MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition

Licai Sun[1&2], Zheng Lian[2], Bin Liu[1&2], Jianhua Tao[3]

[1]UCAS [2]CASIA [3]Tsinghua University

University of Chinese Academy of Sciences · CASIA · Tsinghua University

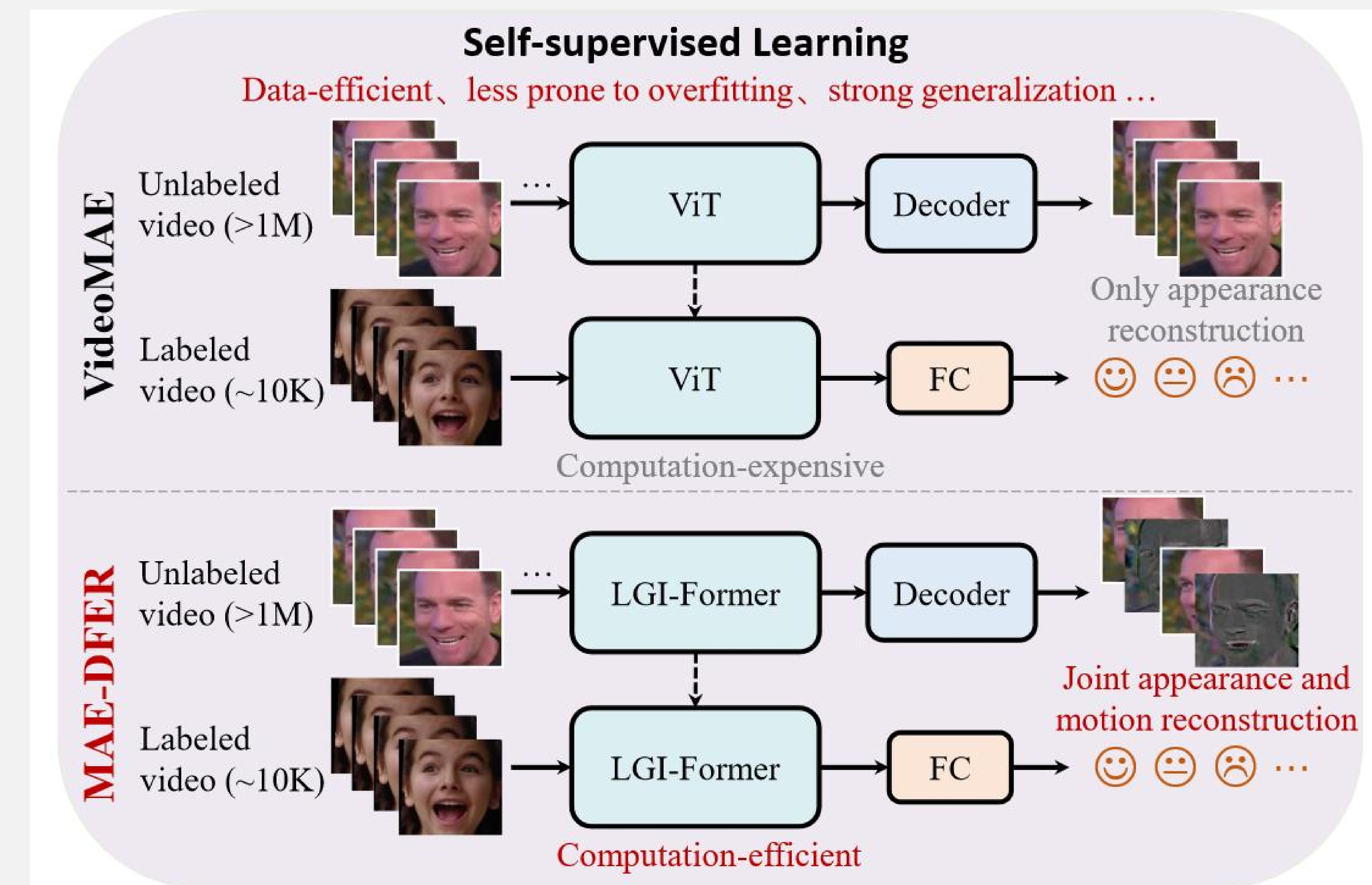
ACM multimedia · Ottawa, Canada · OCT 29 – NOV 3, 2023

