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

Christos Koutlis and Symeon Papadopoulos
Information Technologies Institute @ CERTH
Thessaloniki, Greece
{ckoutlis,papadop}@iti.gr

A. Generalization

Tabs. 1 and 2 show cross-dataset generalization performance for DFD and TFL. Trained on AV-Deepfake1M, a set ten times larger with higher quality generated content, the model achieves strong performance on LAV-DF, with 93.33 AUC (+7.03 compared to DiMoDif) and 43.3 AP@0.75 (+17.1 compared to DiMoDif). Trained on LAV-DF AuViRe performs on par to DiMoDif (-1.51 AUC, +2.1 AP@0.75). The failure of both models at 90/95% IoU is attributed to the x2 larger forged segments of LAV-DF vs. AV-Deepfake1M.

B. Hyperparameter tuning

Fig. 1 presents our hyperparameter tuning experimental results. These are obtained by training and evaluating AuViRe on LAV-DF on a grid of 36 experiments defined by $d_a \in \{32, 64, 128, 256\}$, $l^r_{down} = l^r_{up} \in \{1, 2, 3\}$, and $l^e_{retain} = l^e_{down} \in \{1, 2, 3\}$. High AP is achieved with $d_a \in \{32, 64\}$ but with lower AR scores. $d_a \in \{128, 256\}$ strike a balance between AP and AR with the best configuration obtaining average rank 6.7 defined by $d_a = 128$, $l^r_{down} = l_{up} = 3$, $l^e_{retain} = l^e_{down} = 2$. The same analysis on AV-Deepfake1M (validation set) is illustrated in Fig. 2.

Table 1. Generalization performance on Deepfake Detection task.

		Tested on			
		LAV-DF		AVD1M	
Trained on	Model	AP	AUC	AP	AUC
AVD1M	DiMoDif	94.98	86.30	-	96.30
	AuViRe	**97.66**	**93.33**	-	**99.78**
LAV-DF	DiMoDif	99.94	99.84	-	**67.21**
	AuViRe	**99.98**	**99.94**	-	65.70

C. Robustness

Tab. 3 reports average backbone $\mathfrak{B}$ feature similarity among original and distorted versions of the same LAV-DF samples, showcasing its high resilience that contributes to AuViRe's robustness.

D. Threshold Calibration

Fig. 3, presents a systematic calibration analysis using the real-world data indicating 0.01 as Ψ_m's optimal probability threshold θ.

E. Real-world performance drop: Examples

Figure 4 provides further qualitative examples pinpointing the most prominent contributing factors to the performance drop from a controlled to the real-world uncontrolled environment.

Table 2. Generalization performance on Temporal Forgery Localization task.

Trained on	Tested on	Method	AP				AR					
			@0.5	@0.75	@0.9	@0.95	@100	@50	@30	@20	@10	@5
AVD1M	LAV-DF	DiMoDif	**59.8**	26.2	2.3	0.2	75.3	72.3	69.7	67.3	62.5	56.2
		AuViRe	53.8	**43.3**	**13.9**	**0.9**	**78.0**	**76.5**	**75.2**	**74.1**	**71.4**	**66.8**
LAV-DF	AVD1M	DiMoDif	**15.9**	4.3	0.3	0.02	-	28.9	**27.3**	**25.7**	**22.5**	**19.1**
		AuViRe	14.7	**6.4**	**0.6**	**0.05**	-	**29.3**	26.6	24.6	21.3	18.2

