MARLIN: Masked Autoencoder for facial video Representation LearnINg

In CVPR 2023 (TUE-AM-142)

Zhixi Cai[1], Shreya Ghosh[2], Kalin Stefanov[1], Abhinav Dhall[3,1], Jianfei Cai[1], Hamid Rezatofighi[1], Reza Haffari[1], Munawar Hayat[1]

[1]Monash University, [2]Curtin University, [3]Indian Institute of Technology Ropar

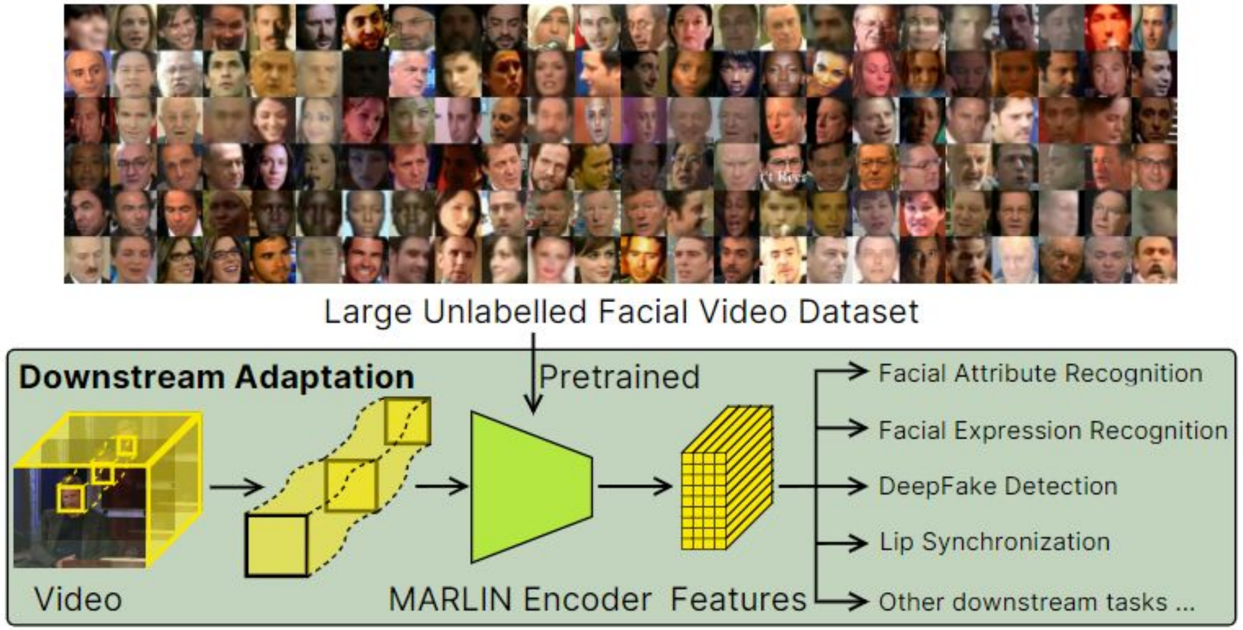
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

MONASH University

Curtin University

Abstract

- Facial representation learning provide important cues for non-verbal human behaviour analysis
- Universal facial representation learning from videos can transfer across
  - Facial Attribute Recognition (FAR)
  - Facial Expression Recognition (FER)
  - DeepFake Detection (DFD)
  - Lip Synchronization (LS)
  - And many more


Large Unlabelled Facial Video Dataset
Downstream Adaptation
Pretrained
Video
MARLIN Encoder
Features
Facial Attribute Recognition
Facial Expression Recognition
DeepFake Detection
Lip Synchronization
Other downstream tasks ...

Challenges and Prior Work

Challenges

- Data collection and annotation - resource expensive and time consuming process
- Spatio-temporal modelling for universal representation

Prior Work

Works closely related to MARLIN are:

- Image-based facial encoding
  - Exploring training dataset properties in terms of size and quality [1]
  - Performing pre-training in visual-linguistic way [2]

[1] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. In ECCV, pages 107-125, 2022.
[2] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General Facial Representation Learning in a Visual-Linguistic Manner. In CVPR, pages 18697–18709, 2022.

Contributions

1. ***MARLIN***, a universal facial encoder that learns robust representations from non-annotated web-crawled facial videos in a self-supervised manner

