06-10 March 2026

AuViRe: Audio-visual Speech Representation Reconstruction for Deepfake Temporal Localization

Christos Koutlis, **Symeon Papadopoulos**
Information Technologies Institute @ CERTH, Thessaloniki, Greece

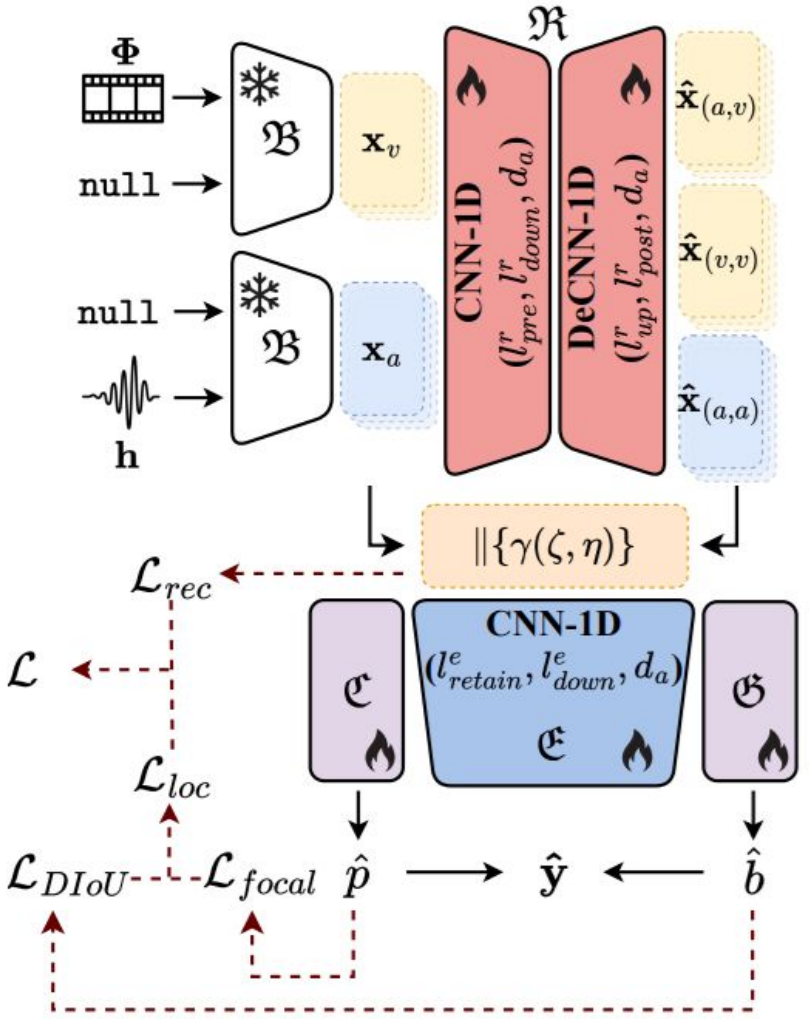
The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2026

Introduction

- **Audio-visual Deepfakes**: Hyper-realistic synthetic media where both the visual stream and the auditory stream are manipulated using Generative AI
- **Current Detection Paradigms & limitations**:
  - ***Coarse-Grained Classification***: Using video-level labels neglecting the need to identify specific forged segments.
  - ***Overfitting & Robustness***: Processing raw signals being prone to overfitting and sensitive to common content distortions.
  - ***Limited Representational Capacity***: Relying on general-purpose feature extractors lacking specialized capacity to capture subtle cross-modal inconsistencies.

Methodology

- Visual and Audio Speech Representations
- Representation Reconstruction Module
- Reconstruction-Discrepancy Encoder
- Predictions
- Objective function

