

**Vehicle Trajectory**

**Forecasting**

Authors: Daniel Ferro, Cansu

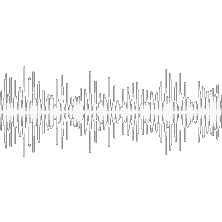
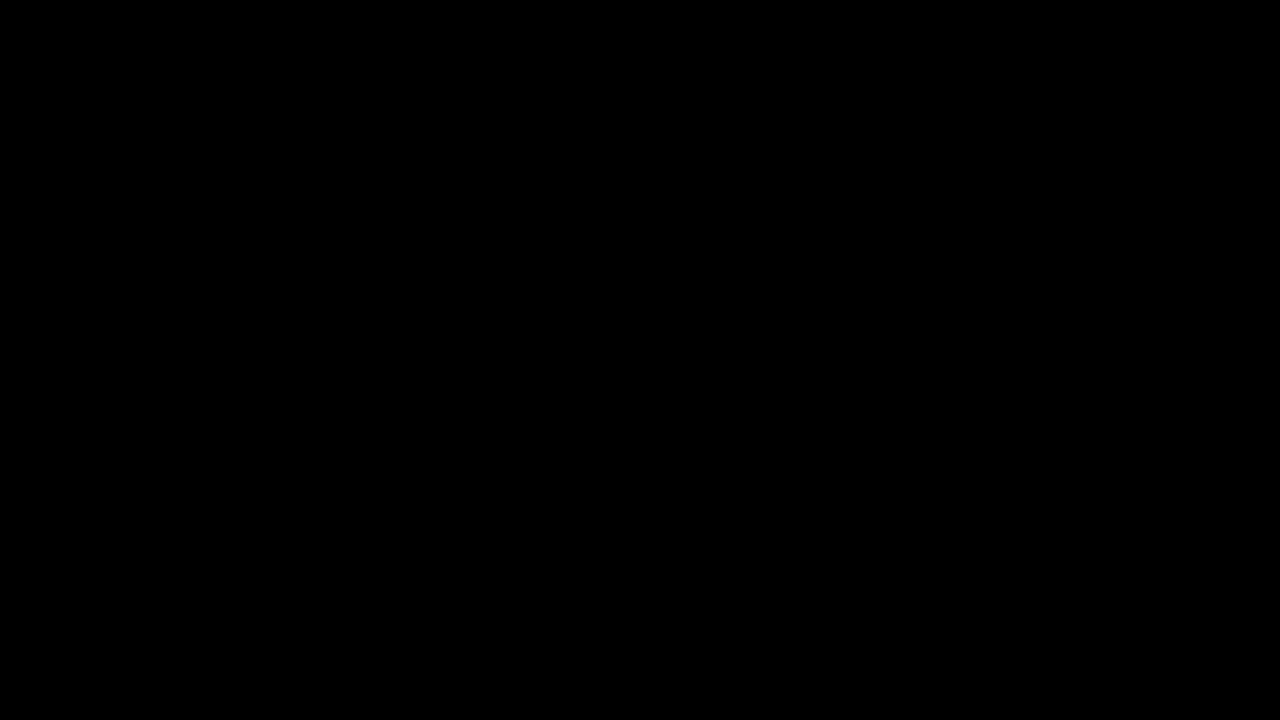
# Outline

* Introduction
* Data preparation
* Proposed methods
* Datasets and Metrics
* Implementation details
* Experimental results