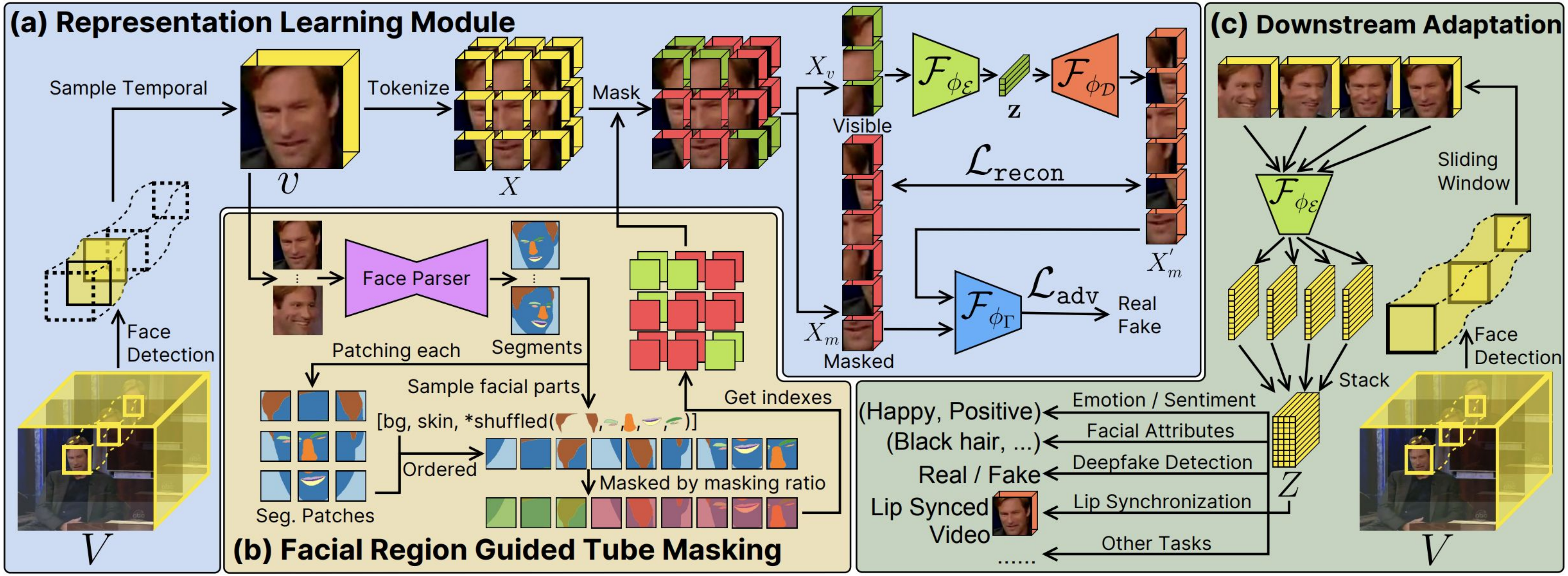
2. ***Fasking***, a facial region-guided tube masking strategy that reconstructs facial regions from densely masked areas. This approach captures both local and global aspects in facial videos, aiding in the acquisition of generic and transferable features

3. Demonstrate generalization capability of MARLIN for ***Facial Attribute Recognition***, ***Facial Expression Recognition***, ***Deepfake Detection***, ***Lip Synchronization***, and even in ***few shot*** settings

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Method Overview

- The pipeline for MARLIN divided into three parts (a), (b) and (c)

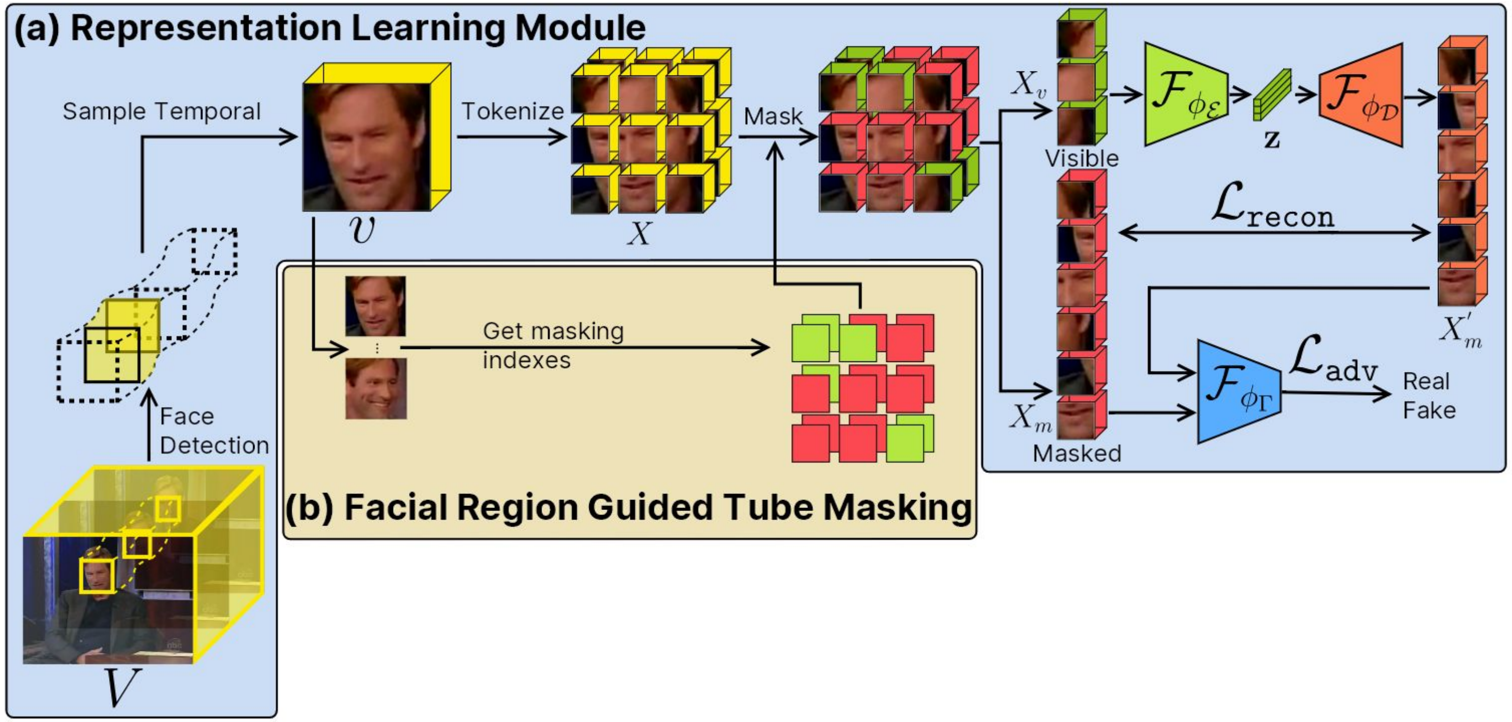


Cai et al. 2023

Method

(a) Representation Learning Module

- Masked autoencoder
- Pre-training performed using
  - Reconstruction loss
  - Adversarial loss