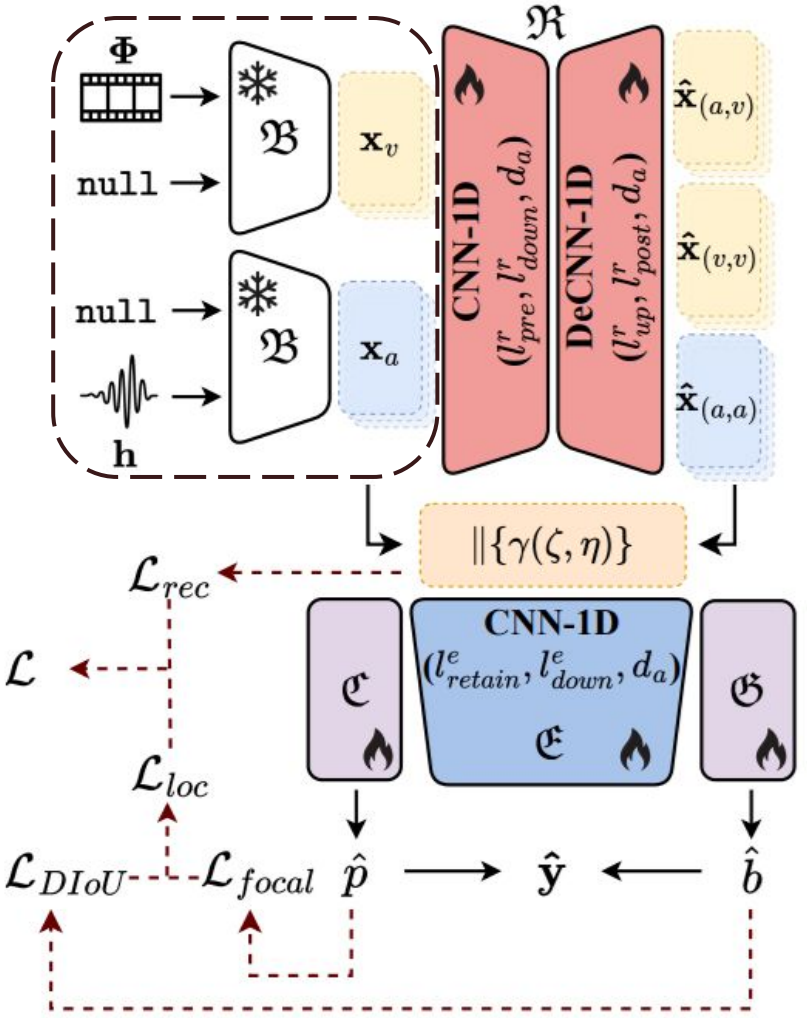


Methodology

- Visual and Audio Speech Representations

$$\mathbf{x}_v = \mathfrak{B}(\mathbf{\Phi}, \mathtt{null}), \quad \mathbf{x}_a = \mathfrak{B}(\mathtt{null}, \mathbf{h})$$

AV-HuBERT is a **self-supervised multimodal framework** that learns **speech representations** by masking and predicting clusters of audio and visual inputs. Its Transformer-based architecture captures **correlations between phonemes and visemes** providing a robust feature set for effective detection of **cross-modal inconsistencies**.