Motivation: Supervised Dillema in DFER

- Current efforts in DFER focus on developing **various deep supervised** models, **but only** achieving ***incremental*** progress due to the ***longstanding lack*** of large-scale high-quality datasets.
- Due to the **ambiguity** and **subjectivity** in facial expression perception, acquiring large-scale high-quality DFER samples is pretty ***time-consuming*** and ***labor-intensive***.

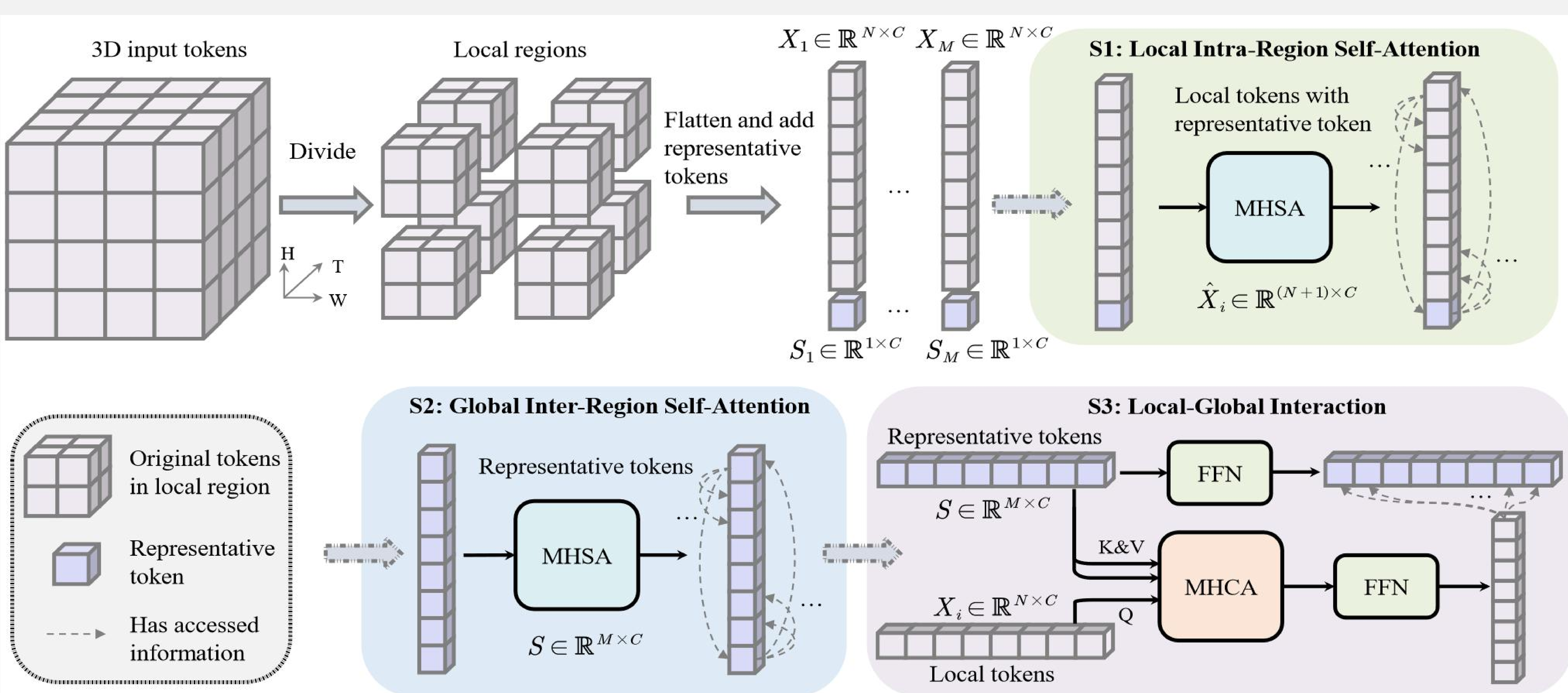
MAE-DFER: Overview

Self-supervised Pre-training + Downstream Fine-tuning

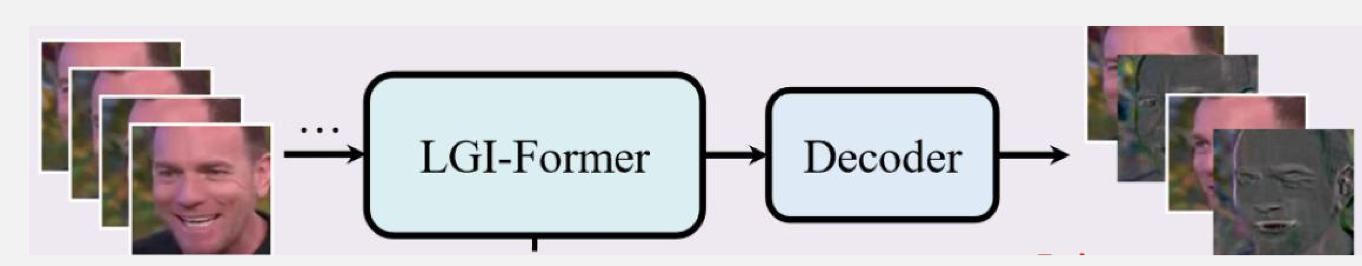


MAE-DFER: Details

Key Module 1: *Efficient* Local-Global Interaction Transfomer (LGI-Former)



Key Module 2: *Joint* Masked Appearance and Motion Modeling



$$\mathcal{L}_{\text{MAE-DFER}} = \lambda \cdot \text{MSE}(\Phi_d(\Phi_e(\mathbf{X}\odot\mathbf{M})), \mathbf{V}_a \odot \Psi(1-\mathbf{M})) + (1-\lambda)\cdot \text{MSE}(\Phi_d(\Phi_e(\mathbf{X}\odot\mathbf{M})), \mathbf{V}_m \odot \Psi(1-\mathbf{M}))$$