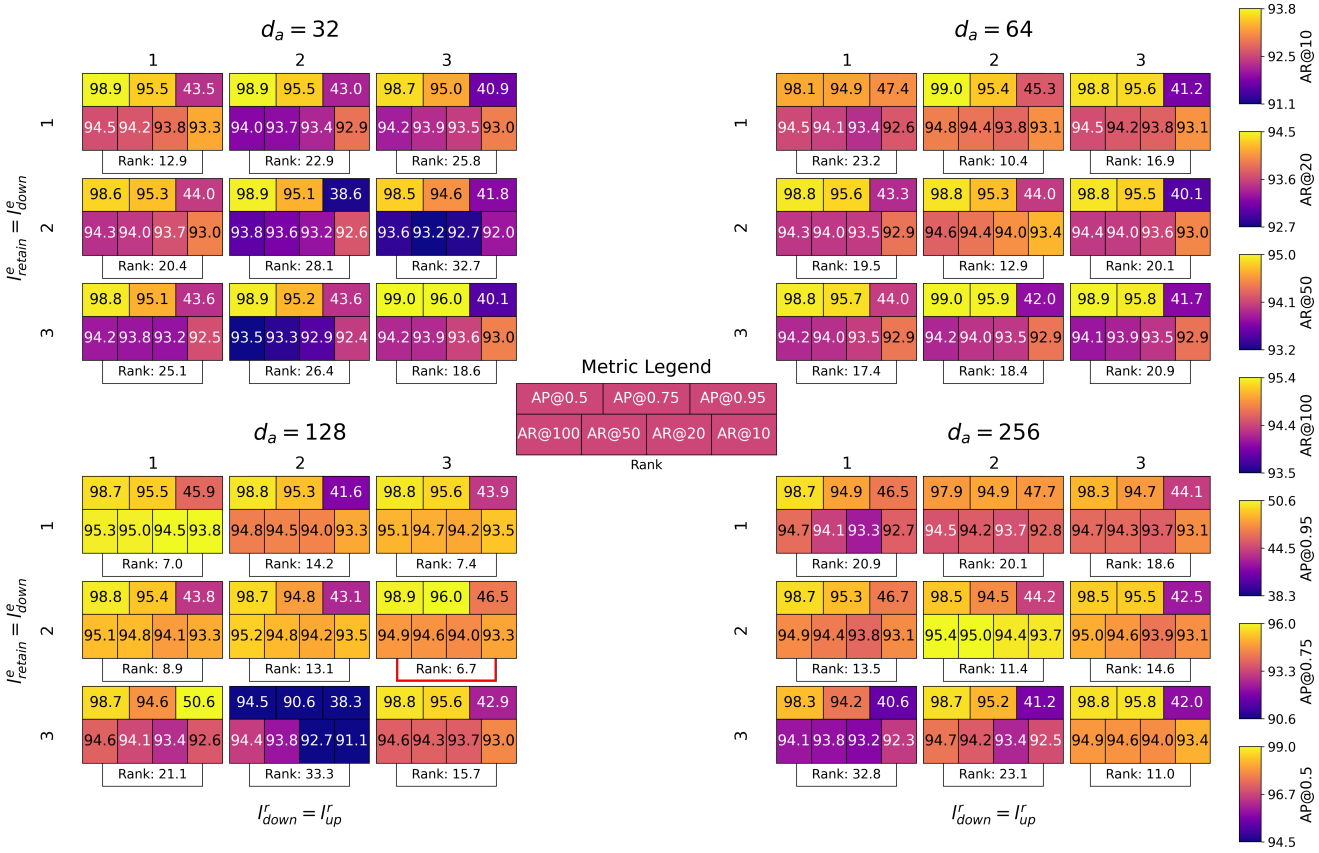

Figure 1. Hyperparameter grid results on LAV-DF. Best rank is highlighted with red.

