Spatiotemporal Face Alignment for Generalizable Deepfake Detection

Alejandro Cobo, Roberto Valle, José M. Buenaposada, Luis Baumela

Grant PID2022-137581OB-I00 funded by:

Outline

1. Introduction
2. Our approach
3. Experimental results
4. Qualitative results
5. Conclusions & future work

Introduction

Manipulation methods:

- Face swapping
- Face reenactment
- Entire face synthesis
- Face editing

Introduction

Manipulation methods:

- **Face swapping**
- Face reenactment
- Entire face synthesis
- Face editing

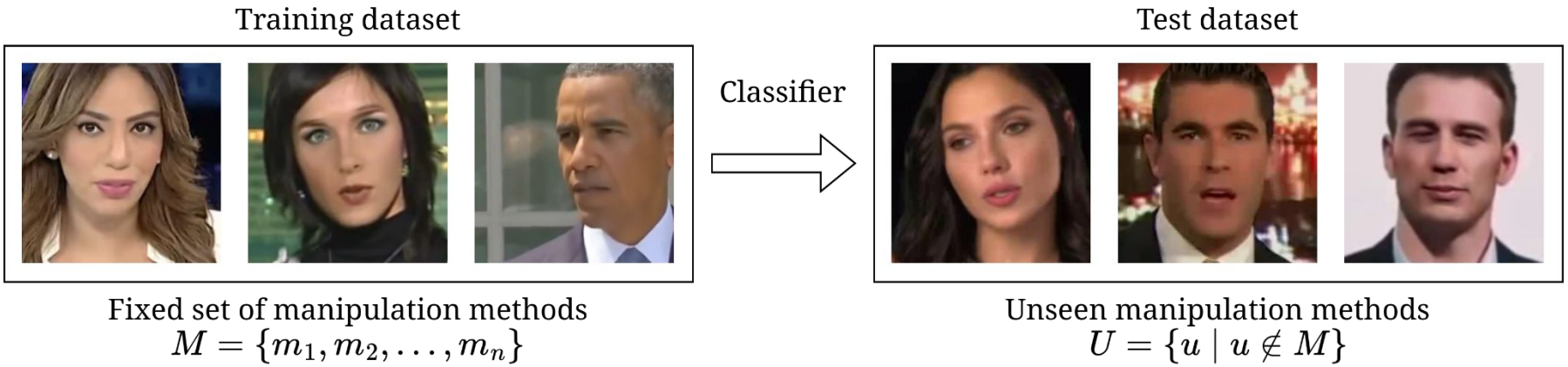
Introduction

Manipulation methods:

- Face swapping
- **Face reenactment**
- Entire face synthesis
- Face editing

Introduction

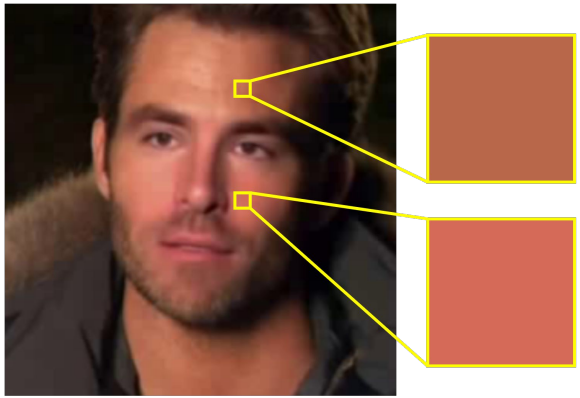
Challenge: generalization to unseen manipulations