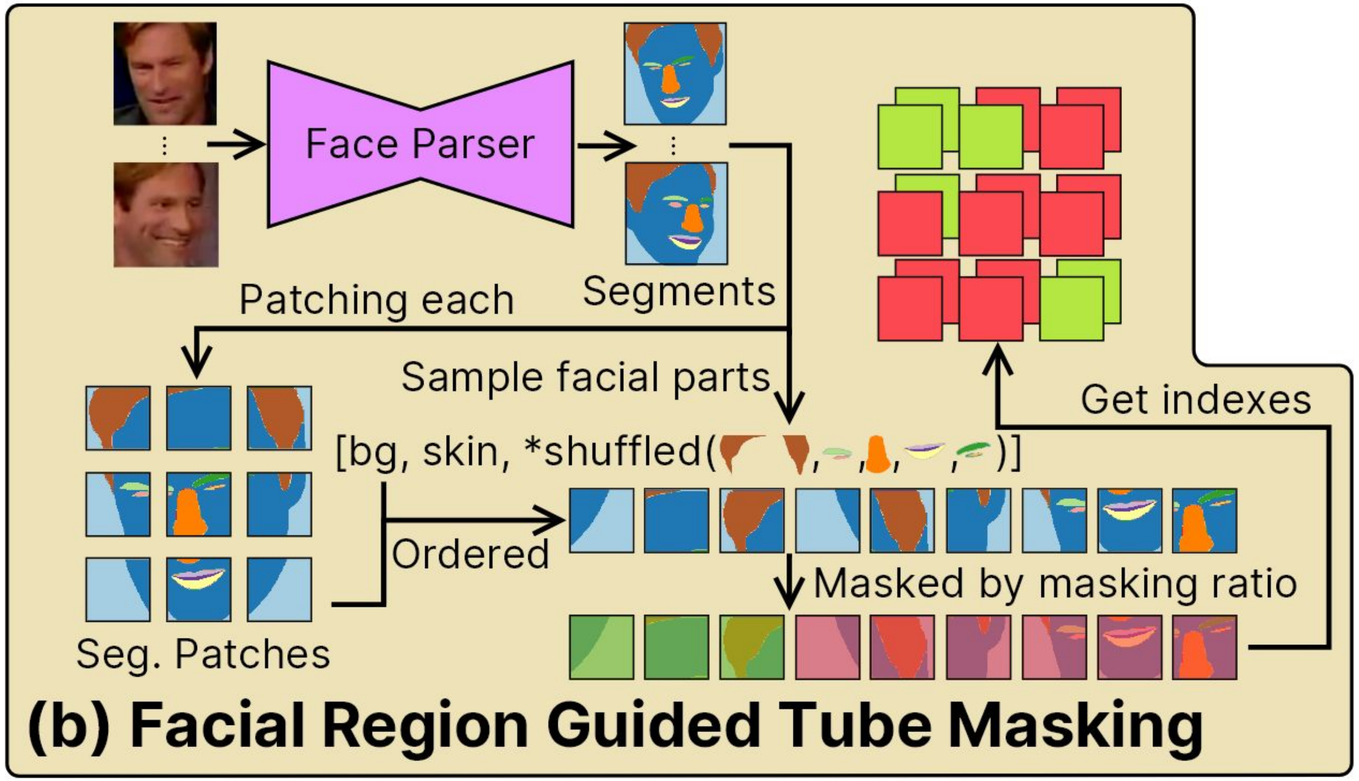


(a) Representation Learning Module
Sample Temporal
Tokenize
Mask
X_v
Visible
$\mathcal{F}_{\phi_\mathcal{E}}$
$\mathbf{z}$
$\mathcal{F}_{\phi_\mathcal{D}}$
v
X
$\mathcal{L}_{\text{recon}}$
Get masking indexes
X'_m
Face Detection
$\mathcal{F}_{\phi_\Gamma}$
$\mathcal{L}_{\text{adv}}$
Real
Fake
X_m
Masked
(b) Facial Region Guided Tube Masking
V

Method

(b) Facial-region guided tube masking (Fasking)

- Used a pre-trained face parser to segment facial components
- Random sample permutation of facial components as a masking priority queue
- Mask the cubes based on the queue