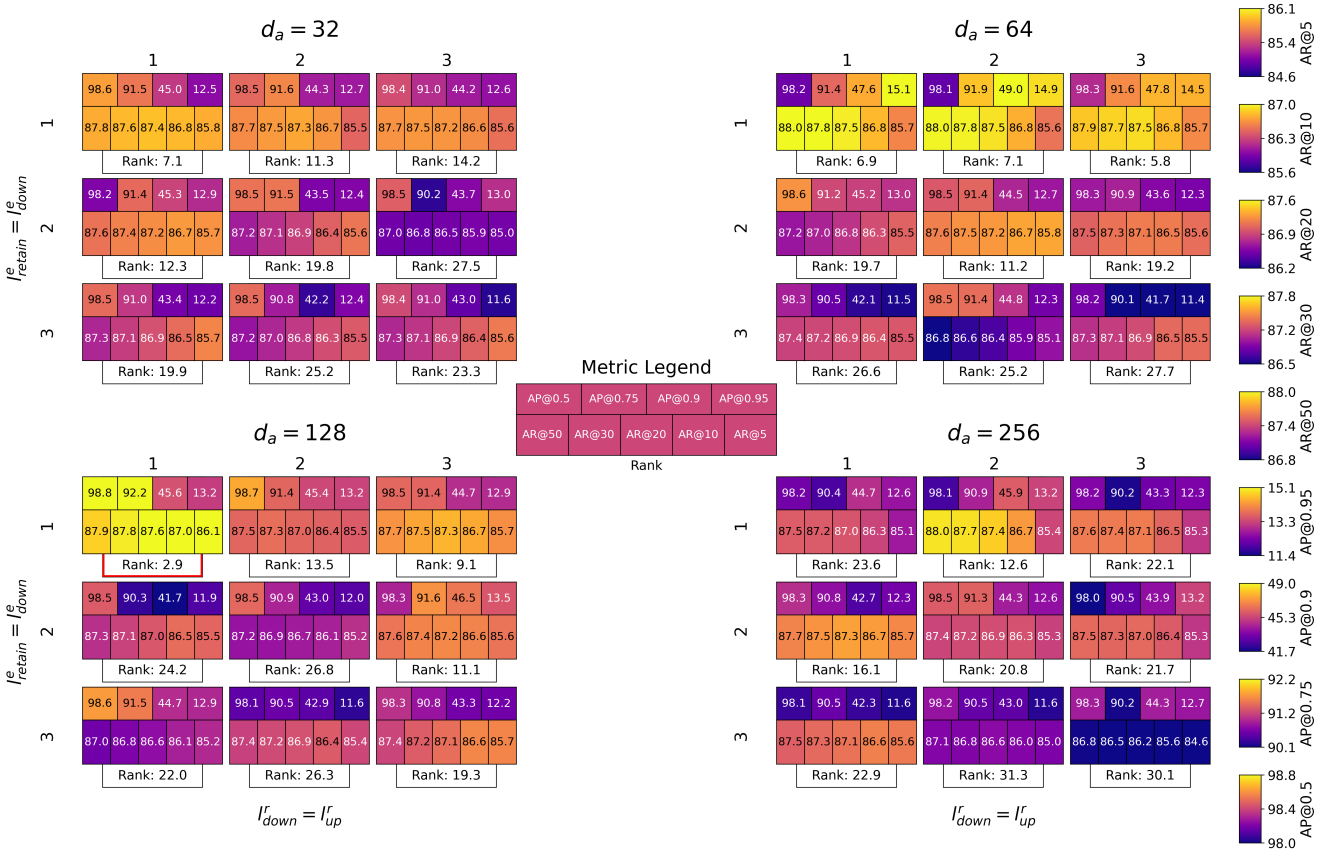

Figure 2. Hyperparameter grid results on AV-Deepfake1M. Best rank is highlighted with red.

Table 3. Average cosine similarity of $\mathfrak{B}$ extracted features, between original and distorted versions of 1,000 LAV-DF unseen samples.

Modality	Distortion	Level 1	Level 2	Level 3	Level 4	Level 5
Visual	Color Saturation	0.96 ± 0.05	0.96 ± 0.05	0.97 ± 0.05	0.96 ± 0.05	0.97 ± 0.05
	Color Contrast	0.95 ± 0.05	0.93 ± 0.06	0.89 ± 0.08	0.83 ± 0.10	0.73 ± 0.10
	Block-wise	0.96 ± 0.04	0.96 ± 0.04	0.94 ± 0.06	0.92 ± 0.08	0.88 ± 0.11
	Gaussian Noise	0.87 ± 0.08	0.79 ± 0.10	0.64 ± 0.13	0.48 ± 0.16	0.09 ± 0.10
	Gaussian Blur	0.94 ± 0.06	0.92 ± 0.08	0.86 ± 0.10	0.79 ± 0.11	0.71 ± 0.13
	JPEG	0.96 ± 0.05	0.94 ± 0.07	0.90 ± 0.08	0.84 ± 0.09	0.75 ± 0.10
	Video Compression	0.89 ± 0.06	0.85 ± 0.06	0.77 ± 0.09	0.67 ± 0.09	0.59 ± 0.10
Audio	Gaussian Noise	0.81 ± 0.07	0.73 ± 0.09	0.66 ± 0.10	0.58 ± 0.11	0.49 ± 0.11
	Pitch Shift	0.88 ± 0.04	0.67 ± 0.18	0.49 ± 0.25	0.35 ± 0.22	0.26 ± 0.20
	Reverberence	1.00 ± 0.03	1.00 ± 0.03	1.00 ± 0.04	0.99 ± 0.04	0.99 ± 0.06
	Audio Compression	0.88 ± 0.17	0.60 ± 0.10	0.51 ± 0.11	0.56 ± 0.19	0.99 ± 0.06