Stage 1:

$$\hat{\mathrm{X}}_i = \mathrm{Concat}(\mathrm{S}_i, \mathrm{X}_i)$$
$$\hat{\mathrm{X}}_i = \mathrm{MHSA}(\mathrm{LN}(\hat{\mathrm{X}}_i)) + \hat{\mathrm{X}}_i$$

Stage 2:

$$\mathrm{S} = \mathrm{Concat}(\mathrm{S}_1, \ldots, \mathrm{S}_M)$$
$$\mathrm{S} = \mathrm{MHSA}(\mathrm{LN}(\mathrm{S})) + \mathrm{S}$$

Stage 3:

$$\mathrm{X}_i = \mathrm{MHCA}(\mathrm{LN}(\mathrm{X}_i), \mathrm{LN}(\mathrm{S})) + \mathrm{X}_i$$
$$\mathrm{X}_i = \mathrm{FFN}(\mathrm{LN}(\mathrm{X}_i)) + \mathrm{X}_i$$
$$\mathrm{S} = \mathrm{FFN}(\mathrm{LN}(\mathrm{S})) + \mathrm{S}$$

Complexity Comparison:

Ours: $O((\frac{1}{M} + \frac{1}{N^2} + \frac{1}{N})K^2)$

ViT: $O(K^2)$ $M \ll K$ $N \ll K$

Quantitative Results

- **MAE-DFER** consistently **outperforms** the previous best **supervised** methods by ***significant*** margins (**+5~8% UAR** on three in-the-wild datasets and **+7~12% WAR** on three lab-controlled datasets)

