Personalized Audio-to-Expression Diffusion Models are Robust Zero-Shot Face Forgery Detectors

Kaede Shiohara[1] Toshihiko Yamasaki[1] Vladislav Golyanik[2]

[1]The University of Tokyo [2]Max Planck Institute for Informatics

Contact & Project Page

Kaede Shiohara

Project Assistant Prof. at the University of Tokyo

shiohara@cvm.t.u-tokyo.ac.jp

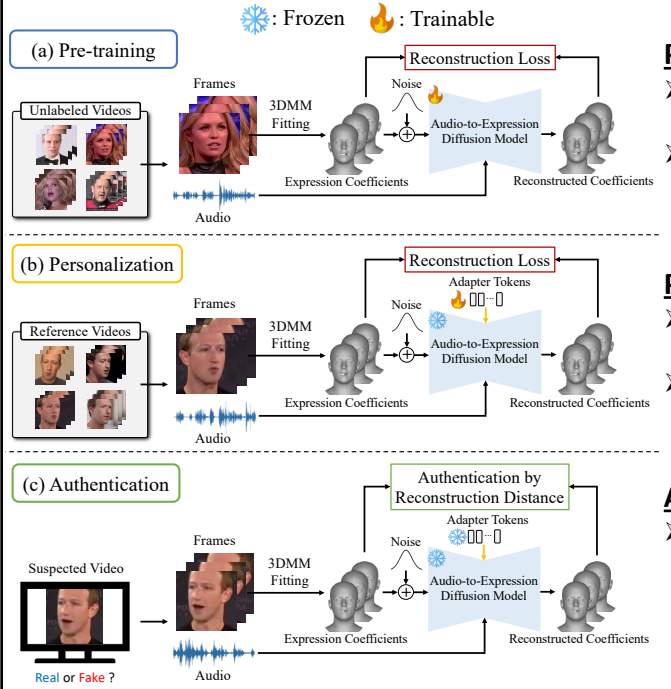
1. Challenge: Generalization to Unseen Manipulations

Supervised Methods: **easy** to train but **not generalized**

Self-supervised Methods: **ideally generalized** but **difficult to train**

ExposeAnyone (Ours): **self-supervised**, **generalized**, and **robust**

2. How? Use an Audio-to-Expression Diffusion Model



Pre-training

- Construct audio-expression datasets e.g., VoxCeleb2
- Use FLAME expression (only 53 dim)

Personalization

- Fine-tune the model on person-of-interest videos (e.g., Zuckerberg)
- Only adapter tokens are optimized

Authentication

- Reconstruct suspected videos
  If rec. error is large → Fake
  small → Real

because the model only learned to reconstruct expressions of the real person

3. Important Techniques

3DMM Refinement

- SPECTRE (a feed-forward per-frame FLAME predictor) generates **entangled** expression and shape

Thus, we disentangle expression from shape by
1. assigning **a single shape** to each video
2. optimizing the shape and expression params

Method	Test Set AUC (%) on DFDCP
SPECTRE	46.82
SPECTRE + Refinement	93.45

Content-agnostic Authentication

- Direct use of the reconstruction error **does not work** due to the variation of rec errors **depending on contents**

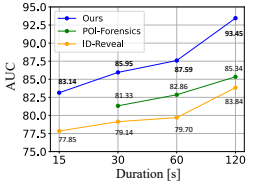
Thus, we focus on the difference between two rec errors:

$$\text{Fakeness} = \frac{\text{Rec. Error } \textbf{w/ Adapter}}{\text{Rec. Error } \textbf{w/o Adapter}}$$

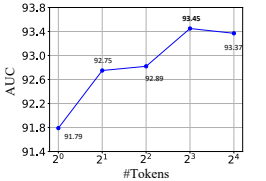
Method	Test Set AUC (%) on DFDCP
Vanilla Reconstruction Error	61.34
Content-agnostic Authentication	93.45

4. Analysis

Duration of reference videos



Number of adapter tokens



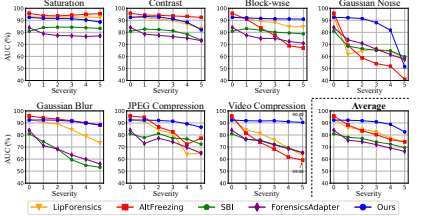
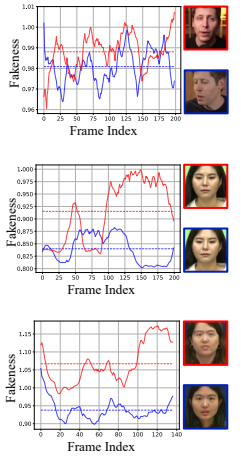
5. Generalization Ability, Robustness, and Fakeness Visualizations

Learning Type	Method	Reference	Test Set AUC (%) DF-TIMIT	DFDCP	KoDF	IDForge	Avg
Supervised	LipForensics	-	96.74	69.89	95.98	93.05	88.92
	AltFreezing	-	**99.82**	69.33	96.41	**95.79**	90.34
	DFD-FCG	-	99.02	90.04	**97.40**	77.56	91.00
Pseudo-Supervised	Face X-ray	-	76.01	67.77	48.84	54.76	61.35
	SBI	-	84.71	88.51	87.84	80.94	85.50
Self-Supervised	ID-Reveal	✓	60.41	83.84	61.64	74.77	70.17
	POI-Forensics	✓	85.65	85.34	64.91	72.60	77.13
	SpeechForensics	-	71.35	63.61	82.98	93.66	77.90
	Ours (VoxCeleb2)	✓	99.49	93.06	87.73	93.25	93.38
	Ours	✓	99.72	**93.45**	95.31	92.40	**95.22**

Deepfake detection across 4 datasets

Method	Test Set AUC (%) on Each Subject @ijustine	@mcuban	@sama	Avg
DFD-FCG	36.11	53.47	56.25	48.61
SBI	45.14	45.83	50.69	47.22
SpeechForensics	54.86	64.58	63.89	61.11
Ours	**98.61**	**84.72**	**100**	**94.44**
DIRE	11.11	56.25	43.75	37.04
B-Free	65.97	71.53	76.39	71.30

Sora2-generated video detection



Robustness to corruptions