MARLIN: Masked Autoencoder for facial video Representation LearnINg

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

Contributions

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

2. We propose ***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. Through thorough analysis, we demonstrate that MARLIN learns rich, consistent, and versatile facial representations, performing well across various tasks such as ***Facial Attribute Recognition***, ***Lip Synchronization***, and even in ***few shot*** settings.

Related Works

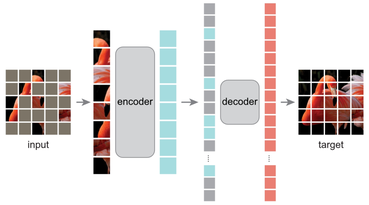


Fig. 1. Masked Autoencoder (MAE) [1]

Reference

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, PiotrDoll´ar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16000–16009, 2022.

[2] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, In Advances in Neural Information Processing Systems, 2022.

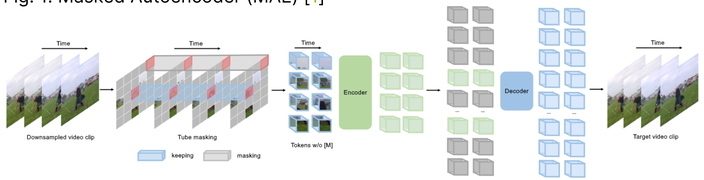


Fig. 2. VideoMAE [2]

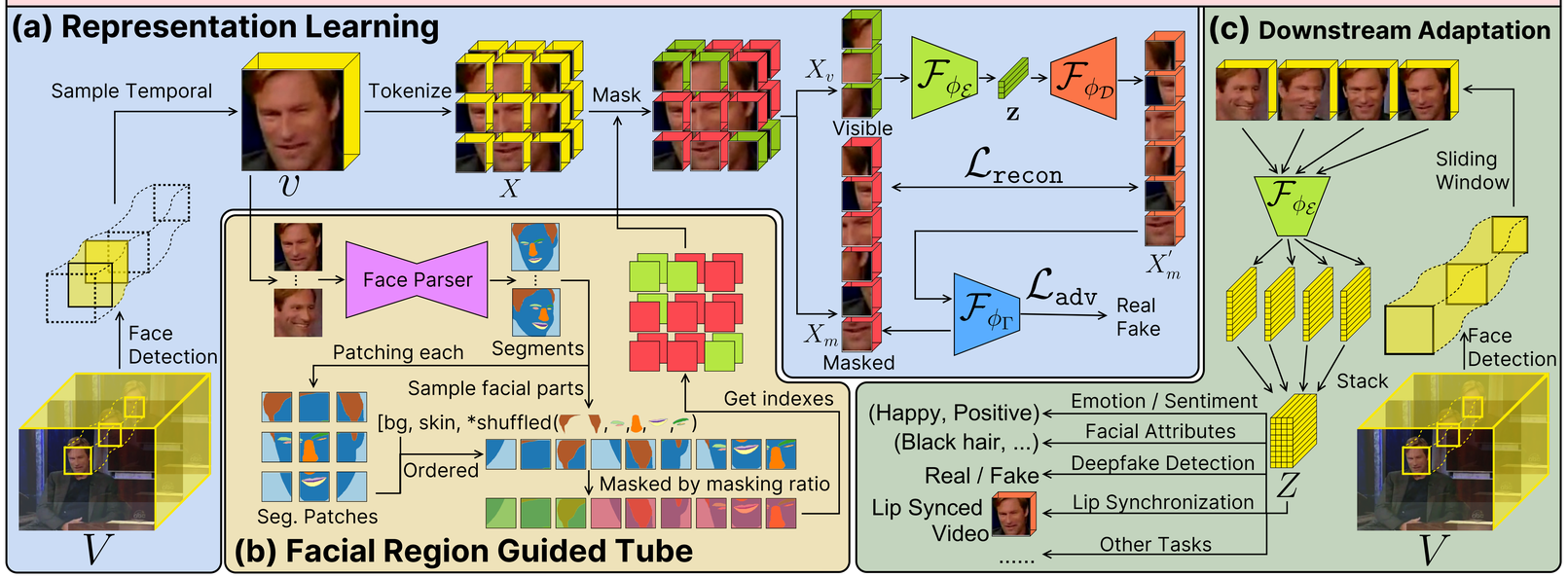
Contact

zhixi.cai@monash.edu
shreya.ghosh@curtin.edu.au
kalin.stefanov@monash.edu
abhinav@iitrpr.ac.in
jianfei.cai@monash.edu
hamid.rezatofighi@monash.edu
gholamreza.haffari@monash.edu
munawar.hayat@monash.edu

Arxiv

GitHub

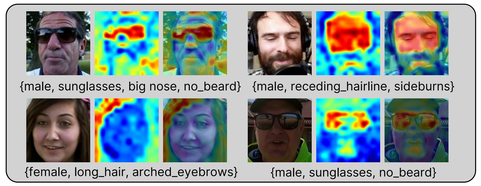
Method



Results

Facial Attribute Recognition

Method	Appearance Acc.↑	Appearance AUC↑	Action Acc.↑	Action AUC↑	Overall 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



Lip Syncronization

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

Few-Shot Adaptation

Data→ / Task→ / Anno.%	MOSEI [7] Emo. Acc.↑	MOSEI [7] 7-Sen. Acc.↑	MOSEI [7] 2-Sen. Acc.↑	FF++ [62] DeepFake AUC↑	CelebV-HQ [85] Appr. AUC↑	CelebV-HQ [85] Act. 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