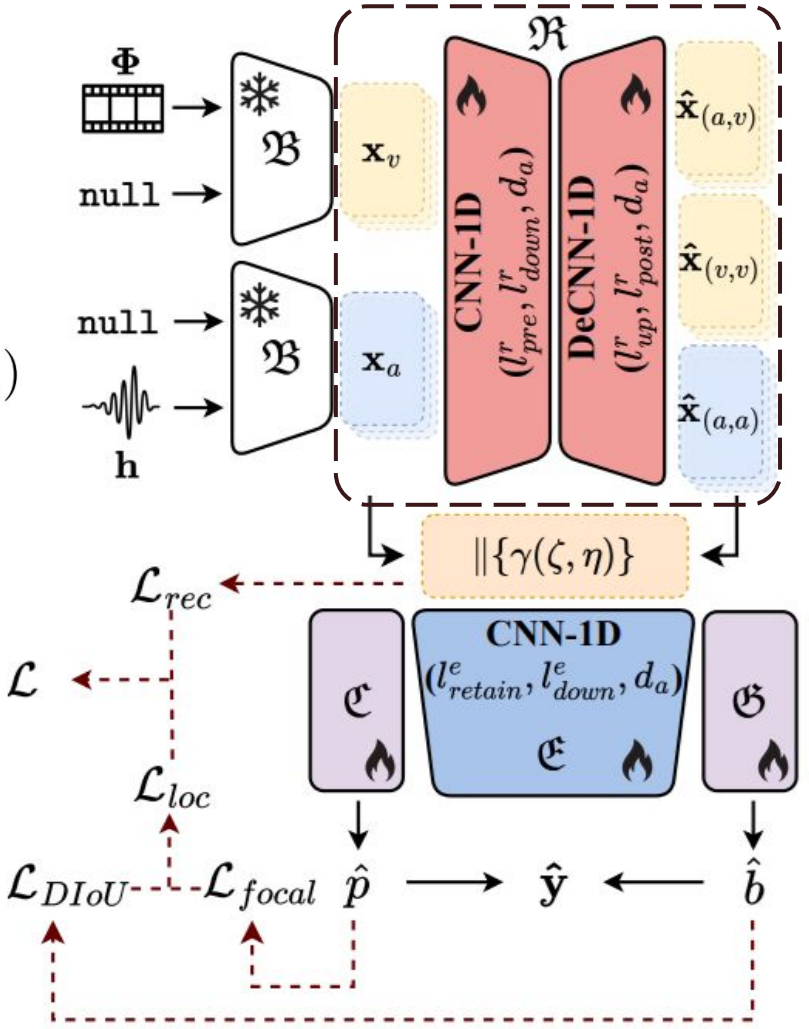


Methodology

- Representation Reconstruction Module

$$\hat{\mathbf{x}}_{(a,v)} = \mathfrak{R}(\mathbf{x}_a), \quad \hat{\mathbf{x}}_{(v,v)} = \mathfrak{R}(\mathbf{x}_v), \quad \hat{\mathbf{x}}_{(a,a)} = \mathfrak{R}(\mathbf{x}_a)$$

The **reconstruction module** predicts **visual from audio** representations (also audio from audio and visual from visual) and measures the **reconstruction discrepancy** between the predicted and ground-truth features, which significantly increases when **cross-modal artifacts** are present.