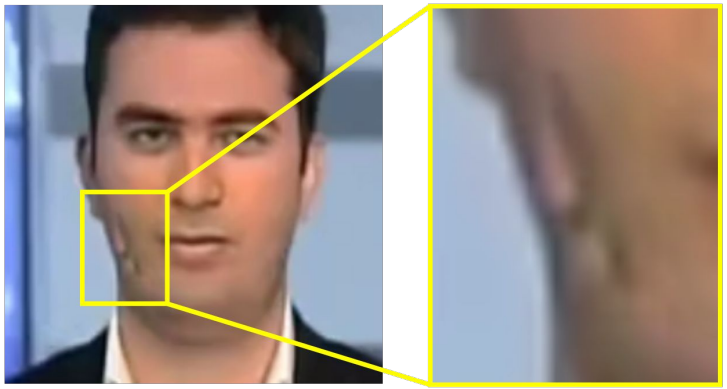
Introduction

Spatial clues

Photometric inconsistencies



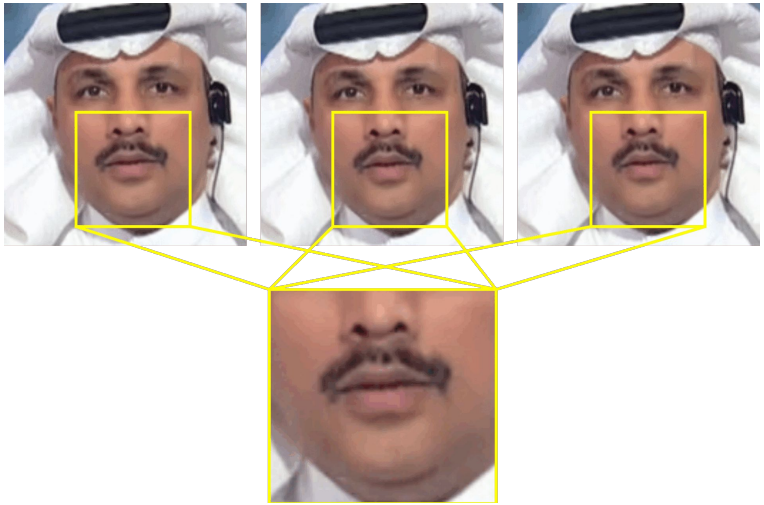
Blending artifacts



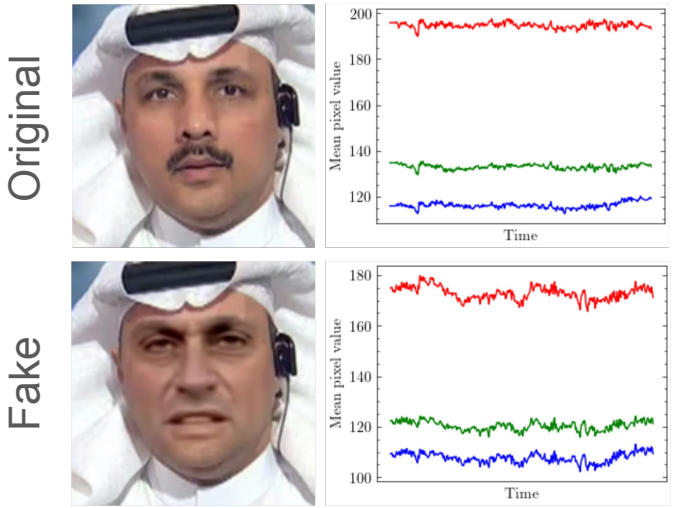
Introduction

Temporal clues

Facial attributes



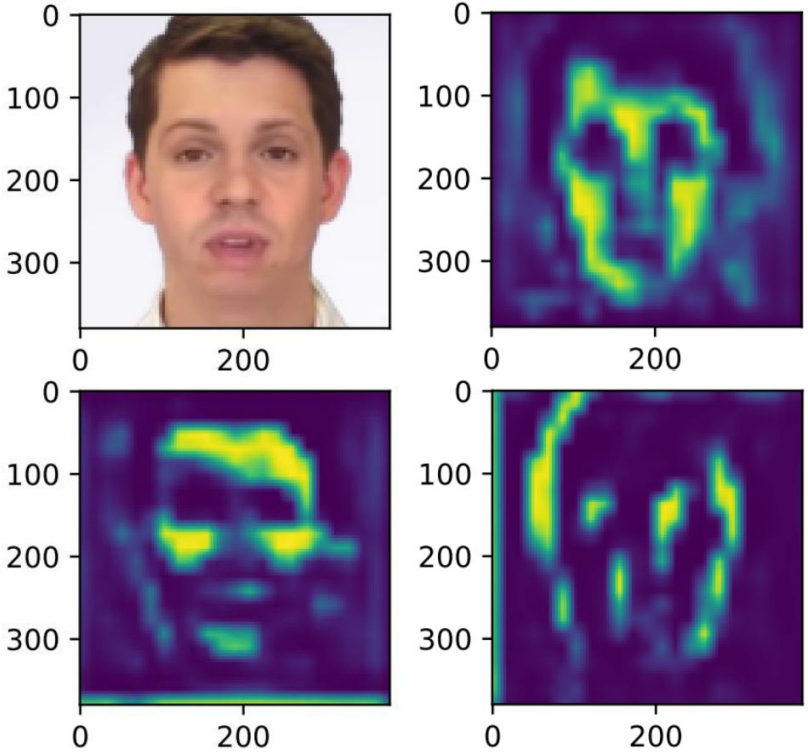
Biological signals



Introduction

Image-based methods:

- **Multi-Att**
- Face X-Ray

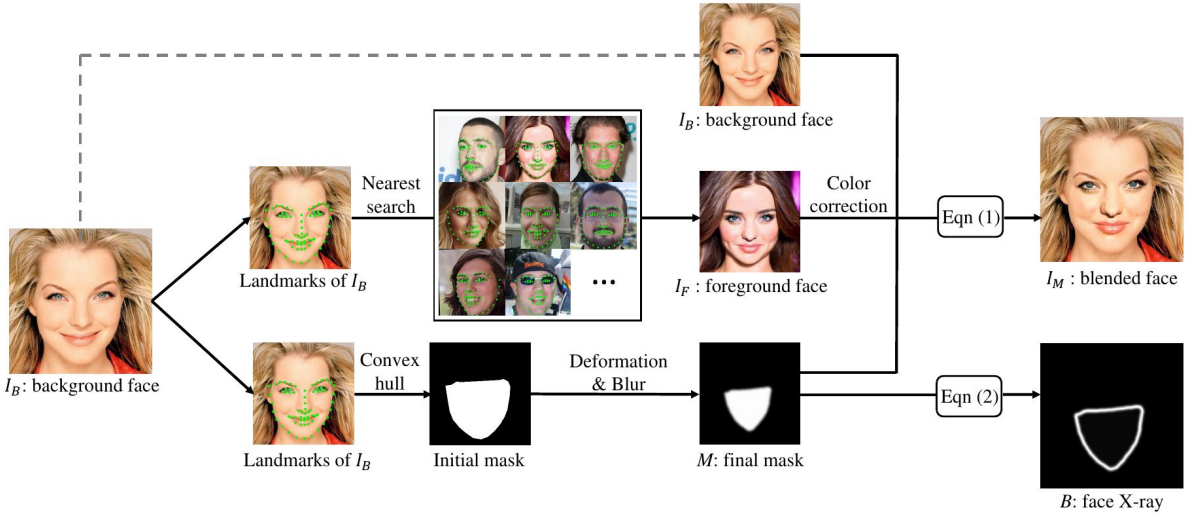


H. Zhao et al., “Multi-attentional Deepfake Detection” (CVPR 2021)

Introduction

Image-based methods:

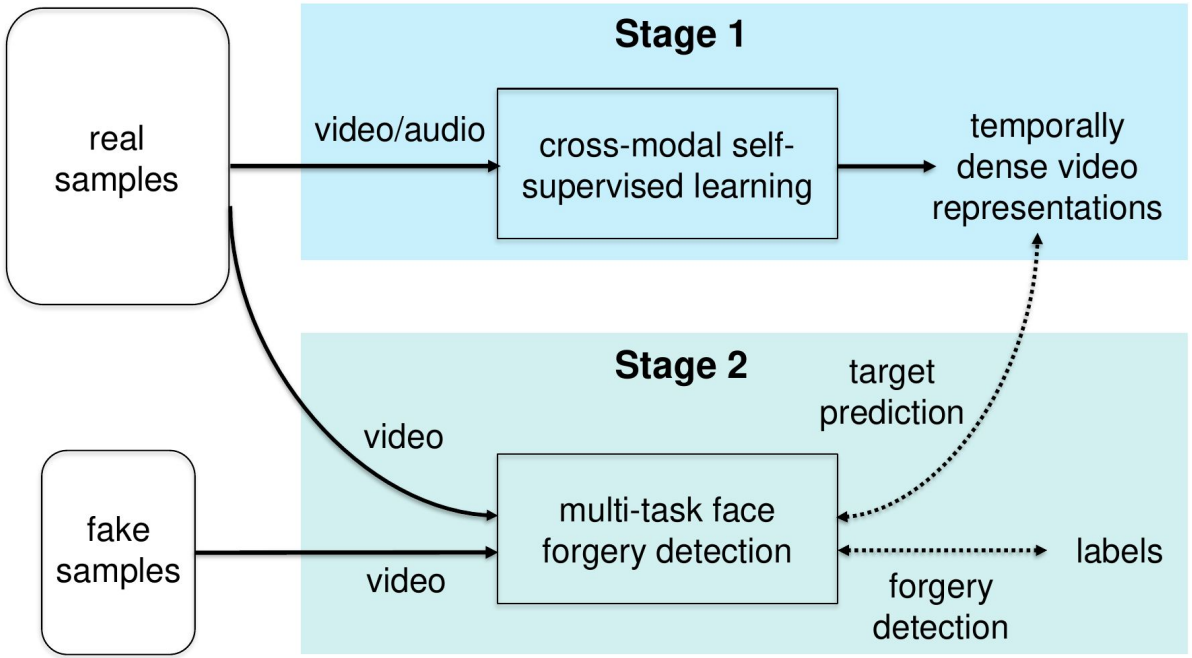
- Multi-Att
- **Face X-Ray**



Introduction

Video-based methods:

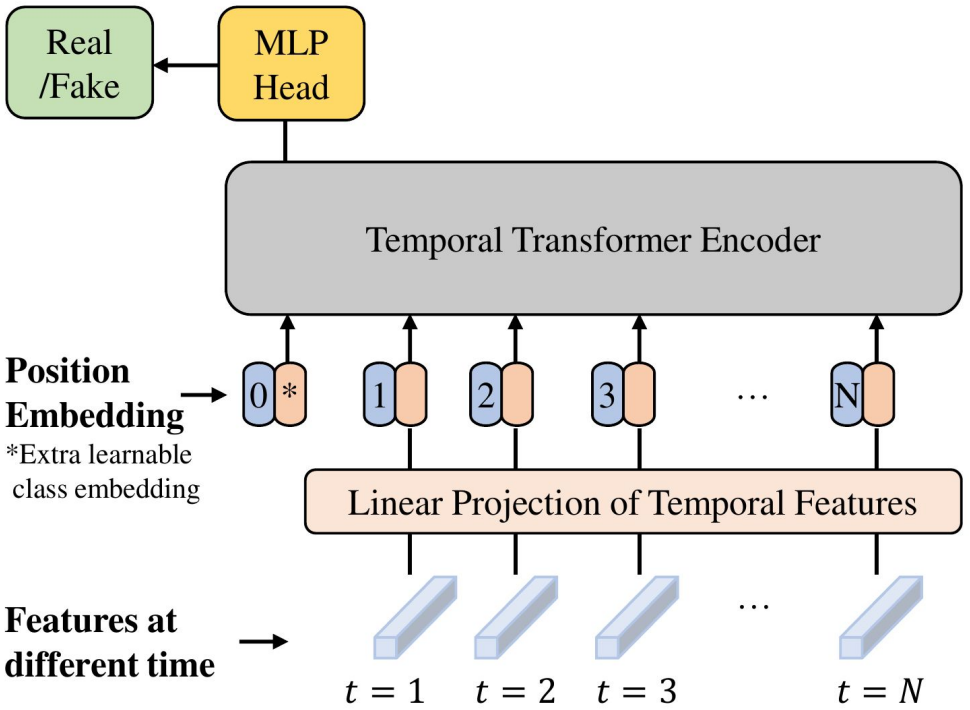
- **RealForensics**
- FTCN
- TALL-Swin



Introduction

Video-based methods:

- RealForensics
- **FTCN**
- TALL-Swin

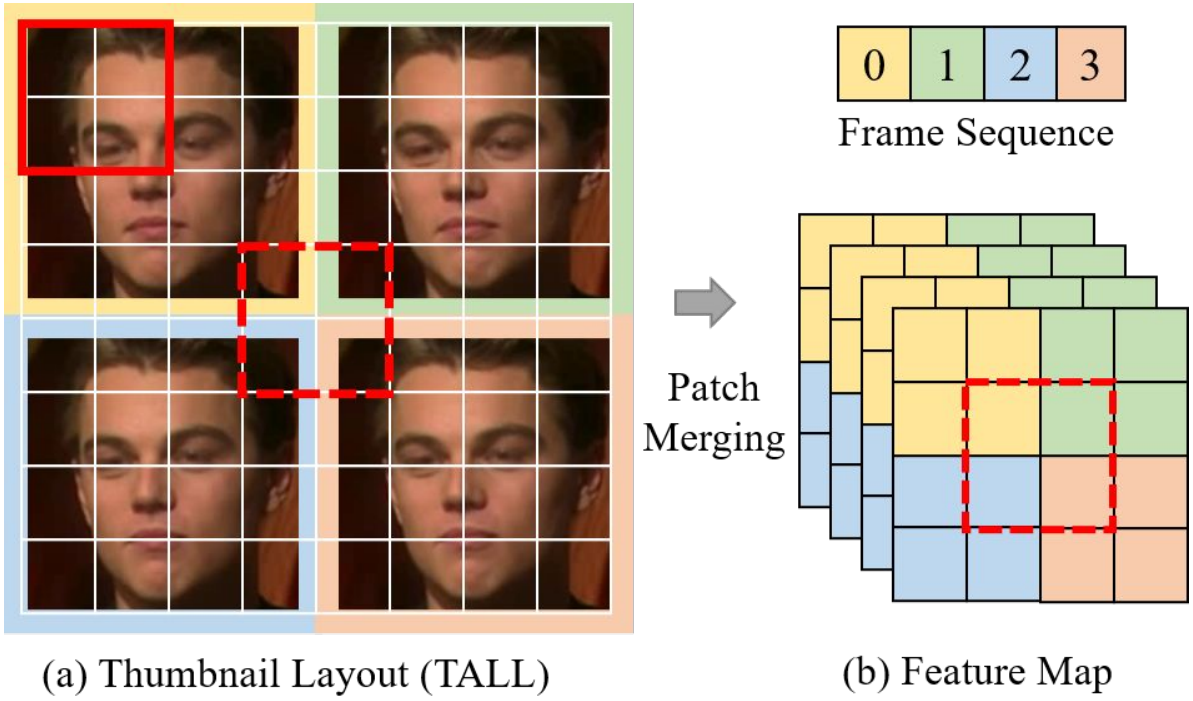


Y. Zheng et al., “Exploring Temporal Coherence for More General Video Face Forgery Detection” (ICCV 2021)

Introduction

Video-based methods:

- RealForensics
- FTCN
- **TALL-Swin**



Y. Xu et al., “TALL: Thumbnail Layout for Deepfake Video Detection” (ICCV 2023)

Introduction

State-of-the-art methods:

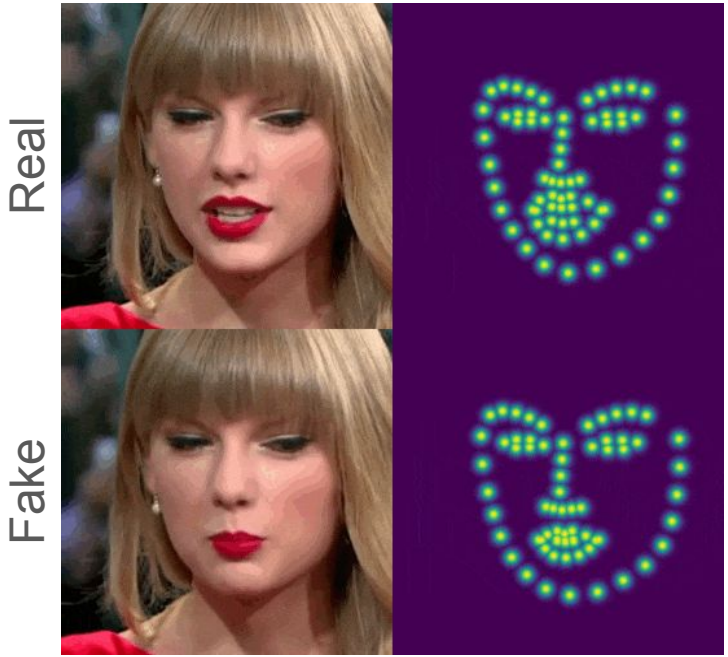
- Image-based methods cannot spot temporal clues.
- Most video-based methods do not explicitly track facial movements.

Our approach:

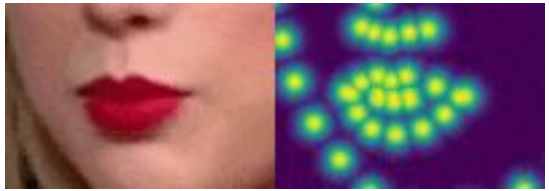
- Video backbone to detect temporal inconsistencies.
- Leverage face alignment to track facial movements.

Our approach

Spatiotemporal face alignment (**SFA**)