# 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

Auditory content

Visual 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

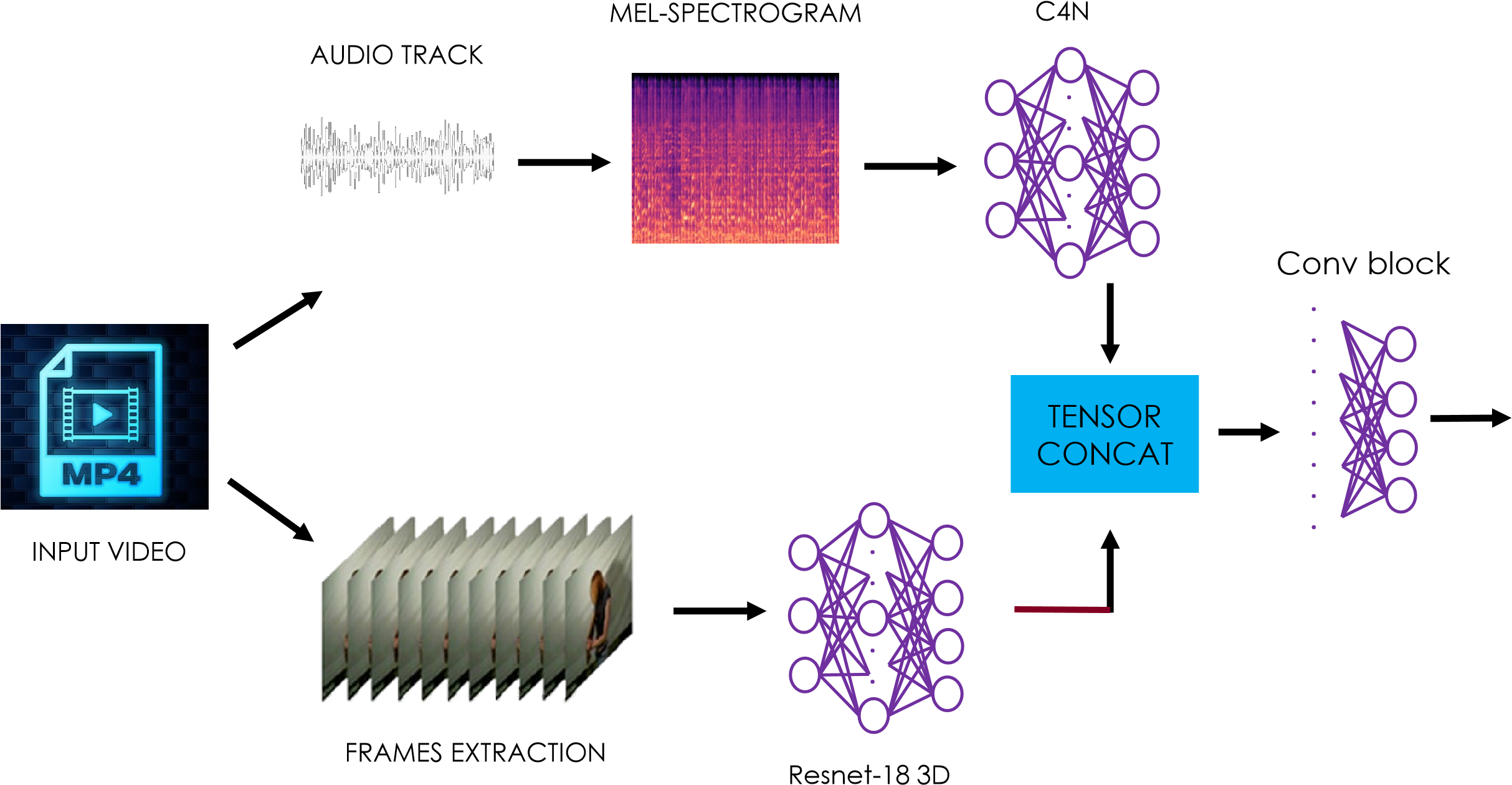
[21] Shahzad, et al. "*Lip sync matters: A novel multimodal forgery detector*." 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE,

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



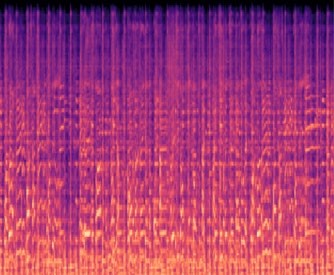
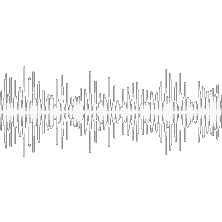
AR VR

AR VF

AF VR

AF VF

# Baselines (1/2): DeepMerge



ENSEMBLE

MODULE

A

R

V

R

A

R

V

F

A

F

V

R

A

F

V

F

**.**

**.**

**.**

**.**

**.**

**.**

**.**

**.**

INPUT VIDEO

AUDIO TRACK

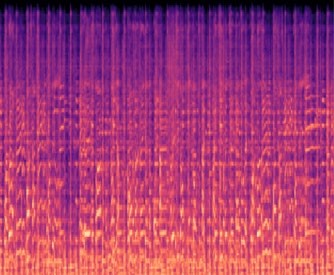
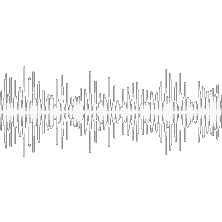
FRAMES EXTRACTION