Methodology

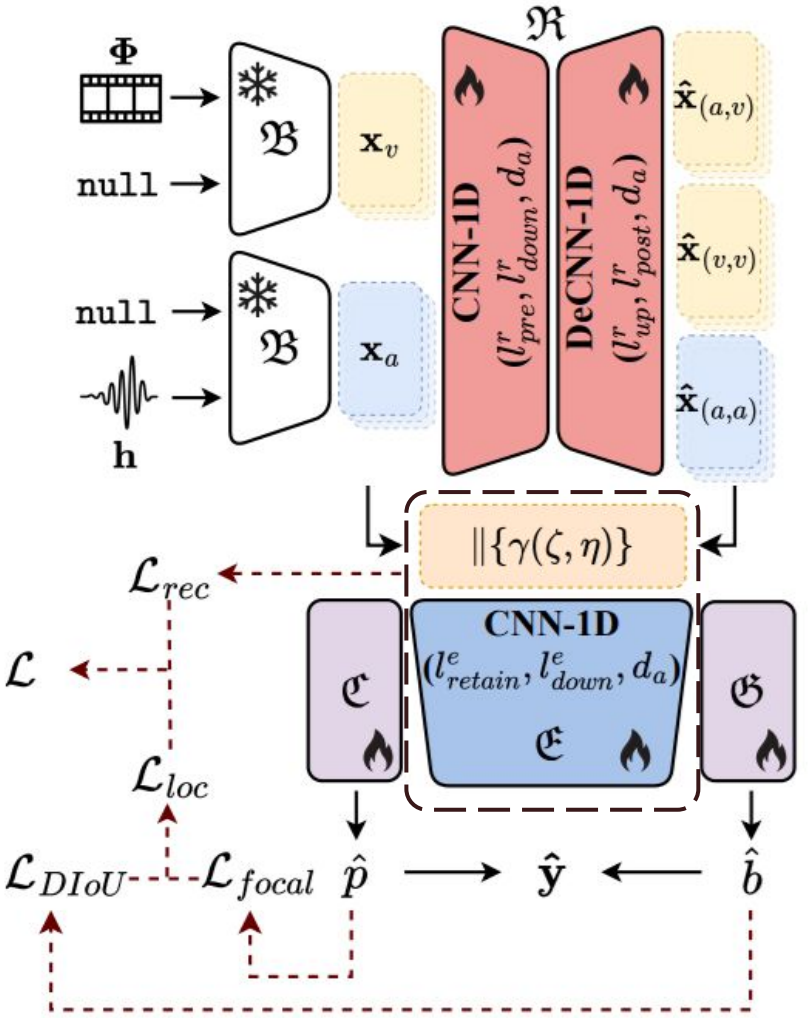
- Reconstruction-Discrepancy Encoder

$$\gamma(\zeta,\eta) = \hat{\mathbf{x}}_{(\zeta,\eta)} - \mathbf{x}_{\eta}$$

$$\mathbf{x} = \gamma(a,v) \| \gamma(v,v) \| \gamma(a,a)$$

$$\mathbf{f} = \mathfrak{E}(\mathbf{x})$$

The **encoder** processes the concatenated **residuals** between predicted and ground-truth representations, effectively **amplifying cross-modal discrepancies** to extract high-dimensional features specifically tuned for temporal forgery localization.

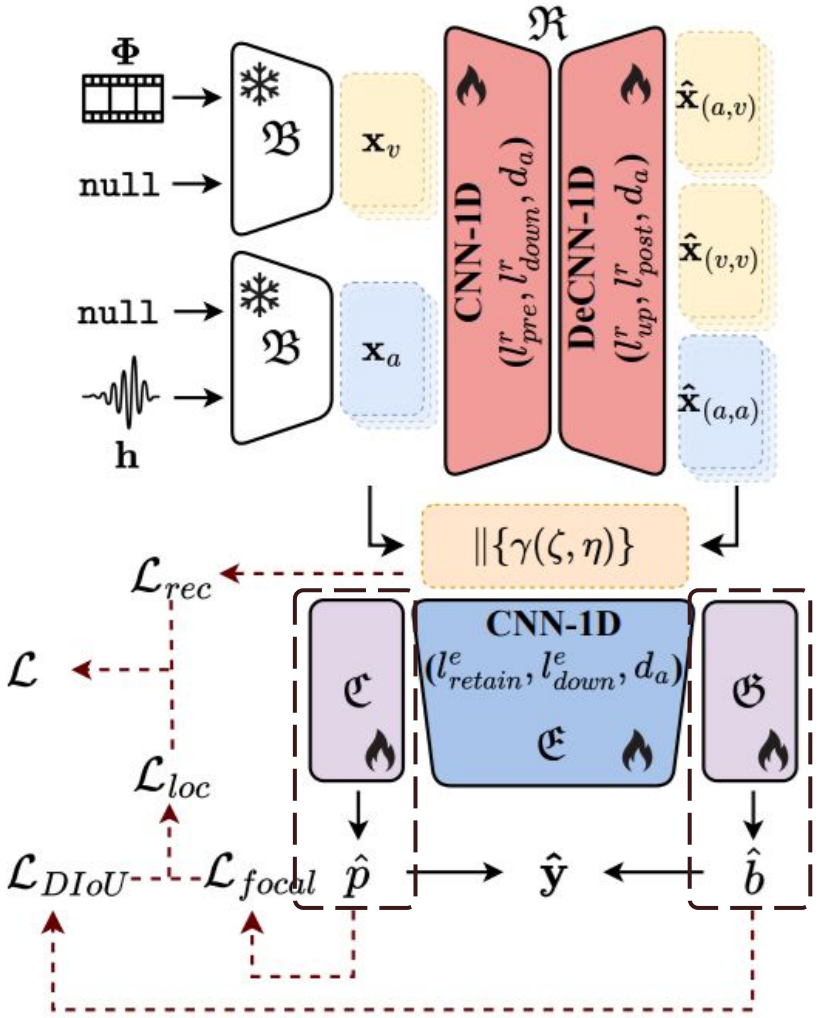


Methodology

- Predictions

$$\hat{p} = \mathfrak{C}(\mathbf{f}), \quad \hat{b} = \mathfrak{G}(\mathbf{f})$$

The **dual prediction heads** process the forgery-specific features simultaneously performing **frame-wise classification** to detect manipulations and **boundary regression** to localize the start and end points of forged segments.



