Audio Information Research
Laboratory

A Multi-Stream Fusion Approach with One-Class Learning for Audio-Visual Deepfake Detection

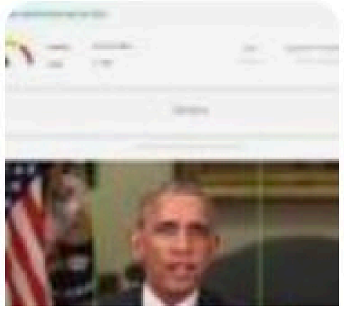
Kyungbok Lee, You Zhang, Zhiyao Duan

University of Rochester, USA

Deepfake

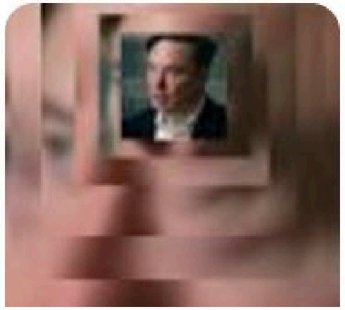
reutersinstitute.politics.ox.ac.uk

Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail



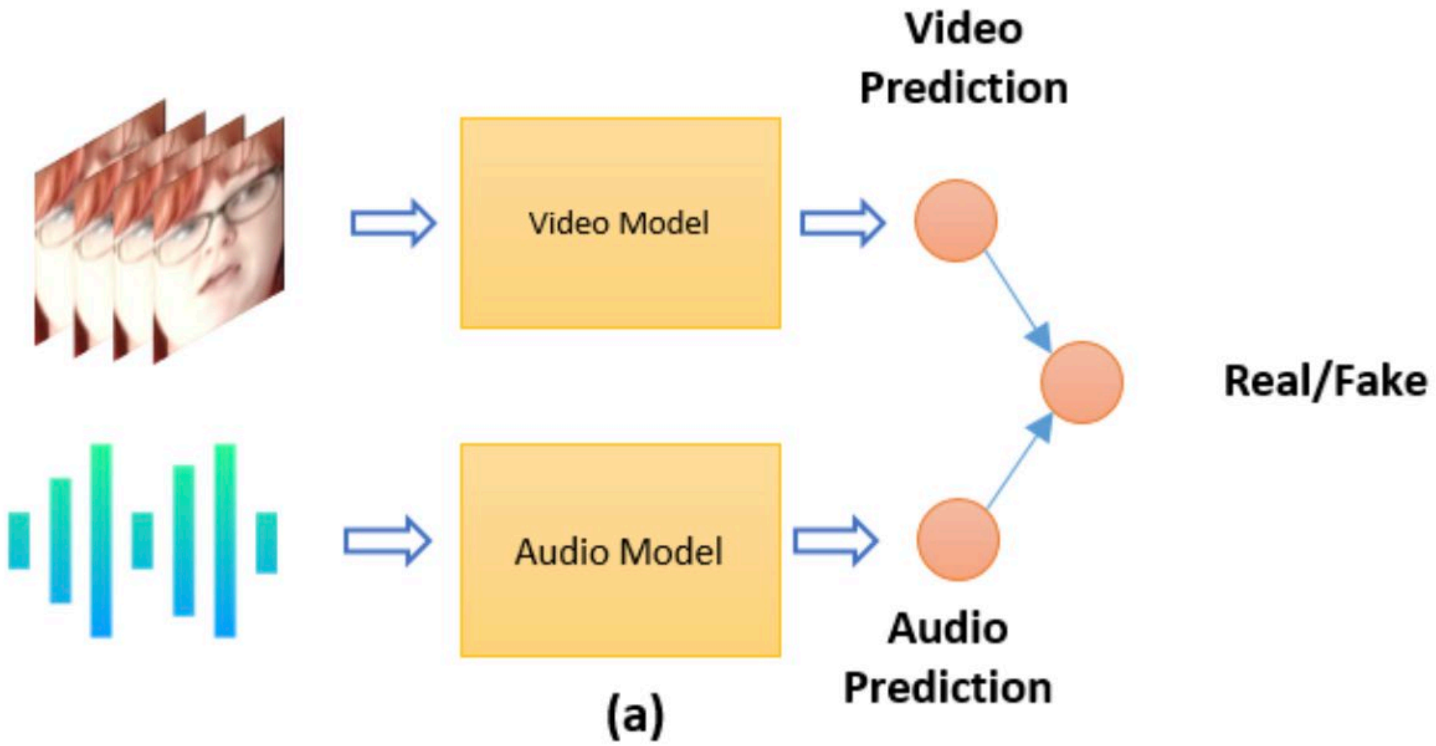
The New York Times

How ‘Deepfake Elon Musk’ Became the Internet’s Biggest Scammer



Audio-visual deepfake detection

- Use Both audio and visual modalities to detect whether a video is real or fake