DFEW: +6.30% UAR

Method	#Params (M)	FLOPs (G)	UAR	WAR
Supervised methods				
C3D [55]	78	39	42.74	53.54
R(2+1)D-18 [56]	33	42	42.79	53.22
3D ResNet-18 [21]	33	8	46.52	58.27
EC-STFL [25]	-	8	45.35	56.51
ResNet-18+LSTM [69]	-	8	51.32	63.85
ResNet-18+GRU [69]	-	8	51.68	64.02
Former-DFER [69]	18	9	53.69	65.70
CEFLNet [33]	13	-	51.14	65.35
EST [35]	43	-	53.43	65.85
STT [37]	-	-	54.58	66.65
NR-DFERNet [30]	-	6	54.21	68.19
DPCNet [65]	51	10	57.11	66.32
IAL [29]	19	10	55.71	69.24
M3DFEL [60]	-	2	56.10	69.25
Self-supervised methods				
VideoMAE [54]	86	81	58.49	70.61
VideoMAE [54] †	86	81	**63.60**	**74.60**
MAE-DFER (ours)	85	50	63.41	74.43

MAFW: +8.34% UAR

Method	#Params (M)	FLOPs (G)	UAR	WAR
Supervised methods				
ResNet-18 [23]	11	-	25.58	36.65
ViT [13]	-	-	32.36	45.04
C3D [55]	78	39	31.17	42.25
ResNet-18+LSTM [32]	-	-	28.08	39.38
ViT+LSTM [32]	-	-	32.67	45.56
C3D+LSTM [32]	-	-	29.75	43.76
Former-DFER [69]	18	9	31.16	43.27
T-ESFL [32]	-	-	33.28	48.18
Self-supervised methods				
VideoMAE [54]	86	81	38.43	51.74
VideoMAE [54] †	86	81	40.87	53.51
MAE-DFER (ours)	85	50	**41.62**	**54.31**