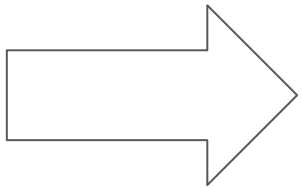
Unnatural movements



Our approach

Motion heatmaps

Temporal aggregation

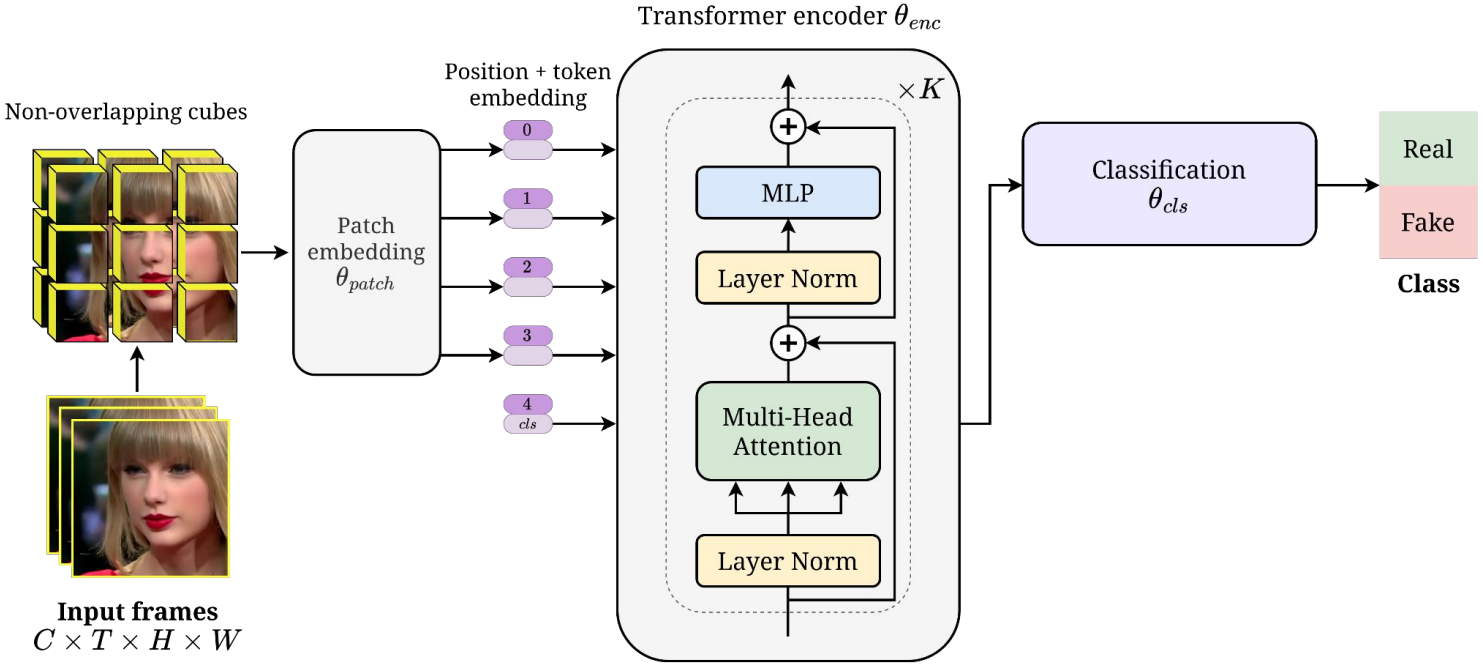


Our approach

Encoding motion direction

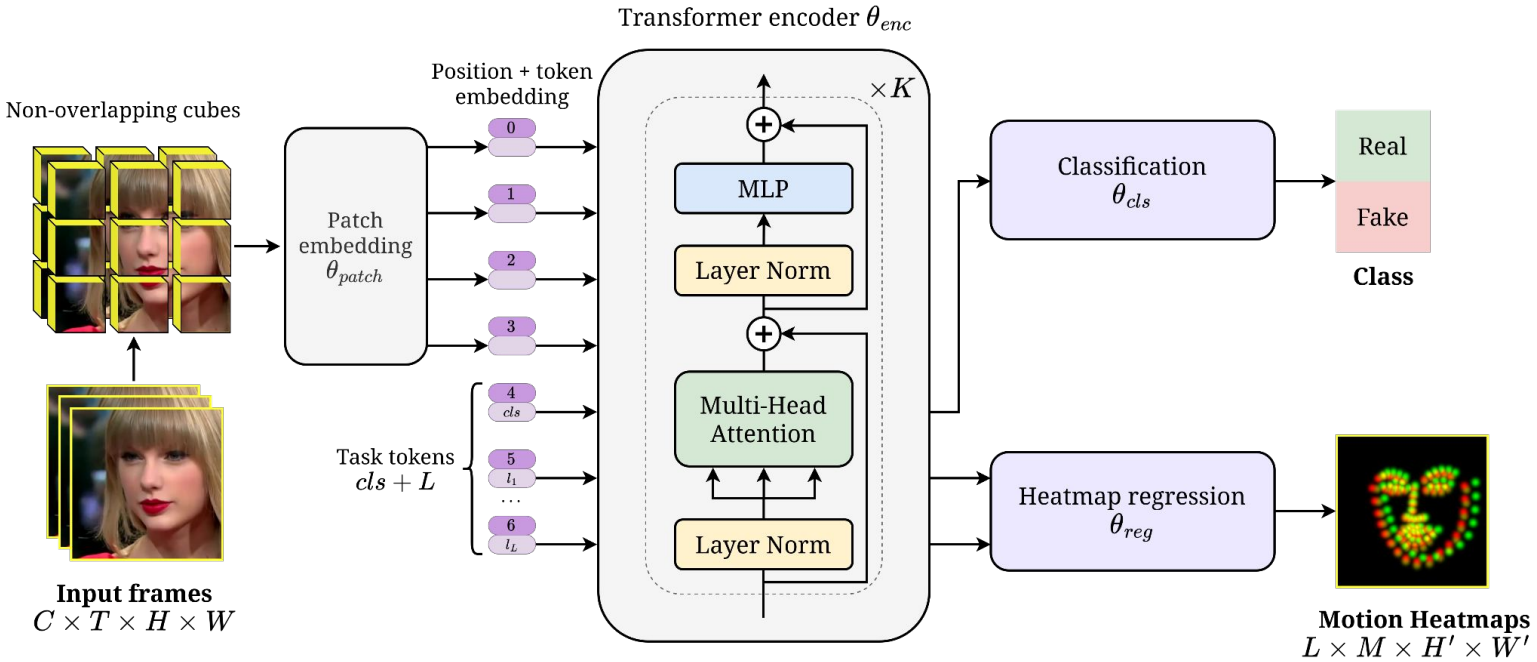
Our approach

Baseline / single-task



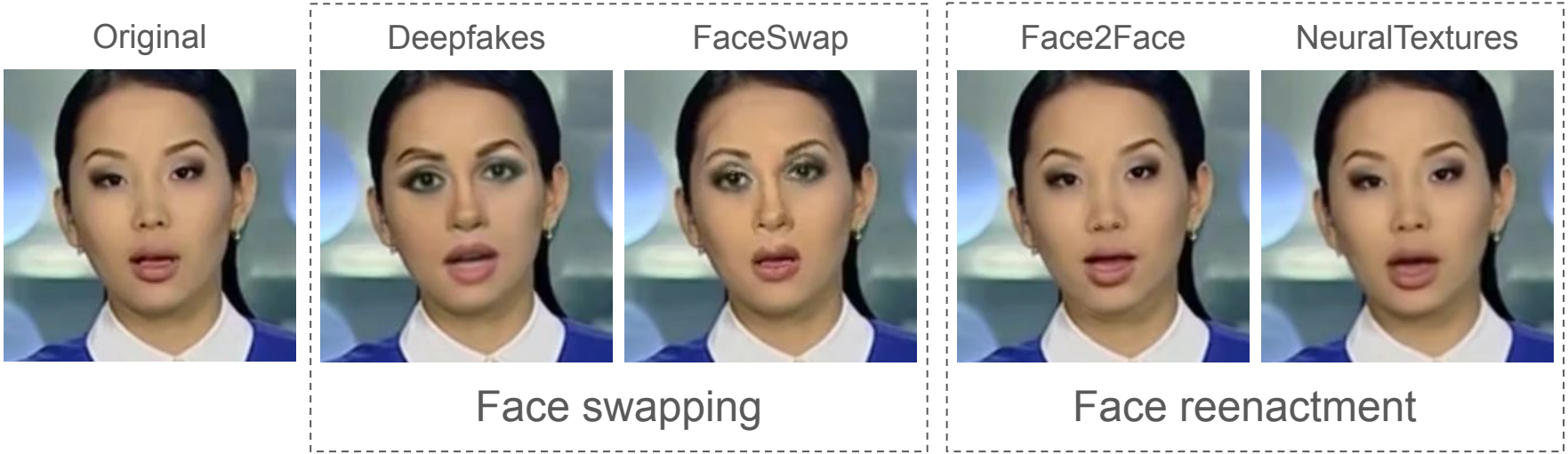
Our approach

Multi-task



Experimental results

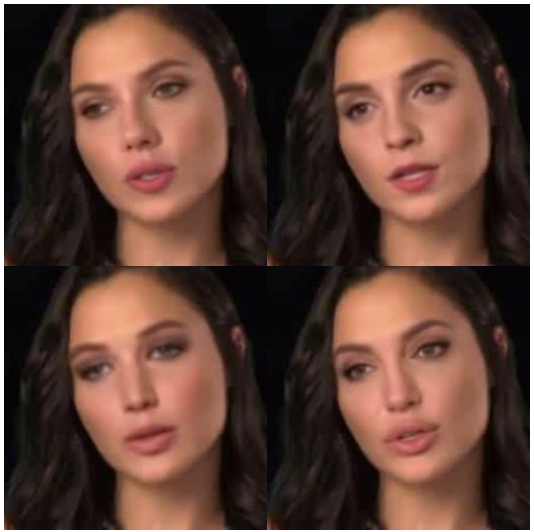
Train dataset: FaceForensics++



Experimental results

Test datasets

Celeb-DF

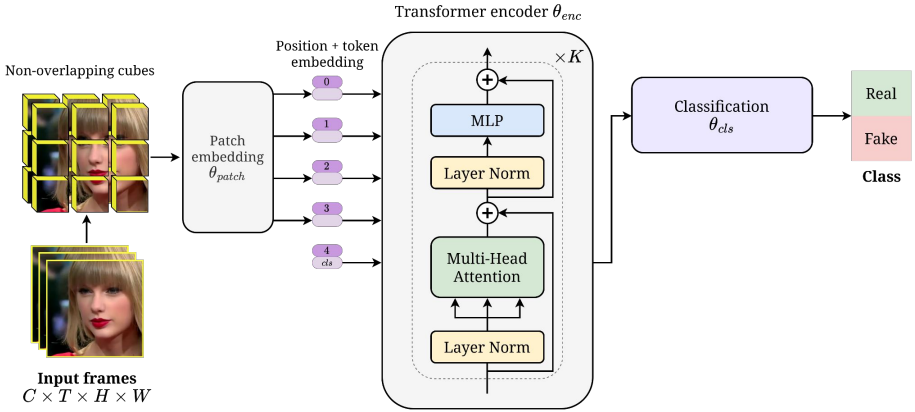


DeeperForensics

FaceShifter

Experimental results

Baseline / single-task

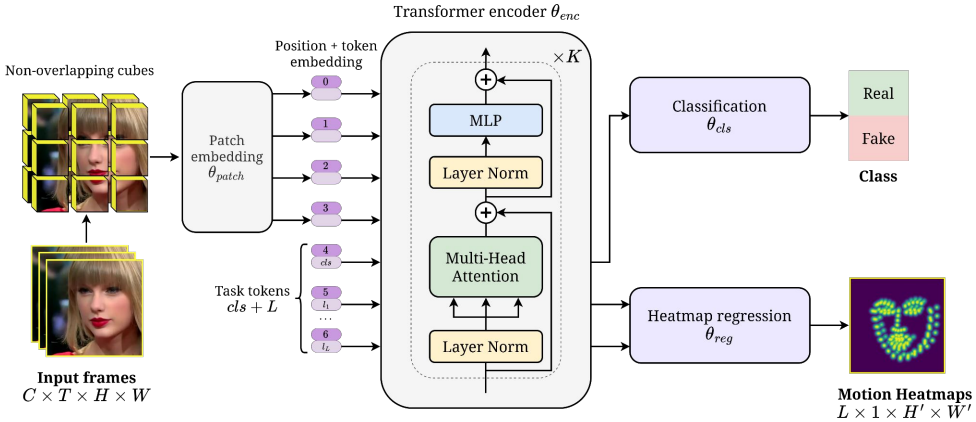


Video-level AUC (%)

Method	Cross-dataset		Cross-manipulation		Avg.
	CDF	DFDCP	FSh	DFo	
Baseline	87.17	78.45	99.55	98.50	90.92