Zhou, Y., & Lim, S. N. (2021). Joint audio-visual deepfake detection.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*

What is the limitation of current studies?

Motivation:

Robustness:

Current SOTA models perform poorly on unseen generation methods while new methods are invented rapidly.

Interpretability:

Current SOTA models do not provide interpretability on the model’s decision .

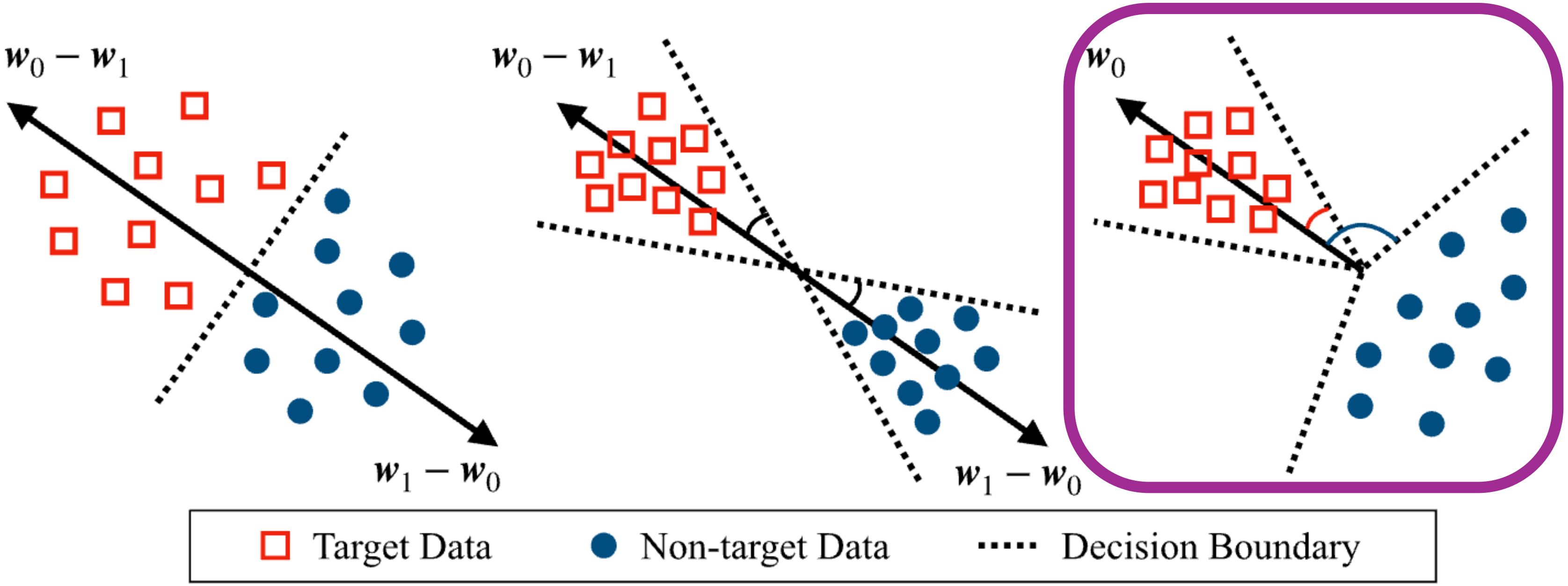
Contributions

- **New problem formation (new dataset)**
- **Implement OC-softmax on audio-visual detection problem**

Framework

OC-Softmax

OC-softmax pulls target data (real) together and spreads non-target data (fake). This feature **enhances robustness** to unseen attacks.