C4N

Resnet-18 3D

MEL-SPECTROGRAM

# Baselines (2/2): DeepFakeCVT



TENSOR

CONCAT

A

R

V

R

A

R

V

F

A

F

V

R

A

F

V

F

**.**

**.**

**.**

**.**

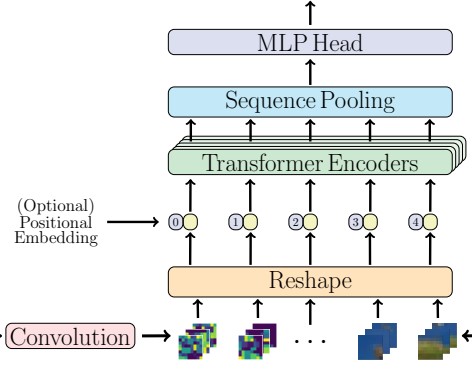
**.**

**.**

**.**

**.**

CCT\_3D [35]



INPUT VIDEO

AUDIO TRACK

FRAMES EXTRACTION

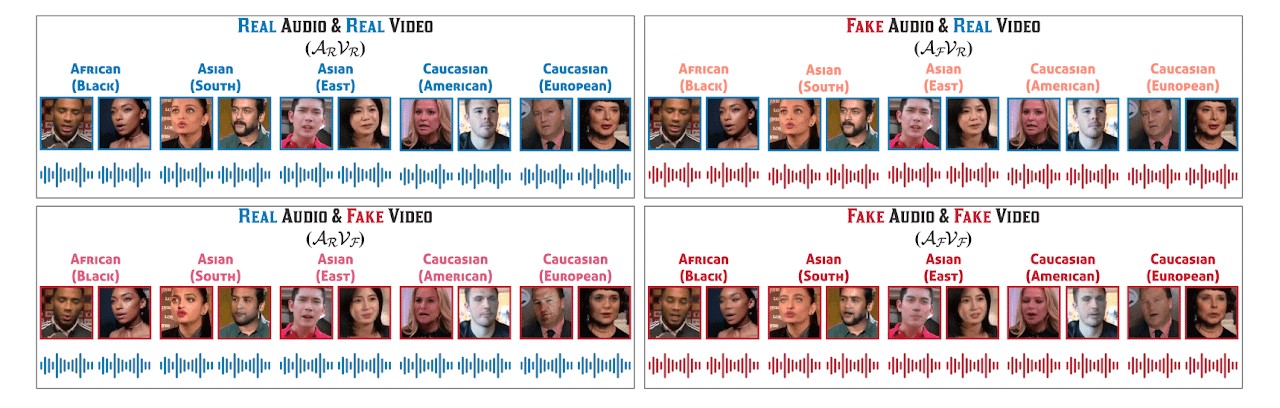
C4N

Resnet-18 3D

MEL-SPECTROGRAM

[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

𝑇𝑇𝑇𝑇+𝑇𝑇𝑇𝑇

𝑇𝑇𝑇𝑇+𝑇𝑇𝑇𝑇+𝐹𝐹𝑇𝑇+𝐹𝐹𝑇𝑇

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

𝑇𝑇𝑇𝑇

𝑇𝑇𝑇𝑇 + 𝐹𝐹𝑇𝑇

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

𝑇𝑇𝑇𝑇

𝑇𝑇𝑇𝑇 + 𝐹𝐹𝐹𝐹

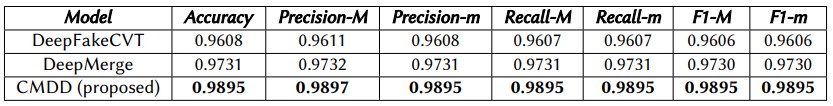
* **F1-score**: the Harmonic mean between Precision and Recall