Methodology

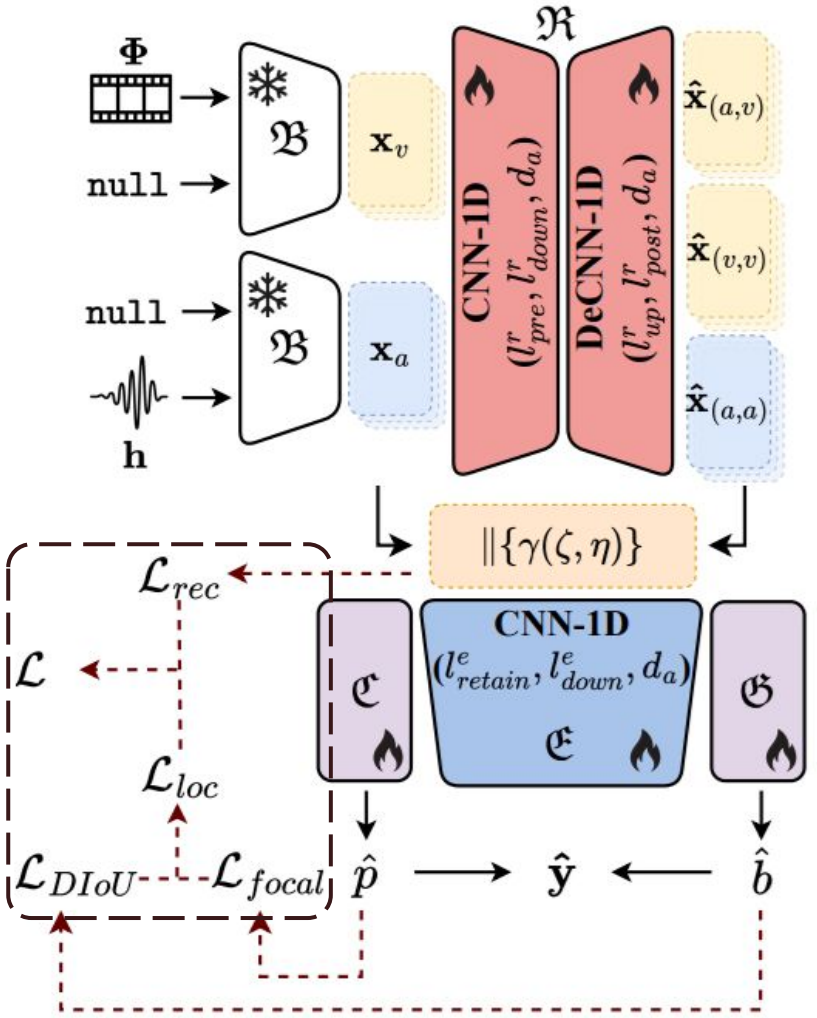
- Objective function

$$\mathcal{L}_{loc} = \frac{1}{l}\sum_{\lambda}^{l}\sum_{\tau}^{t}\Big(\mathcal{L}_{focal}(\hat{p}^{\tau,\lambda}, p^{\tau}) + p^{\tau}\cdot\mathcal{L}_{DIoU}(\hat{b}^{\tau,\lambda}, b^{\tau})\Big) \Big/ \max\Big\{\sum_{\tau}^{t} p^{\tau}, 1.0\Big\}$$

$$\mathcal{L}_{rec} = \frac{\bigwedge_{\tau=1}^{t}\neg p^{\tau}}{t\cdot d}\sum_{\tau}^{t}\sum_{\delta}^{d}\sum_{\zeta,\eta\in\{a,v\}}|\hat{\mathbf{x}}_{\zeta,\eta} - \mathbf{x}_{\eta}|$$

$$\mathcal{L} = \frac{\mathcal{L}_{loc} + \mathcal{L}_{rec}}{2}$$

The **multi-task objective** function balances **frame-level classification**, **boundary regression**, and **reconstruction**.