Zhang, Y., Jiang, F., & Duan, Z. (2021). One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, *28*, 937-941.

Dataset

Training set **Four test sets**

- We created a training set along with four test sets to evaluate **performance** on **unseen attacks** across four fake video categories.

Result

Model	RAFV	FAFV	FARV	Unsynced
Multilabel [30]	52.50 ± 2.50	88.12 ± 2.19	50.50 ± 1.80	49.50 ± 1.62
Multimodal-dissonance [18]	48.62 ± 6.81	62.12 ± 5.94	57.62 ± 1.88	49.62 ± 3.19
AVDF [29]	50.88 ± 0.96	86.38 ± 1.14	51.38 ± 1.63	**49.88 ± 2.30**
MRDF-CE [9]	54.38 ± 2.84	88.25 ± 0.83	47.25 ± 0.83	47.50 ± 0.61
MRDF-Margin [9]	55.12 ± 1.02	86.88 ± 1.85	47.62 ± 1.19	47.88 ± 1.14
MSOC (Ours)	**60.25 ± 2.19**	**89.88 ± 3.15**	**74.38 ± 5.41**	45.25 ± 1.64

- **The metric is Accuracy %**
- **Random Guess performance is 50%,**

Result

Model	RAFV	FAFV	FARV	Unsynced
Multilabel [30]	52.50 ± 2.50	88.12 ± 2.19	50.50 ± 1.80	49.50 ± 1.62
Multimodal-dissonance [18]	48.62 ± 6.81	62.12 ± 5.94	57.62 ± 1.88	49.62 ± 3.19
AVDF [29]	50.88 ± 0.96	86.38 ± 1.14	51.38 ± 1.63	$\mathbf{49.88 \pm 2.30}$
MRDF-CE [9]	54.38 ± 2.84	88.25 ± 0.83	47.25 ± 0.83	47.50 ± 0.61
MRDF-Margin [9]	55.12 ± 1.02	86.88 ± 1.85	47.62 ± 1.19	47.88 ± 1.14
MSOC (Ours)	$\mathbf{60.25 \pm 2.19}$	$\mathbf{89.88 \pm 3.15}$	$\mathbf{74.38 \pm 5.41}$	45.25 ± 1.64

- We believe MSOC excels particularly well on the Audio-only fake dataset (**FARV**) compared to other models because fake **audio is entirely generated**, enabling **OC-softmax to distinguish between real and fake instances more effectively**.

