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

Licai Sun
sunlicai2019@ia.ac.cn
School of Artificial Intelligence, University of Chinese Academy of Sciences
Institute of Automation, Chinese Academy of Sciences
Beijing, China

Zheng Lian
lianzheng2016@ia.ac.cn
Institute of Automation, Chinese Academy of Sciences
Beijing, China

Bin Liu
liubin@nlpr.ia.ac.cn
Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China

Jianhua Tao
jhtao@tsinghua.edu.cn
Department of Automation, Tsinghua University
Beijing National Research Center for Information Science and Technology, Tsinghua University
Beijing, China

ABSTRACT

Dynamic facial expression recognition (DFER) is essential to the development of intelligent and empathetic machines. Prior efforts in this field mainly fall into supervised learning paradigm, which is severely restricted by the limited labeled data in existing datasets. Inspired by recent unprecedented success of masked autoencoders (e.g., VideoMAE), this paper proposes MAE-DFER, a novel self-supervised method which leverages large-scale self-supervised pre-training on abundant unlabeled data to largely advance the development of DFER. Since the vanilla Vision Transformer (ViT) employed in VideoMAE requires substantial computation during fine-tuning, MAE-DFER develops an efficient local-global interaction Transformer (LGI-Former) as the encoder. Moreover, in addition to the standalone appearance content reconstruction in VideoMAE, MAE-DFER also introduces explicit temporal facial motion modeling to encourage LGI-Former to excavate both static appearance and dynamic motion information. Extensive experiments on six datasets show that MAE-DFER consistently outperforms state-of-the-art supervised methods by significant margins (e.g., +6.30% UAR on DFEW and +8.34% UAR on MAFW), verifying that it can learn powerful dynamic facial representations via large-scale self-supervised pre-training. Besides, it has comparable or even better performance than VideoMAE, while largely reducing the computational cost (about 38% FLOPs). We believe MAE-DFER has paved a new way for the advancement of DFER and can inspire more relevant research in this field and even other related tasks. Codes and models are publicly available at https://github.com/sunlicai/MAE-DFER.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612365

KEYWORDS

Dynamic facial expression recognition, masked autoencoder

ACM Reference Format:
Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3612365

1 INTRODUCTION

Facial expressions, as an important aspect of nonverbal communication, play a significant role in interpersonal interactions [10]. In the past two decades, automatic facial expression recognition (FER) has drawn widespread attention due to its crucial role in developing intelligent and empathetic machines that can interact with humans in a natural and intuitive way [11, 44, 45]. FER also has a wide spectrum of practical applications in areas such as healthcare [3], education [66], and entertainment [53]. According to the input data type, FER can be divided into two categories, i.e., static FER (SFER) and dynamic FER (DFER) [31]. SFER takes static facial images as input, while DFER aims to recognize expressions in dynamic image sequences or videos. Since SFER overlooks the critical temporal information for the interpretation of facial expressions, this paper mainly focuses on DFER.

DFER is dominated by the supervised learning paradigm. Researchers have developed various deep neural networks for this task, including 2D/3D convolutional neural networks (CNN) [15, 25, 27], recurrent neural networks (RNN) [14, 52, 65], and more advanced Transformer-based architectures [29, 35, 37, 61, 69]. Although supervised methods have achieved remarkable success, the limited training samples in existing DFER datasets (typically around 10K, which is much smaller than those in other research areas such as general image/video classification and face recognition, see details in Table 1) severely restrict their further advancement (e.g., training large video Transformers). A straightforward idea to address this issue is to increase the dataset scale. However, collecting and