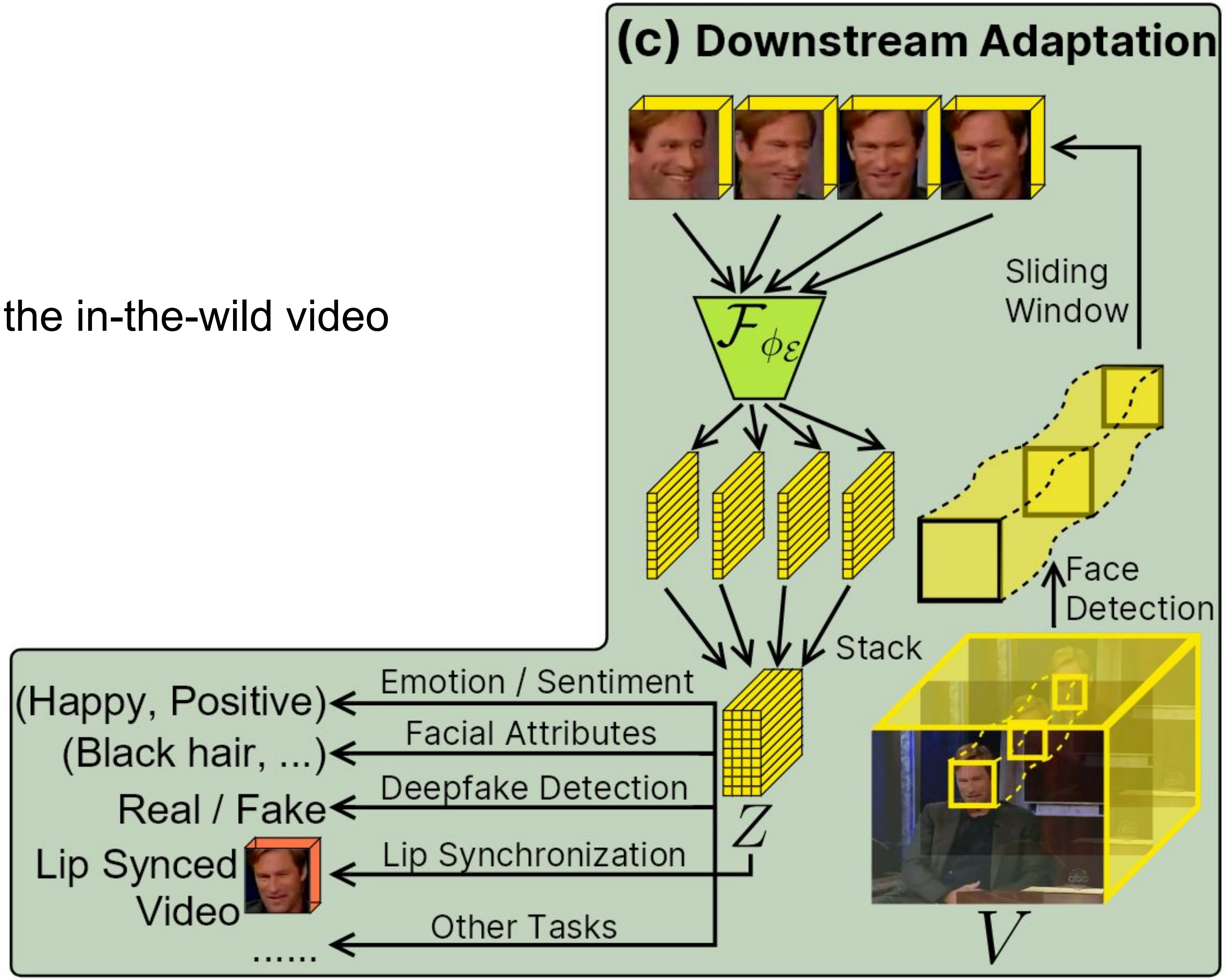


(b) Facial Region Guided Tube Masking

Method

(c) Downstream adaptation

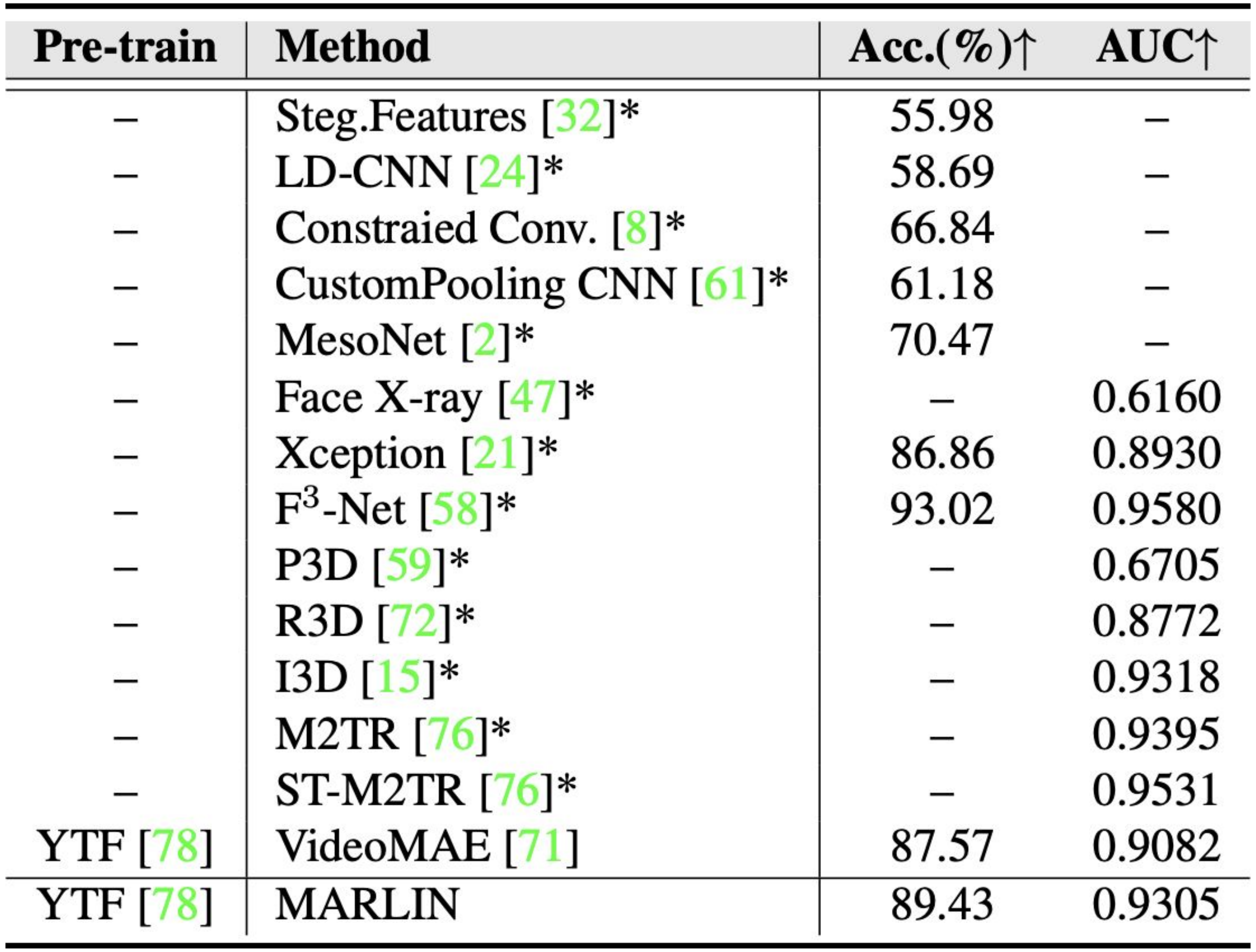
- We use sliding window to get clips of the in-the-wild video input for extracting features
- Two modes for adapting
  - Finetune the encoder and classifier head
  - Freeze the encoder and only train the head