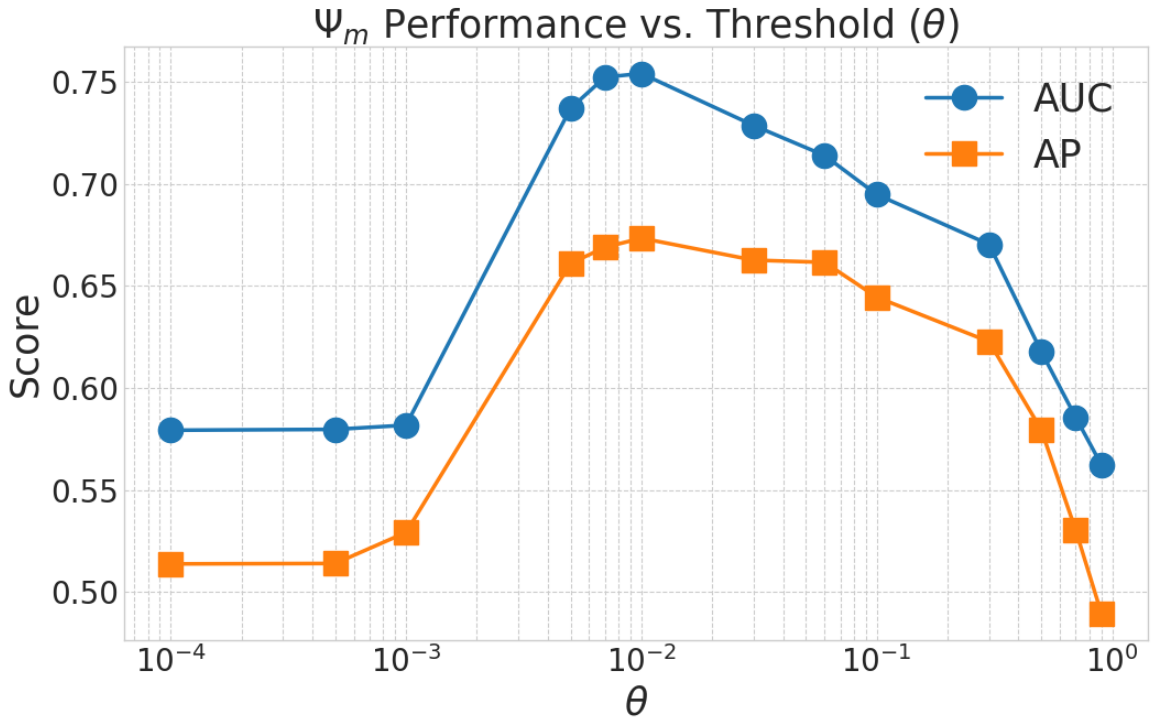

Figure 3. Calibration of Ψ_m's θ parameter.