Experimental results

Multi-task (1 motion channel)

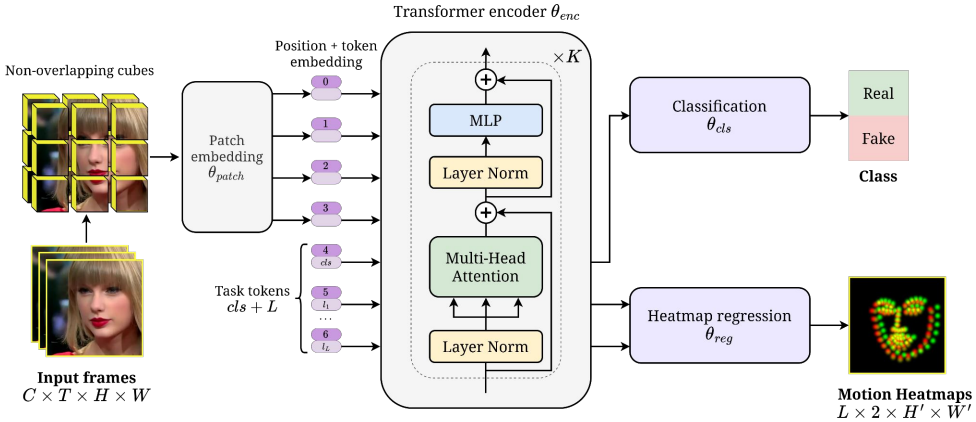


Video-level AUC (%)

Method	Cross-dataset		Cross-manipulation		Avg.
	CDF	DFDCP	FSh	DFo	
Baseline	87.17	78.45	99.55	98.50	90.92
$M = 1$	**89.70**	79.84	99.82	99.17	92.13

Experimental results

Multi-task (2 motion channels)



Video-level AUC (%)

Method	Cross-dataset		Cross-manipulation		Avg.
	CDF	DFDCP	FSh	DFo	
Baseline	87.17	78.45	99.55	98.50	90.92
$M = 1$	**89.70**	79.84	99.82	99.17	92.13
$M = 2$	89.52	**80.58**	**99.84**	**99.24**	**92.30**

Experimental results

Video-level AUC (%)

	Method	Cross-dataset		Cross-manipulation		Avg.
		CDF	DFDCP	FSh	DFo	
Image methods	FWA	69.50	-	65.50	50.20	-
	PatchForensics	69.60	-	57.80	81.80	-
	Xception	73.70	-	72.00	84.50	-
	CNN-aug	75.60	-	65.70	74.40	-
	Multi-Att	75.70	-	66.00	77.70	-
	Face X-Ray	79.50	**80.92**	92.80	86.80	85.01
	SLADD	79.70	-	-	-	-
Video methods	CNN-GRU	69.80	-	80.80	74.10	-
	LipForensics	82.40	-	97.10	97.60	-
	ISTVT	84.10	74.20	99.30	98.60	-
	FTCN	86.90	74.00	98.80	98.80	89.63
	RealForensics	86.90	-	99.70	99.30	-
	AltFreezing	89.50	70.91 *	99.40	99.30	89.78
	TALL-Swin	**90.79**	-	99.67	**99.62**	-
	SFA (ours)	89.52	80.58	**99.84**	99.24	**92.30**

Qualitative results

Baseline

Multi-task

Conclusions

- Face alignment improves deepfake detection
  - Accurate detection of temporal artifacts
  - Increased attention to more relevant facial regions

Future work

- Generalization to future manipulation methods
- Interpretability with pseudo-fake generation techniques
- Localization of manipulated regions