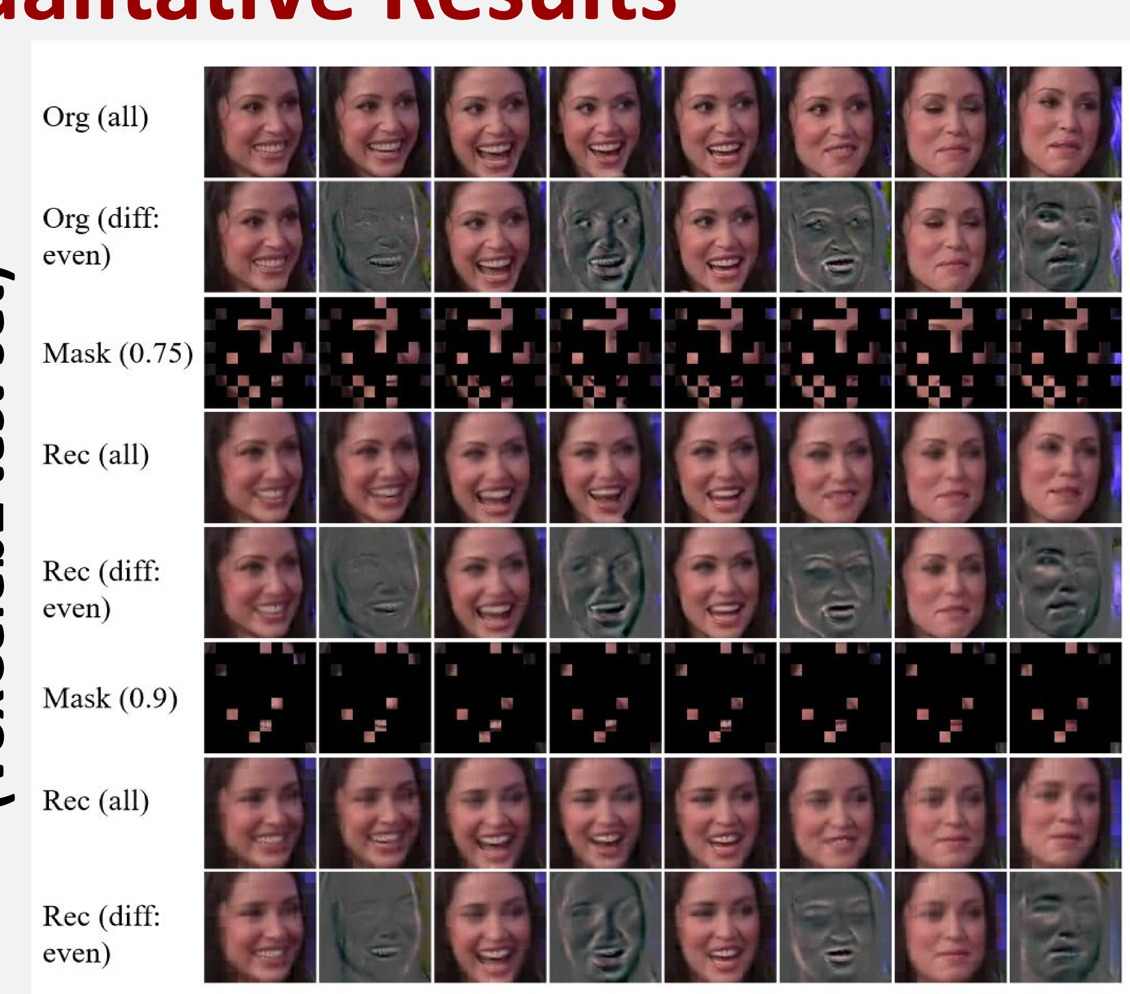
CREMA-D: +10.58% WAR
RAVDESS: +12.57% WAR
eNTERFACE05: +7.02% WAR

CREMA-D				RAVDESS				eNTERFACE05		
Method	Modality	UAR	WAR	Method	Modality	UAR	WAR	Method	UAR	WAR
VO-LSTM [19]	Video	-	66.80	VO-LSTM [19]	Video	-	60.50	3DCNN [4]	-	41.05
Goncalves et al. [20]	Video	-	62.20	3D ResNeXt-50 [50]	Video	-	62.99	3DCNN-DAP [4]	-	41.36
Lei et al. [28]	Video	64.68	64.76	AV-LSTM [19]	Video+Audio	-	65.80	STA-FER [43]	-	42.98
AV-LSTM [19]	Video+Audio	-	72.90	AV-Gating [19]	Video+Audio	-	67.70	TSA-FER [42]	-	43.72
AV-Gating [19]	Video+Audio	-	74.00	MCBP [50]	Video+Audio	-	71.32	C-LSTM [40]	-	45.29
MulT Base [57]	Video+Audio	-	68.87	MMTM [50]	Video+Audio	-	73.12	EC-LSTM [41]	-	49.26
MulT Large [57]	Video+Audio	-	70.22	MSAF [50]	Video+Audio	-	74.86	FAN [39]	-	51.44
Goncalves et al. [20]	Video+Audio	-	77.30	CFN-SR [18]	Video+Audio	-	75.76	Graph-Tran [68]	-	54.62
MAE-DFER (ours)	Video	**77.33**	**77.38**	MAE-DFER (ours)	Video	**75.91**	**75.56**	MAE-DFER (ours)	**61.67**	**61.64**