Results

Deepfake Detection

- Finetune MARLIN for Deepfake Detection
- We evaluate it in FaceForensics++ dataset
- Supervised methods*



Pre-train	Method	Acc.(%)↑	AUC↑
–	Steg.Features [32]*	55.98	–
–	LD-CNN [24]*	58.69	–
–	Constraied Conv. [8]*	66.84	–
–	CustomPooling CNN [61]*	61.18	–
–	MesoNet [2]*	70.47	–
–	Face X-ray [47]*	–	0.6160
–	Xception [21]*	86.86	0.8930
–	F^3-Net [58]*	93.02	0.9580
–	P3D [59]*	–	0.6705
–	R3D [72]*	–	0.8772
–	I3D [15]*	–	0.9318
–	M2TR [76]*	–	0.9395
–	ST-M2TR [76]*	–	0.9531
YTF [78]	VideoMAE [71]	87.57	0.9082
YTF [78]	MARLIN	89.43	0.9305

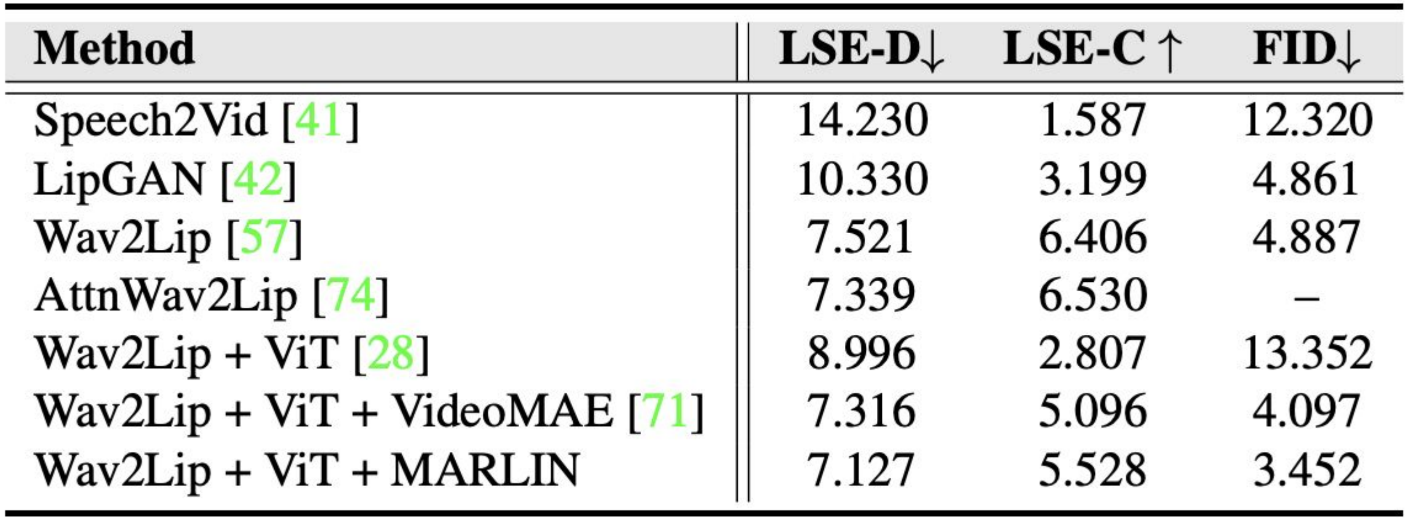
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

