Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning

Ștefan Smeu[1], Dragoș-Alexandru Boldișor[1], Dan Oneață[2], Elisabeta Oneață[1],




[1]Bitdefender, Romania [2]Politehnica, Bucharest
bit-ml.github.io

Motivation

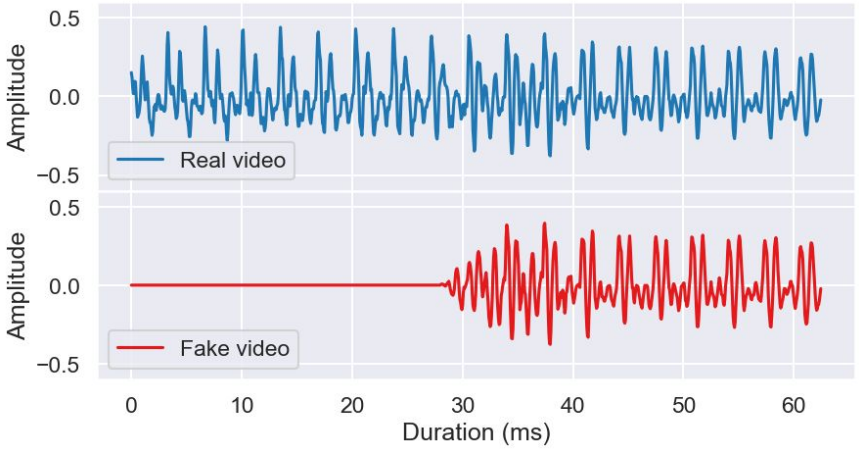
- Deepfake detection models often **exploit unintended artifacts and biases** in datasets.
  - audio or visual artifacts
  - audio-video desynchronisation
  - post-processing artifacts
  - others
- Such **reliance on spurious features undermines model robustness and generalizability**.

Findings

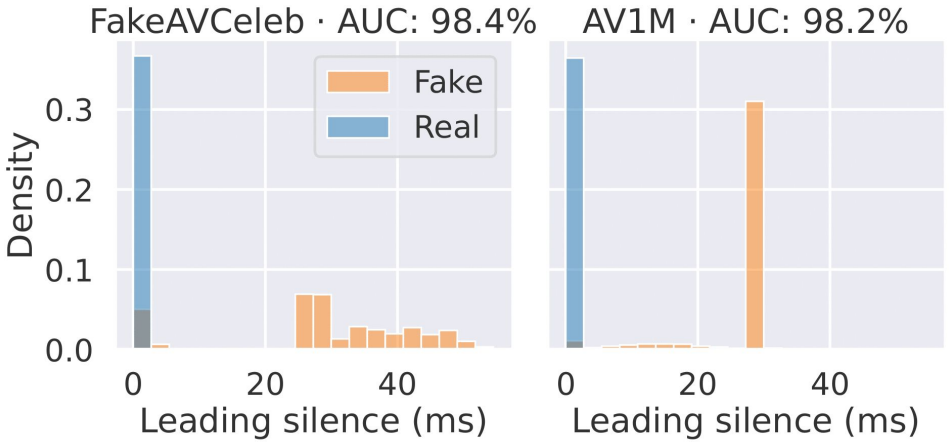
- **Leading silence in fake videos** is **a spurious shortcut** in audio-visual deepfake datasets.
- **Supervised models overfit to this**, failing to generalize when the silence is removed.

Found biases in audio-deepfake datasets

Leading silence:

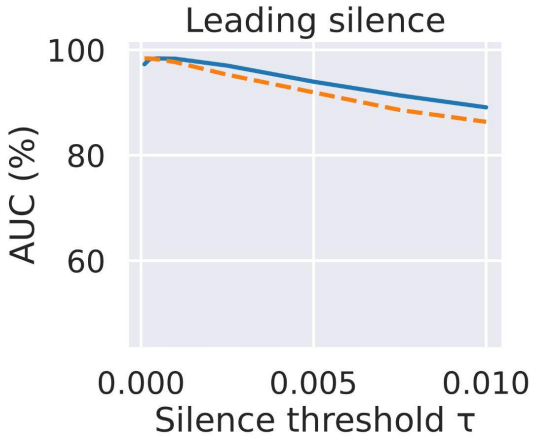


Beginning of a real video and its fake counterpart, before alteration begins.