Experimental setup

- **Datasets**: LAV-DF, AV-Deepfake1M, Real-world Analysis (371 videos curated by fact-checkers)
- **Implementation**: Adam, lr 0.001, bs 64, epochs 100, early stop
- **Real-world experiment**: discard single-stream videos, video chunking, discard small or non-talking faces, prediction aggregation
- **Evaluation**: AP@{} / AR@{} (TFL), AUC / AP (DFD)

Results

Method	Modality	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
MDS [10]	$\mathcal{AV}$	12.8	1.6	0.0	37.9	36.7	34.4	32.2
AVFusion [3]	$\mathcal{AV}$	65.4	23.9	0.1	63.0	59.3	54.8	52.1
ActionFormer [37]	$\mathcal{V}$	95.3	90.2	23.7	88.4	89.6	90.3	90.4
BA-TFD [7]	$\mathcal{AV}$	76.9	38.5	0.3	66.9	64.1	60.8	58.4
BA-TFD+ [6]	$\mathcal{AV}$	96.3	85.0	4.4	81.6	80.5	79.4	78.8
UMMAFormer [39]	$\mathcal{AV}$	<u>98.8</u>	<u>95.5</u>	<u>37.6</u>	92.4	92.5	92.5	<u>92.1</u>
DiMoDif [20]	$\mathcal{AV}$	95.5	87.9	20.6	<u>94.2</u>	<u>93.7</u>	<u>92.7</u>	91.4
AuViRe (ours)	$\mathcal{AV}$	**98.9**	**96.0**	**46.5**	**94.9**	**94.6**	**94.0**	**93.3**

Table 1. Temporal forgery localization results on LAV-DF [7]. Modality denotes the model's input type with $\mathcal{V}$ being visual and $\mathcal{A}$ audio. **Bold** indicates best and <u>underline</u> second to best performance.

Method	Modality	AP@0.5	AP@0.75	AP@0.9	AP@0.95	AR@50	AR@30	AR@20	AR@10	AR@5
MesoInception4 [1]	$\mathcal{V}$	08.50	05.16	01.89	00.50	39.27	39.22	39.00	35.78	24.59
ActionFormer+VideoMAEv2 [33, 37]	$\mathcal{V}$	20.24	05.73	00.57	00.07	19.97	19.93	19.81	19.11	17.80
BA-TFD [7]	$\mathcal{AV}$	37.37	6.34	0.19	0.02	45.55	40.37	35.95	30.66	26.82
BA-TFD+ [6]	$\mathcal{AV}$	44.42	13.64	0.48	0.03	48.86	44.51	40.37	34.67	29.88
UMMAFormer [39]	$\mathcal{AV}$	51.64	28.07	7.65	1.58	44.07	43.93	43.45	42.09	40.27
DiMoDif [20]	$\mathcal{AV}$	<u>86.93</u>	<u>75.95</u>	<u>28.72</u>	<u>5.43</u>	<u>81.57</u>	<u>80.85</u>	<u>80.25</u>	<u>78.84</u>	<u>76.64</u>
AuViRe (ours)	$\mathcal{AV}$	**96.5**	**89.3**	**42.9**	**11.7**	**86.0**	**85.8**	**85.5**	**84.9**	**83.8**

Table 2. Temporal forgery localization results on AV-Deepfake1M [5]. Modality denotes the model's input type with $\mathcal{V}$ being visual and $\mathcal{A}$ audio. **Bold** indicates best and <u>underline</u> second to best performance. *Reports validation performance.

Results

Method	Modality	AUC
F^3-Net [28]	$\mathcal{V}$	52.0
MDS [10]	$\mathcal{AV}$	82.8
EfficientViT [11]	$\mathcal{V}$	96.5
BA-TFD [7]	$\mathcal{AV}$	99.0
UMMAFormer [39]	$\mathcal{AV}$	99.8
DiMoDif [20]	$\mathcal{AV}$	<u>99.84</u>
AuViRe (ours)	$\mathcal{AV}$	**99.94**

Table 3. Video-level deepfake detection results on LAV-DF [7]. Modality denotes the model's input type with $\mathcal{V}$ being visual and $\mathcal{A}$ audio. **Bold** indicates best and <u>underline</u> second to best performance.