MONASH University

Results

Deepfake Generation

- Adapt MARLIN for Facial Video Generation
- We evaluate it in LRS2 dataset



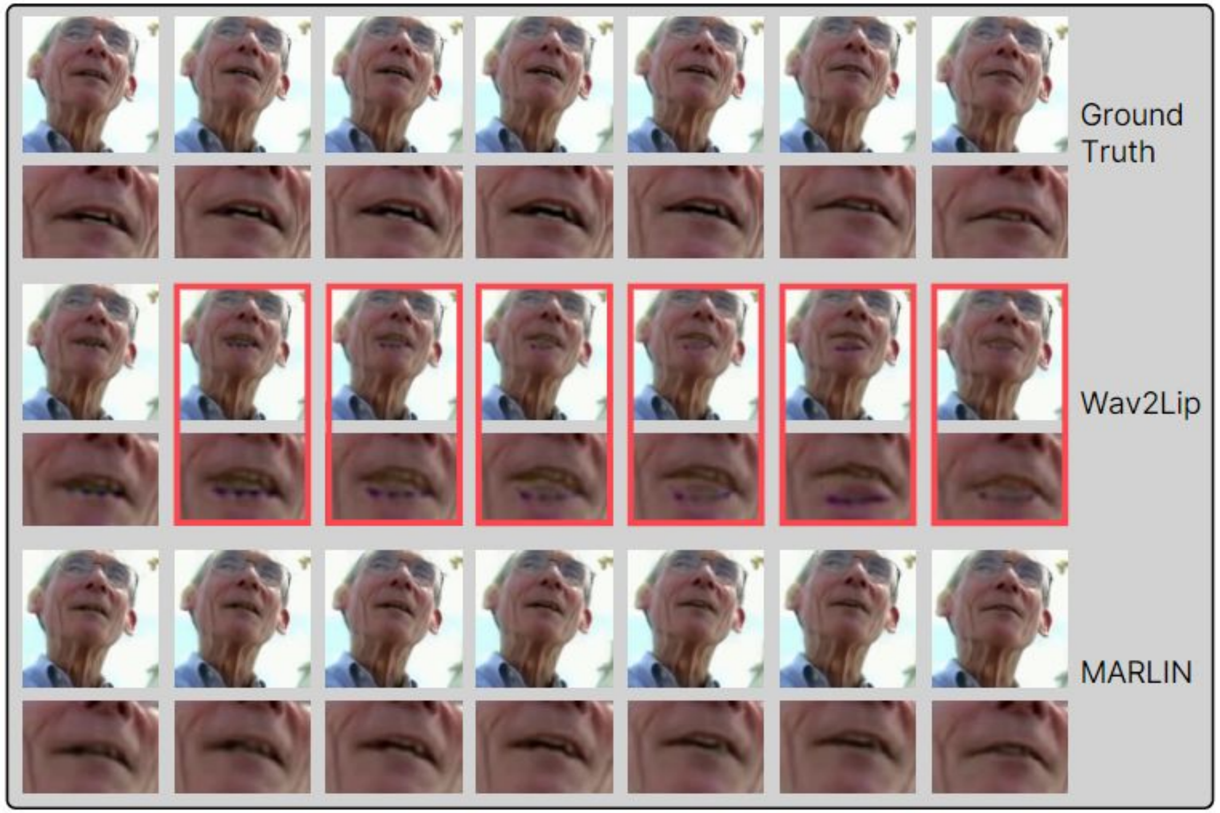
Method	LSE-D↓	LSE-C↑	FID↓
Speech2Vid [41]	14.230	1.587	12.320
LipGAN [42]	10.330	3.199	4.861
Wav2Lip [57]	7.521	6.406	4.887
AttnWav2Lip [74]	7.339	6.530	–
Wav2Lip + ViT [28]	8.996	2.807	13.352
Wav2Lip + ViT + VideoMAE [71]	7.316	5.096	4.097
Wav2Lip + ViT + MARLIN	7.127	5.528	3.452

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Results

Deepfake Generation

- Adapt MARLIN for Facial Video Generation
- Evaluate it in LRS2 dataset

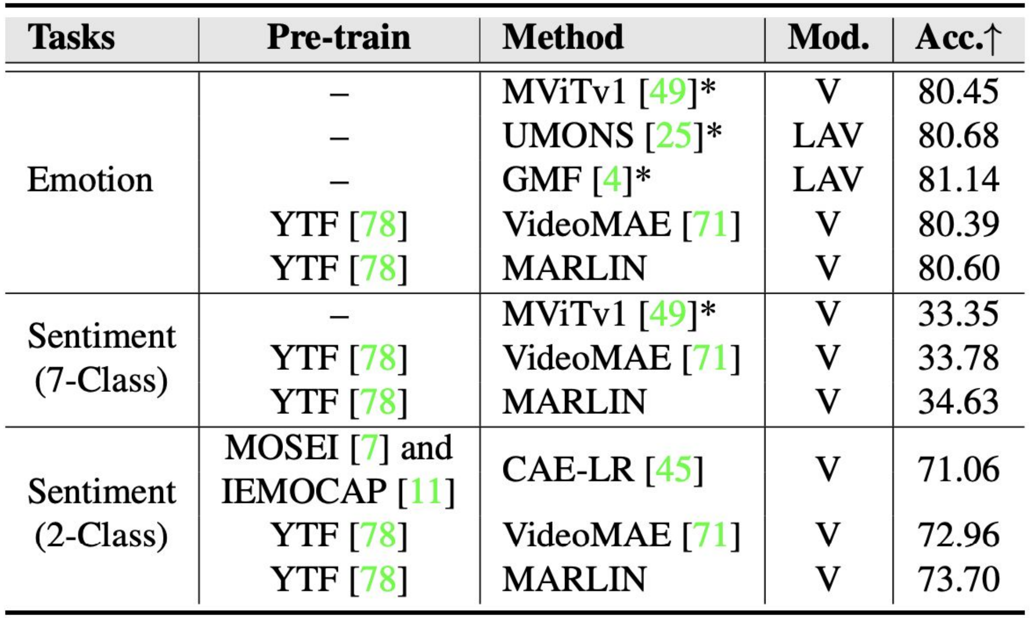


JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Results

Facial Expression/Sentiment Recognition

- Facial Expression Recognition
- Facial Sentiment Recognition
- Evaluated on MOSEI dataset



Tasks	Pre-train	Method	Mod.	Acc.↑
Emotion	–	MViTv1 [49]*	V	80.45
	–	UMONS [25]*	LAV	80.68
	–	GMF [4]*	LAV	81.14
	YTF [78]	VideoMAE [71]	V	80.39
	YTF [78]	MARLIN	V	80.60
Sentiment (7-Class)	–	MViTv1 [49]*	V	33.35
	YTF [78]	VideoMAE [71]	V	33.78
	YTF [78]	MARLIN	V	34.63
Sentiment (2-Class)	MOSEI [7] and IEMOCAP [11]	CAE-LR [45]	V	71.06
	YTF [78]	VideoMAE [71]	V	72.96
	YTF [78]	MARLIN	V	73.70