Leading silence distribution in FakeAVCeleb and AV-Deepfake1M

Performant baseline: leading silence classifier



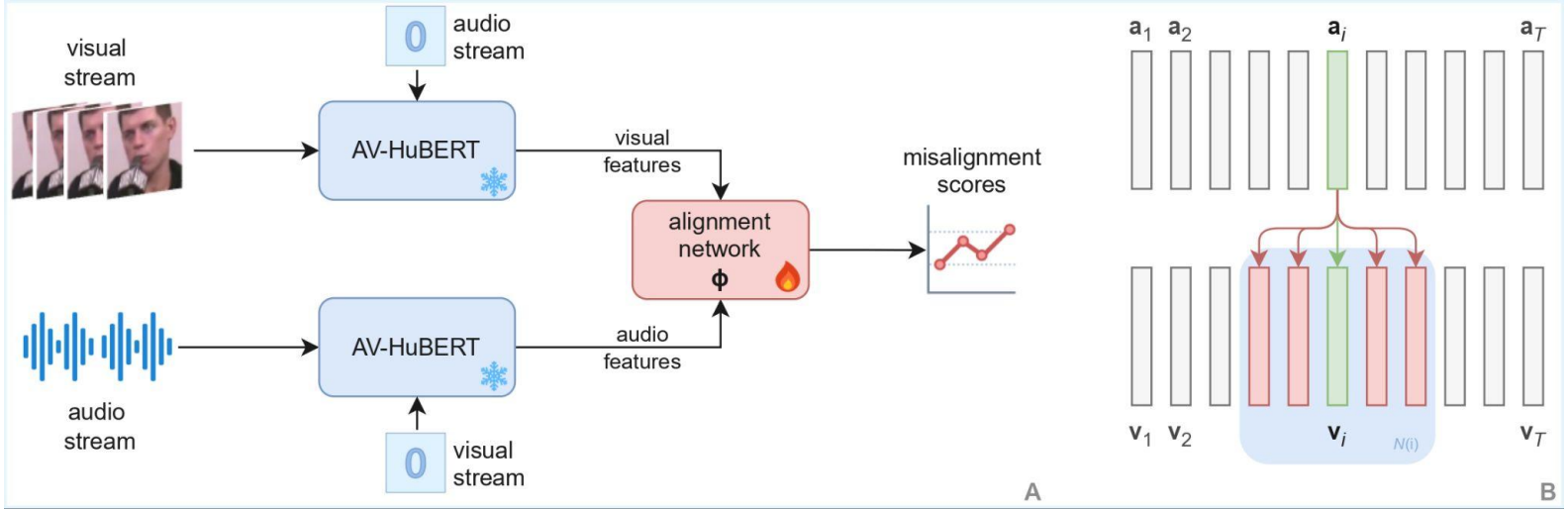
Metric: AUC		FakeAVCeleb		AV-Deepfake1M	
Method	Mod.	Trim: ✗	Trim: ✓	Trim: ✗	Trim: ✓
Silence classifier	A	98.4	54.8 ⇊ 43.6	98.2	50.6 ⇊ 47.6
RawNet2	A	99.9	97.3 ↓ 2.6	99.9	88.1 ⇊ 11.8
MDS (sup)	AV	90.4	73.8 ⇊ 16.6	99.2	54.9 ⇊ 44.3
AVAD (unsup)	AV	95.2	95.2 ≅ 0.0	52.9	52.9 ≅ 0.0

Over **98% AUC** for a small enough audio magnitude threshold

Performance of most supervised methods drops when leading silence is removed!

AVH-Align

One way to **circumvent** these unintended artifacts and biases is to **train only on real videos.**
We leverage pretrained self-supervised audio-visual features of AV-HuBERT.



