



CMDD
(Proposed)

Results: state of the art

Position	*Model*	*Year*	*Accuracy*
1	DLC [4]	2023	0.997
2	S-Capsule Forensics [2]	2023	0.992
3	**CMDD**	**2023**	**0.989**
4	**DeepMerge**	**2023**	**0.973**
5	MIS-AVoiDD [30]	2021	0.962
6	**DeepFakeCVT**	**2023**	**0.960**
7	PVASS-MDD [31]	2023	0.957
8	AVAD [31]	2023	0.942
9	AV-Lip-Sync [21]	2022	0.940
10	AVFakeNet [32]	2023	0.934
11	Multimodaltrace [33]	2023	0.929

- Differently from other methods, our solution detects fake content using the first second of the audio only and the corresponding video frames
- The proposed solution shows a rapid tendency to overfit the training data

Results: balanced dataset

Position	*Model*	*Year*	*Accuracy*
1	**CMDD**	**2023**	**0.989**
2	**DeepMerge**	**2023**	**0.973**
3	**DeepFakeCVT**	**2023**	**0.960**
4	AV-Lip-Sync [21]	2022	0.940
5	MFD-Ensemble [34]	2022	0.790

Differences with our developed multimodal models:

- AV-Lip-Sync [21] uses a tool to generate synthetic lip sequences based on the audio track of each video and compare the generated and the real sequences
- MFD [34] is an ensamble method similar to DeepMerge

Conclusion and future works

- We propose a multimodal method that achieves state of the art performance

Future works:

- Directly give the input to the Vision Transformer by reducing the depth and the number of convolutional layers in DeepFakeCVT
- Study different fusion techniques for audio and video modalities

Contacts

Alcor Lab

WEBSITE
https://alcorlab.diag.uniroma1.it/

EMAIL
alcor@diag.uniroma1.it

Personal contacts

Luca Maiano, PhD
maiano@diag.uniroma1.it