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

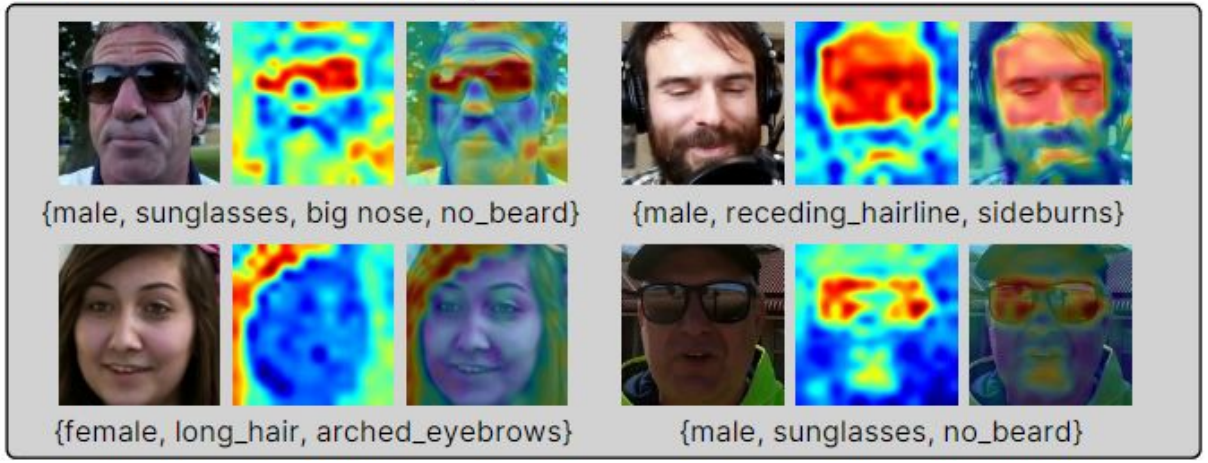
Results

Facial Expression/Sentiment Recognition

- Facial Attribute Recognition
- Evaluated on CelebV-HQ dataset

- Grad cam visualization for facial attributes recognition

Facial Attribute Recognition

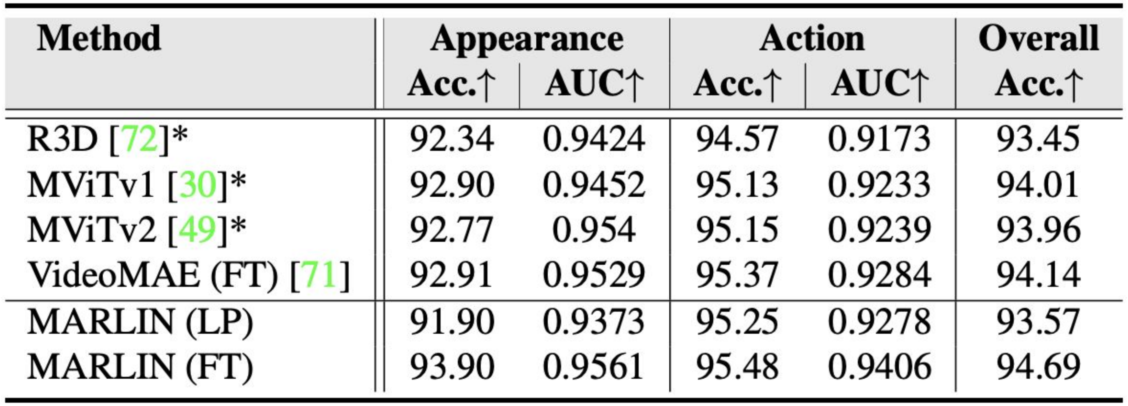


{male, sunglasses, big nose, no_beard}

{male, receding_hairline, sideburns}

{female, long_hair, arched_eyebrows}

{male, sunglasses, no_beard}

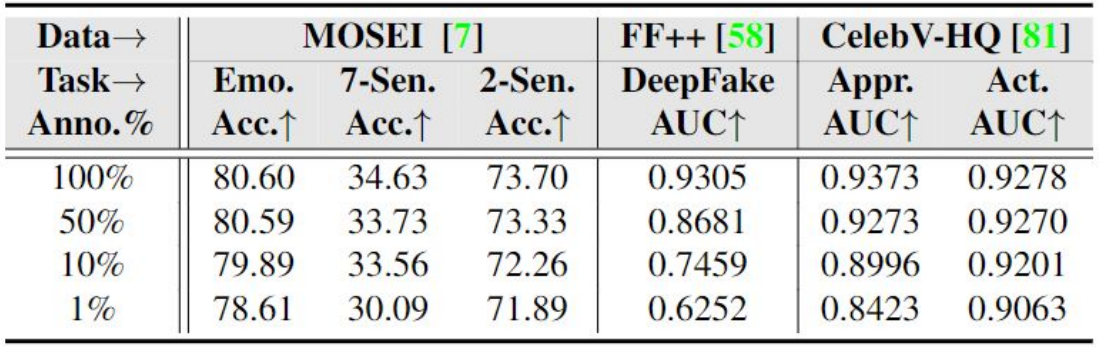


Method	Appearance		Action		Overall
	Acc.↑	AUC↑	Acc.↑	AUC↑	Acc.↑
R3D [72]*	92.34	0.9424	94.57	0.9173	93.45
MViTv1 [30]*	92.90	0.9452	95.13	0.9233	94.01
MViTv2 [49]*	92.77	0.954	95.15	0.9239	93.96
VideoMAE (FT) [71]	92.91	0.9529	95.37	0.9284	94.14
MARLIN (LP)	91.90	0.9373	95.25	0.9278	93.57
MARLIN (FT)	93.90	0.9561	95.48	0.9406	94.69