Method	Modality	AUC
Video-LLaMA (13B) E5 [38]	$\mathcal{AV}$	50.7
LipForensics [17]	$\mathcal{V}$	51.6
Face X-Ray [22]	$\mathcal{V}$	61.5
MesoInception4 [1]	$\mathcal{V}$	50.1
SBI [31]	$\mathcal{V}$	65.8
MDS [10]	$\mathcal{AV}$	56.6
DiMoDif [20]	$\mathcal{AV}$	<u>96.3</u>
AuViRe (ours)	$\mathcal{AV}$	**99.8**

Table 4. Video-level deepfake detection results on AV-Deepfake1M [5]. Modality denotes the model's input type with $\mathcal{V}$ being visual and $\mathcal{A}$ audio. **Bold** indicates best and <u>underline</u> second to best performance.

Results

method	AUC	AP
DiMoDif [20]	61.6	55.7
RealForensics [16]	70.4	59.9
AuViRe (Ψ_m)	**75.5**	**67.2**
AuViRe (Ψ_s)	<u>71.0</u>	<u>62.3</u>

Table 5. Real-world performance.

		Ψ_m		Ψ_s	
Language	Videos (#)	AUC	AP	AUC	AP
English	282	77.9	64.4	73.6	60.0
Welsh	42	86.2	74.4	71.4	56.2
French	24	60.1	68.3	54.5	59.7
Greek	17	90.9	73.5	81.8	74.6
Spanish	16	86.7	94.2	88.3	94.8

Table 6. AuViRe's real-world performance per language.

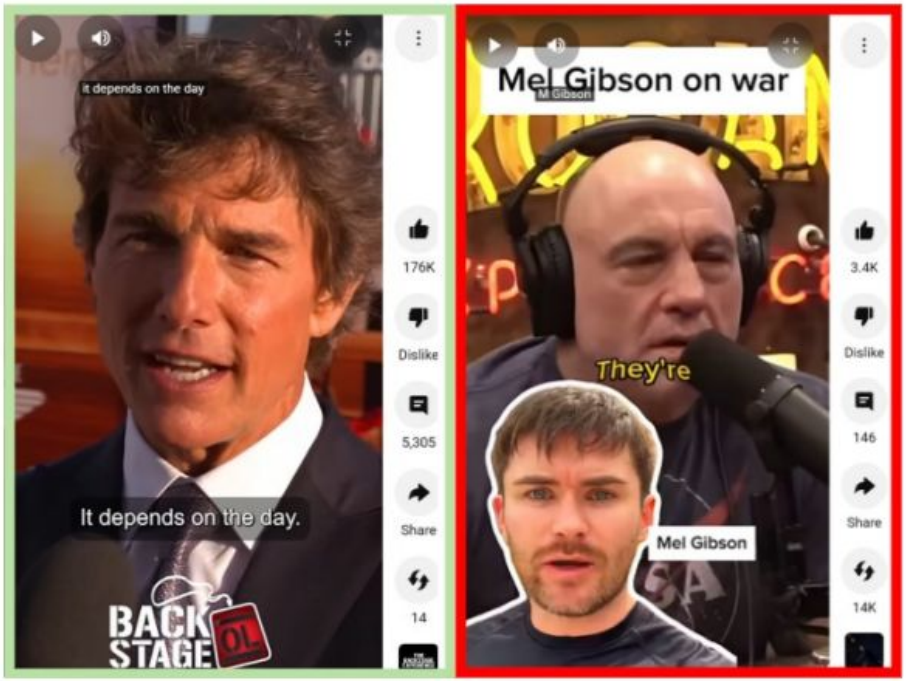


Figure 2. Examples of correctly (left; https://www.youtube.com/shorts/xAPIjhkXF-0) and erroneously (right; https://www.youtube.com/shorts/D9mQGzG9dRk) classified **real** samples.