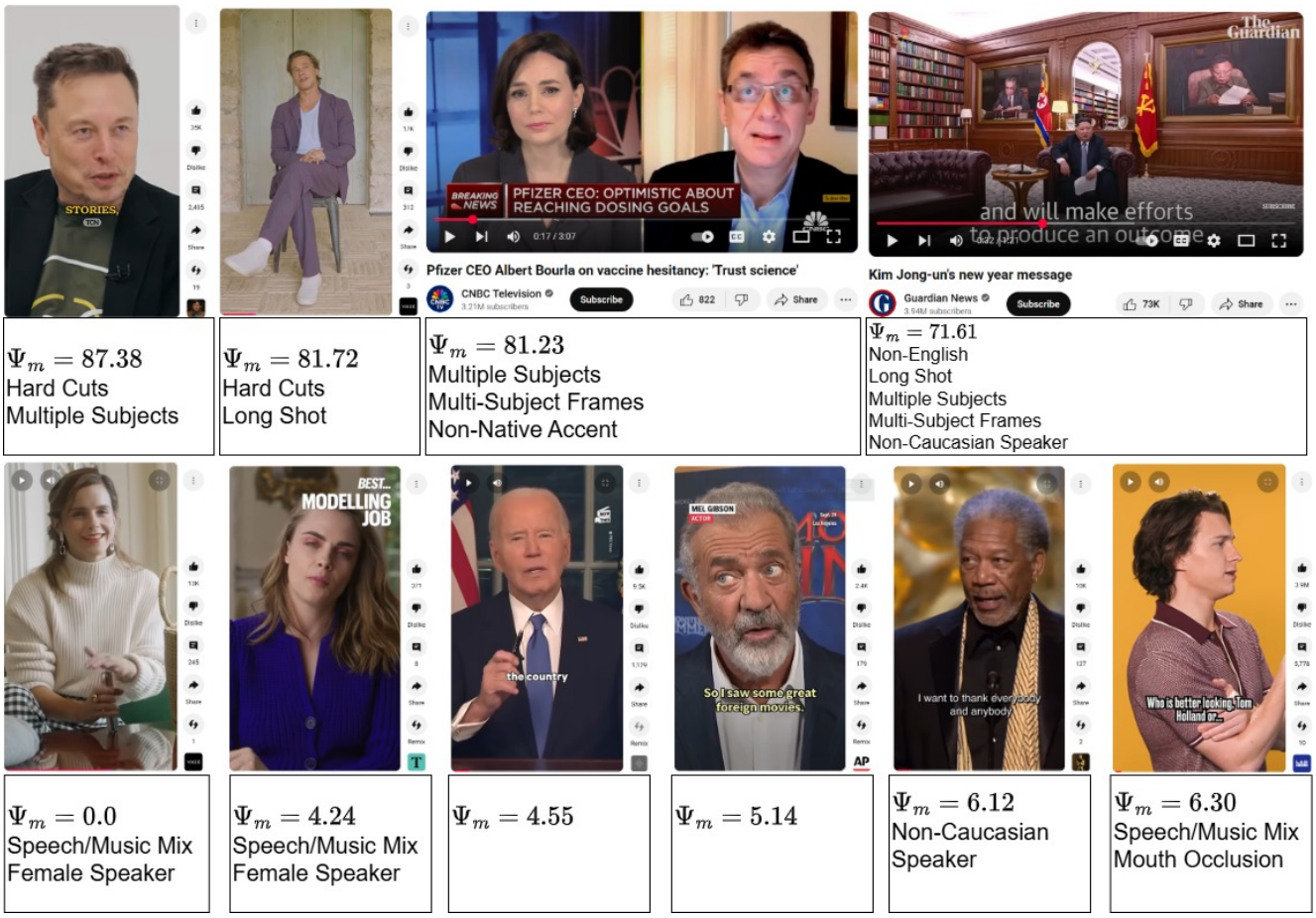

Figure 4. Examples of erroneously (upper panel; 1: https://www.youtube.com/shorts/c15HhmtfI5w, 2: https://www.youtube.com/shorts/cqxkYUfOf9c, 3: https://www.youtube.com/watch?v=P6O6FqYYImk&ab_channel=CNBCTelevision, 4: https://www.youtube.com/watch?v=YLWqqDVWsXo) and correctly (bottom panel; 1: https://www.youtube.com/shorts/ciSd7M3lAco, 2: https://www.youtube.com/shorts/kyf_eCCc5qU, 3: https://www.youtube.com/shorts/EVQo--qKQAI, 4: https://www.youtube.com/shorts/dvfN3WD1_A8, 5: https://www.youtube.com/shorts/JW4FT8RVJ88, 6: https://www.youtube.com/shorts/gH96b2U293Y) classified **real** samples.