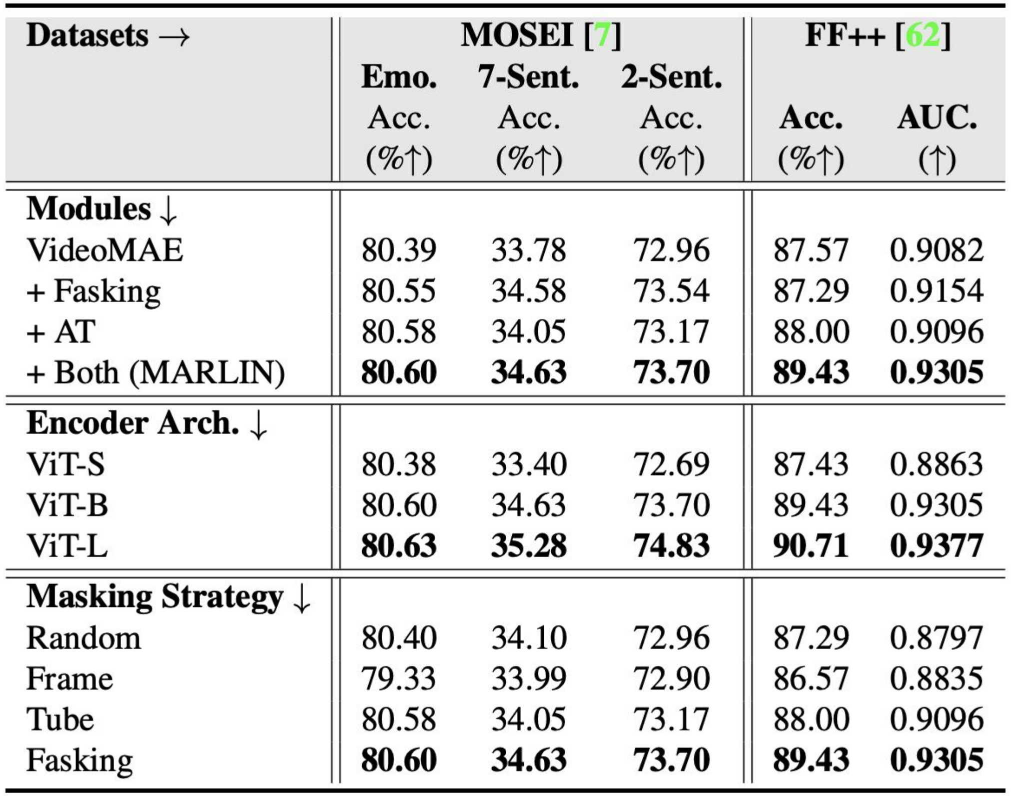
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Ablation Studies

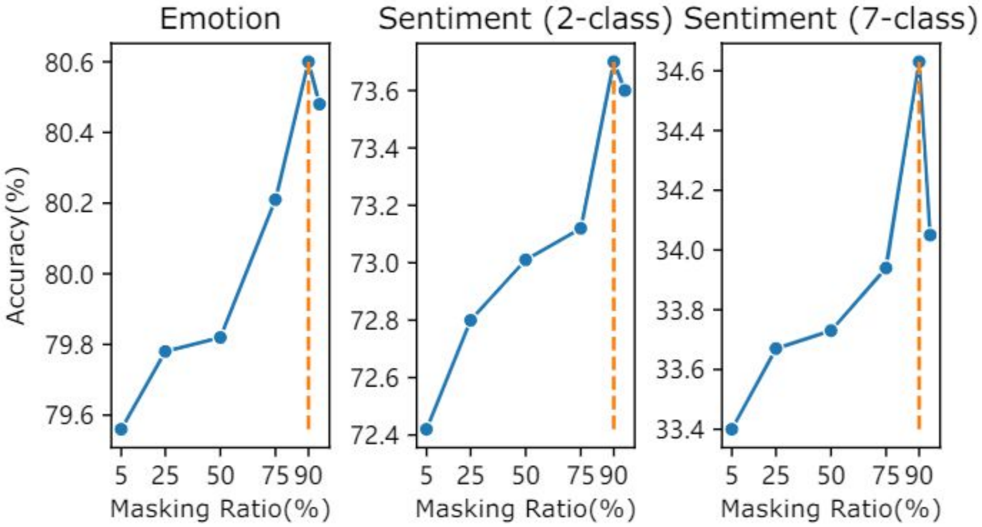
- Contribution of different modules
- Encoder architecture
- Masking strategy



Data→	MOSEI [7]			FF++ [58]	CelebV-HQ [81]	
Task→	Emo.	7-Sen.	2-Sen.	DeepFake	Appr.	Act.
Anno.%	Acc.↑	Acc.↑	Acc.↑	AUC↑	AUC↑	AUC↑
100%	80.60	34.63	73.70	0.9305	0.9373	0.9278
50%	80.59	33.73	73.33	0.8681	0.9273	0.9270
10%	79.89	33.56	72.26	0.7459	0.8996	0.9201
1%	78.61	30.09	71.89	0.6252	0.8423	0.9063



Datasets →	MOSEI [7]			FF++ [62]	
	Emo. Acc. (%↑)	**7-Sent.** Acc. (%↑)	**2-Sent.** Acc. (%↑)	**Acc.** (%↑)	**AUC.** (↑)
Modules ↓					
VideoMAE	80.39	33.78	72.96	87.57	0.9082
+ Fasking	80.55	34.58	73.54	87.29	0.9154
+ AT	80.58	34.05	73.17	88.00	0.9096
+ Both (MARLIN)	**80.60**	**34.63**	**73.70**	**89.43**	**0.9305**
Encoder Arch. ↓					
ViT-S	80.38	33.40	72.69	87.43	0.8863
ViT-B	80.60	34.63	73.70	89.43	0.9305
ViT-L	**80.63**	**35.28**	**74.83**	**90.71**	**0.9377**
Masking Strategy ↓					
Random	80.40	34.10	72.96	87.29	0.8797
Frame	79.33	33.99	72.90	86.57	0.8835
Tube	80.58	34.05	73.17	88.00	0.9096
Fasking	**80.60**	**34.63**	**73.70**	**89.43**	**0.9305**



JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank You

Contact: zhixi.cai@monash.edu

Github: https://github.com/ControlNet/MARLIN