Results

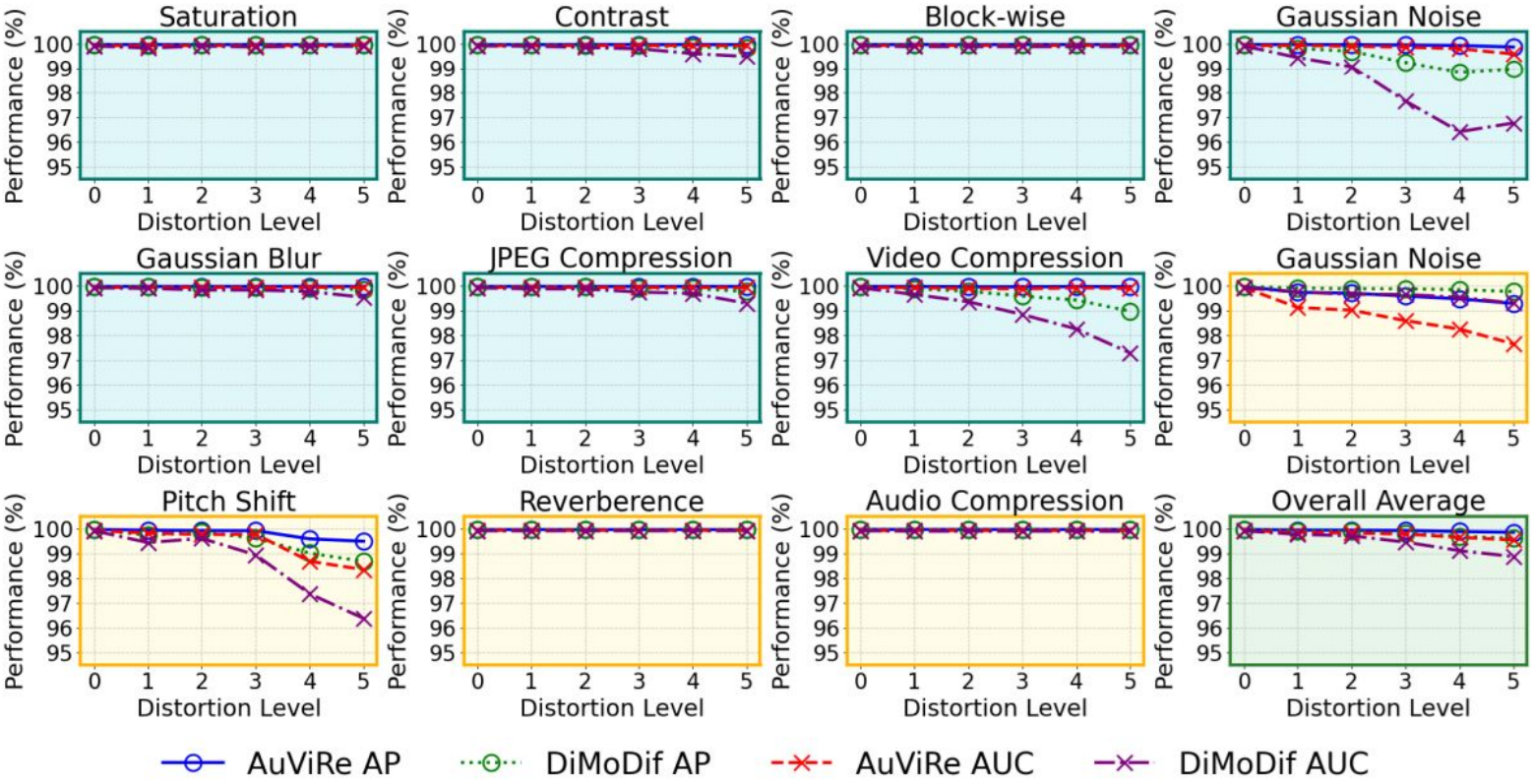


Figure 3. Robustness analysis using visual (cyan plots) and audio (yellow plots) distortions. Overall average robustness is also reported.

Computational efficiency

- Average `processing-time/video-duration` ratio 0.58 (std 0.25)
- 12.1M learnable parameters
- 1.0 GFLOP for a forward pass
- Processes ~43 FPS including feature extraction and simultaneous audio processing

Thank you for your attention!

Pre-print:
https://arxiv.org/abs/2511.18993

GitHub:
https://github.com/mever-team/auvire

Contact
Christos Koutlis
Information Technologies Institute @ CERTH
Email: ckoutlis@iti.gr

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

Information Technologies Institute

mever

WACV 2026
TUCSON, AZ
3/6 - 3/10