Spatiotemporal Face Alignment for Generalizable Deepfake Detection

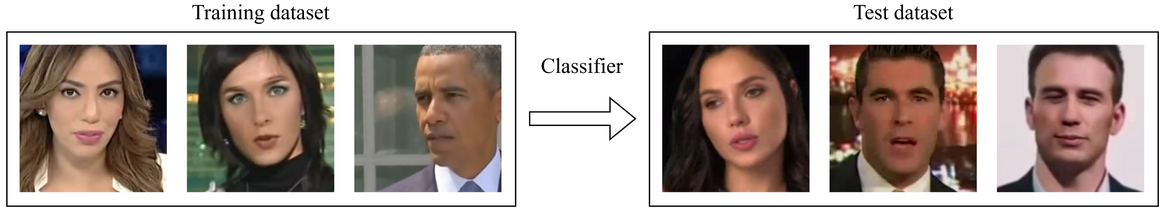
Alejandro Cobo[1], Roberto Valle[1], José M. Buenaposada[2], Luis Baumela[1]

Universidad Politécnica de Madrid[1] Universidad Rey Juan Carlos[2]

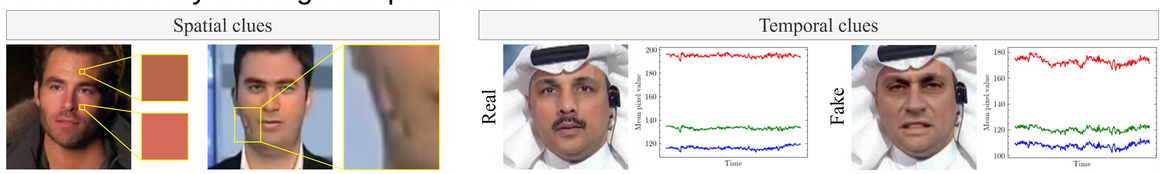
Grant PID2022-137581OB-I00 funded by:

Abstract

Generalization is an essential property for any deepfake detection method to be used with real world data. At test time, the model must handle **unseen** types of manipulations.



Deepfakes can be spotted via **spatial** and **temporal** clues. However, most methods cannot correctly leverage temporal clues.



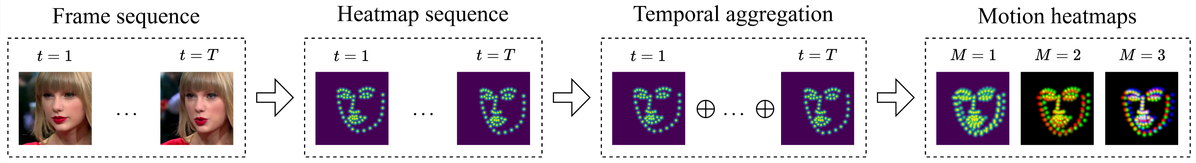
Experiments

Deepfake detection video-level AUC (%). All methods are trained on FF++ (c23).

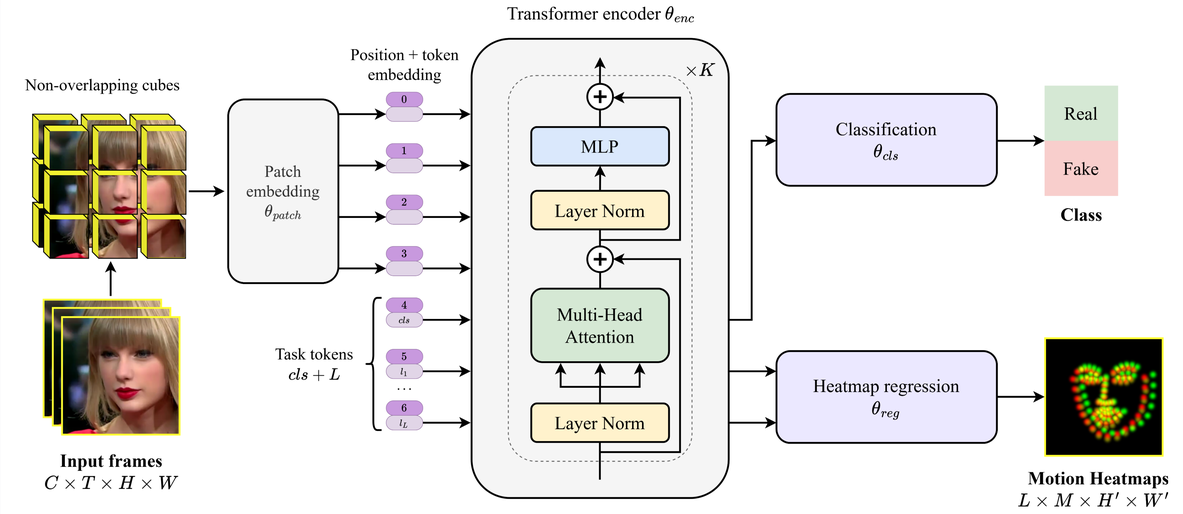
Method	Cross-dataset		Cross-manipulation		Avg.
	CDF	DFDCP	FSh	DFo	
Multi-Att	75.70	-	66.00	77.70	-
Face X-Ray	79.50	**80.92**	92.80	86.80	85.01
FTCN	86.90	74.00	98.80	98.80	89.63
AltFreezing	89.50	70.91 *	99.40	99.30	89.78
TALL-Swin	**90.79**	-	99.67	**99.62**	-
Baseline	87.17	78.45	99.55	98.50	90.92
SFA $(M = 1)$	89.70	79.84	99.82	99.17	92.13
SFA $(M = 2)$	89.52	80.58	**99.84**	99.24	**92.30**

Method

We leverage a **spatiotemporal face alignment** task to focus on the temporal consistency of facial movements. This is done via **motion heatmaps**. Multi-channel heatmaps ($M \geq 2$) can be used to represent the direction of the movement.



Our multi-task framework (**SFA**) combines a deepfake detection task with a face alignment task in a video transformer architecture:



The benefits of our approach are two-fold:

1. The network focuses more on temporal consistency, improving the detection of temporal artifacts.
2. Attention to unrelated parts of the input sequence is reduced.