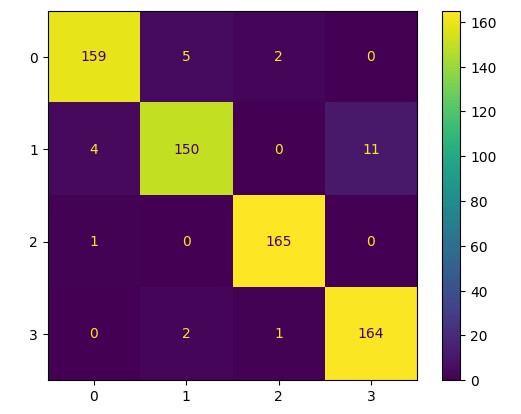
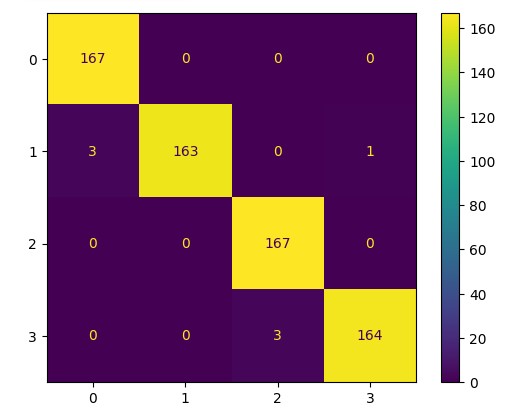
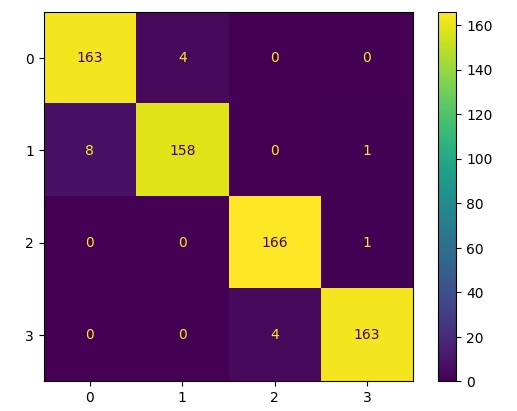
𝑇𝑇𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃 𝑅𝑅𝑃𝑃𝑃𝑃𝑅𝑅𝑅𝑅𝑅𝑅

2

𝑇𝑇𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃𝑃 + 𝑅𝑅𝑃𝑃𝑃𝑃𝑅𝑅𝑅𝑅𝑅𝑅

# Results: baseline evaluation

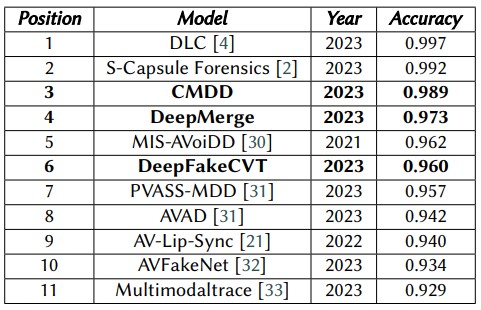




DeepMerge DeepFakeCVT CMDD

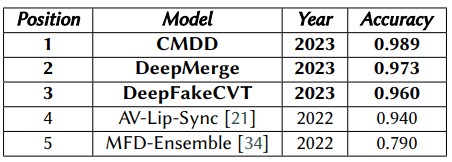
(Proposed)

# Results: state of the art



* 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



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

Multimodal Deepfake Detection

# Conclusion and future works

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

**Future works**:

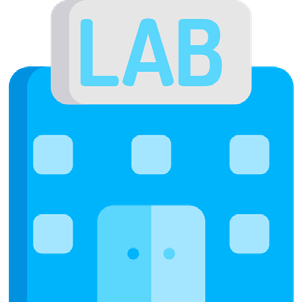
* Directly give the input to the Vision Transformer by reducing the depth and the number of convolutional layers in DeepFakeCVT
* Study different fusion techniques for audio and video modalities

**Alcor Lab**

**WEBSITE**

[https://alcorlab.diag.uniroma1.it](https://alcorlab.diag.uniroma1.it/)

[/](https://alcorlab.diag.uniroma1.it/)



**EMAIL**

alcor@diag.uniroma1.it

**Personal contacts**



**Contacts**

**Luca Maiano, PhD**

maiano@diag.uniroma1.it