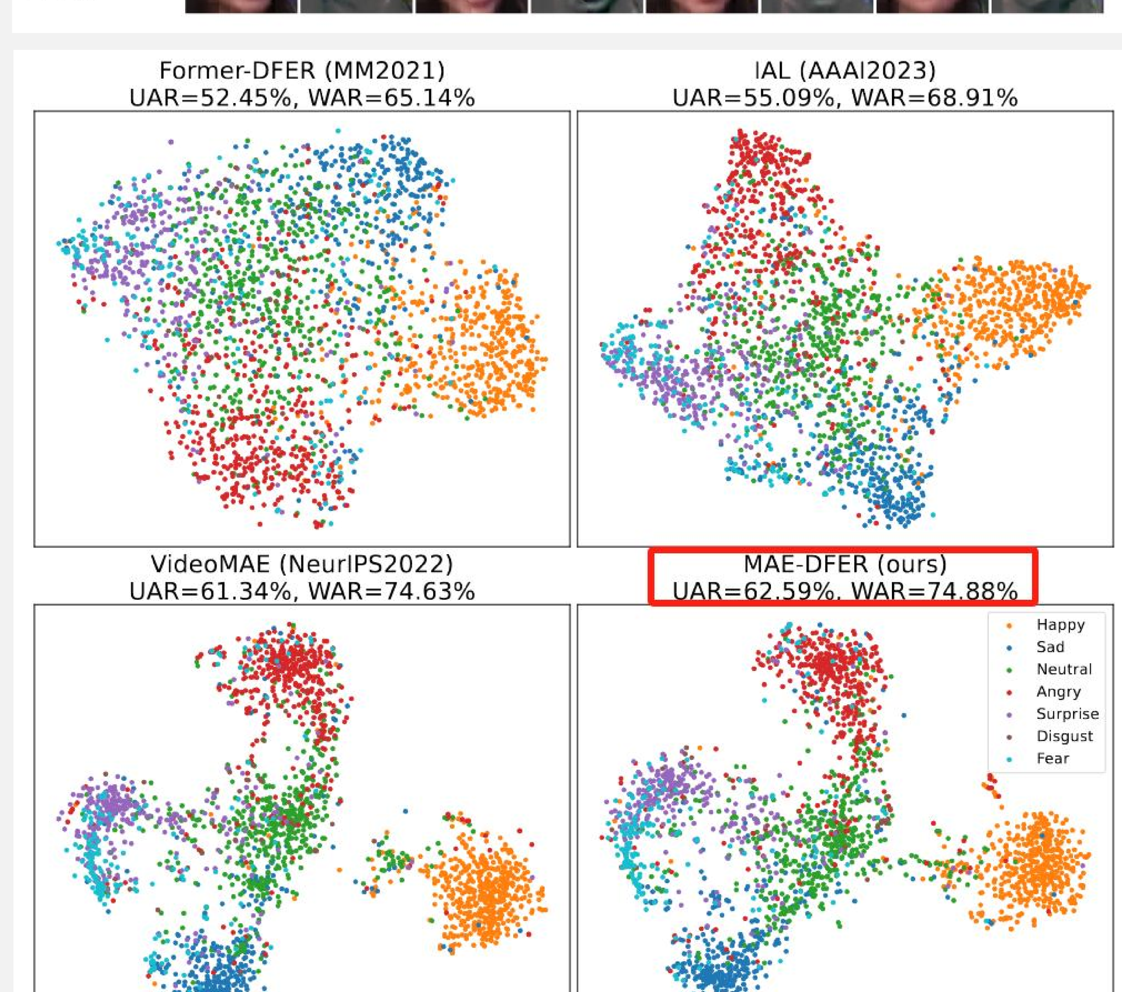
- **MAE-DFER** has ***comparable*** or even ***better*** results than **VideoMAE**, while **largely** reducing the computational cost (**about 38% FLOPs**)

Qualitative Results

Reconstruction Visualization (VoxCeleb2 *test* set)



Embedding Space using t-SNE (DFEW fd1)



Paper:

Code:

Conclusion: **MAE-DFER**, as an early attempt to leverage **large-scale self-supervised pre-training** on unlabeled facial videos, ***has paved a new way*** for the advancement of DFER.