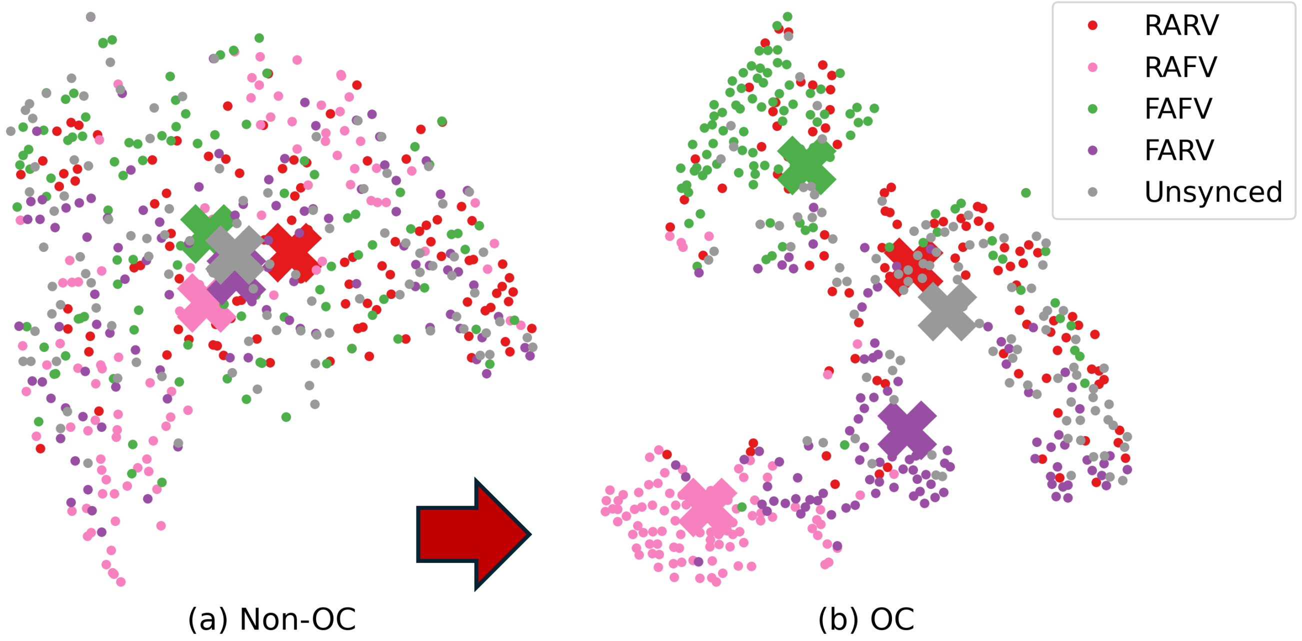
Result

Model	RAFV	FAFV	FARV	Unsynced
Multilabel [30]	52.50 ± 2.50	88.12 ± 2.19	50.50 ± 1.80	49.50 ± 1.62
Multimodal-dissonance [18]	48.62 ± 6.81	62.12 ± 5.94	57.62 ± 1.88	49.62 ± 3.19
AVDF [29]	50.88 ± 0.96	86.38 ± 1.14	51.38 ± 1.63	$\mathbf{49.88 \pm 2.30}$
MRDF-CE [9]	54.38 ± 2.84	88.25 ± 0.83	47.25 ± 0.83	47.50 ± 0.61
MRDF-Margin [9]	55.12 ± 1.02	86.88 ± 1.85	47.62 ± 1.19	47.88 ± 1.14
MSOC (Ours)	$\mathbf{60.25 \pm 2.19}$	$\mathbf{89.88 \pm 3.15}$	$\mathbf{74.38 \pm 5.41}$	45.25 ± 1.64

- However, the model struggles with Unsynchronized fake videos (**Unsynced**), as the **model does not have an explicit module** for detecting synchronization inconsistencies.

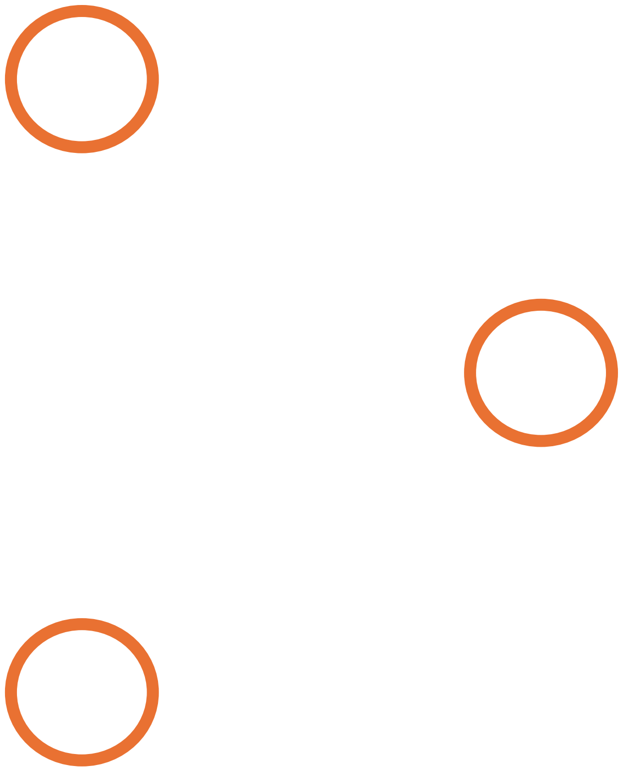
OC-softmax enhances generalizability

- Feature embedding visualization demonstrates that model trained with OC-softmax separates unseen Fake data from real data better



t-SNE feature embedding visualization on four test sets

Interpretability



Conclusion

- Improves **generalizability** against unseen deepfake generation methods

- provides **interpretability**, offering the ability to identify which modality is fake

Thank you