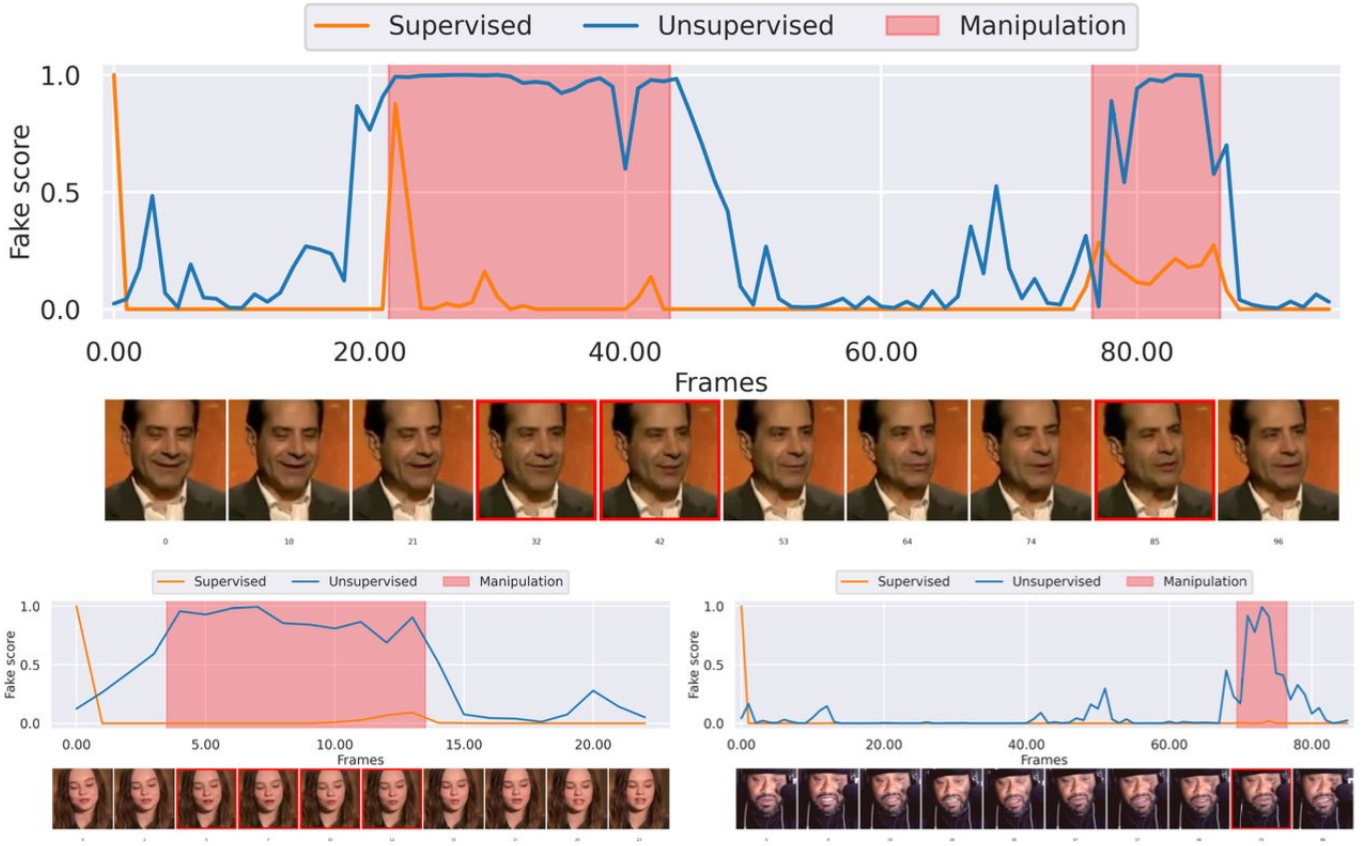
Proposed network to learn alignment between video and audio channels.
Left - shows the overall architecture; Right - illustrates the contrastive learning approach.

Results

Metric: AUC			FakeAVCeleb		AV-Deepfake1M	
Method	Train type	Train data	Trim: ✗	Trim: ✓	Trim: ✗	Trim: ✓
AVH-Align/sup	sup.	FAVC	99.2	99.2 ≅	69.0	63.6 ⇊
AVH-Align/sup	sup.	AV1M	77.5	70.8 ⇊	100.0	83.1 ⇊
AVAD	unsup.	LRS	84.5	84.7 ≅	54.3	54.3 ≅
SpeechForensics	unsup.	VoxCeleb2	**98.8**	**98.8** ≅	68.8	68.2 ≅
AVH-Align	unsup.	VoxCeleb2	94.6	94.6 ≅	**85.9**	**83.5** ↓

AVH-Align achieves the highest results, while maintaining robustness.

Qualitative results



Takeaways

- **Audio-video datasets may inherit unwanted artifacts and biases that differ between real and fakes.**
- **Supervised methods latch onto these artifacts and biases to make predictions.**
- **Learning only from real videos provides improved and robust results.**

Thank You!