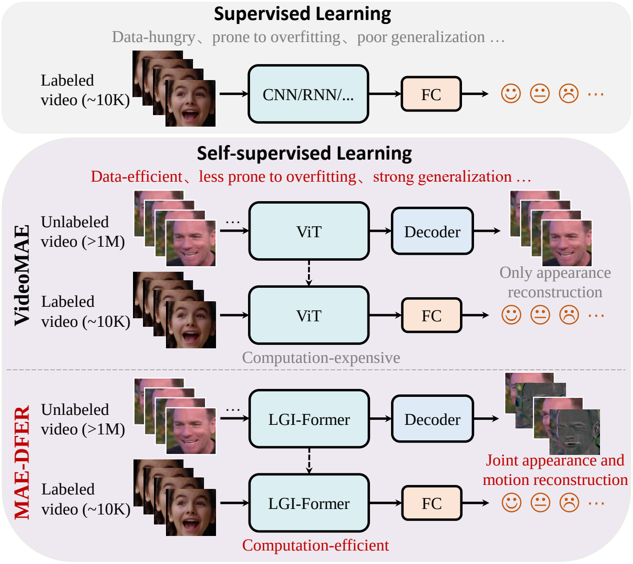

Figure 1: An overview of the proposed MAE-DFER.

annotating large-scale high-quality DFER datasets is pretty time-consuming and labor-intensive, which is mainly due to the sparsity of dynamic facial expressions in videos and the ambiguity and subjectivity in facial expression perception [25, 31, 64]. Considering that there are massive unlabeled facial videos on the Internet, a natural question arises in the mind: *can we exploit them to fully unleash the power of deep neural networks for better DFER?*

The recent progress of self-supervised learning in many deep learning fields [1, 12, 22] indicates that there is a positive answer. Notably, masked autoencoder (MAE) [22] in computer vision develops an asymmetric encoder-decoder architecture for masked image modeling. It successfully pre-trains the vanilla Vision Transformer (ViT) [13] in an end-to-end manner and outperforms the supervised baselines in many vision tasks. Subsequently, VideoMAE [54] extends MAE to the video domain and also achieves impressive results on lots of general video datasets. Motivated by this line of research, we present MAE-DFER (Fig. 1), a novel self-supervised method based on VideoMAE which leverages large-scale self-supervised pre-training on abundant unlabeled facial video data to promote the advancement of DFER. Although VideoMAE has made remarkable success in self-supervised video pre-training, we notice that it still has two main drawbacks: 1) The vanilla ViT encoder employed in VideoMAE requires substantial computation during fine-tuning due to the quadratic scaling cost of global space-time self-attention, which is unaffordable in many resource-constrained scenarios. 2) It only reconstructs video appearance contents during pre-training, thus might be insufficient to model temporal facial motion information which is also crucial to DFER.

To tackle the above issues in VideoMAE, our MAE-DFER presents two core designs accordingly. For the *first* issue, MAE-DFER develops an *efficient* local-global interaction Transformer (LGI-Former) as the encoder. Different from the global space-time self-attention in ViT, LGI-Former first constrains self-attention in local spatiotemporal regions and then utilizes a small set of learnable *representative tokens* to enable efficient local-global information exchange. Concretely, it decomposes the global space-time self-attention into three stages: local intra-region self-attention, global inter-region self-attention, and local-global interaction. In this way, LGI-Former can efficiently propagate global information to local regions and avoid the expensive computation of global space-time attention. For the *second* issue, MAE-DFER introduces *joint* masked appearance and motion modeling to encourage the model to capture both static facial appearance and dynamic motion information. Specifically, in addition to the original appearance content reconstruction branch, it simply utilizes the frame difference signal as another reconstruction target for explicit temporal facial motion modeling. To verify the effectiveness of MAE-DFER, we perform large-scale self-supervised pre-training on the VoxCeleb2 dataset [9], which has more than 1M unlabeled facial video clips collected from YouTube. Then we fine-tune the pre-trained model on six DFER datasets, including three relatively large in-the-wild datasets (DFEW [25], FERV39k [64], and MAFW [32]) and three small lab-controlled datasets (CREMA-D [5], RAVDESS [36], and eNTERFACE05 [38]). The results show that MAE-DFER significantly outperforms the state-of-the-art supervised methods, indicating that it is capable of learning strong and useful dynamic facial representations for DFER. Moreover, compared with VideoMAE, MAE-DFER largely reduces ~**38%** FLOPs while having comparable or even better performance. The main contributions of this paper are summarized as follows:

- We present a novel self-supervised method, MAE-DFER, as an early attempt to leverage large-scale self-supervised pre-training on abundant unlabeled facial video data to advance the development of DFER.
- MAE-DFER improves VideoMAE by developing an efficient LGI-Former as the encoder and introducing joint masked appearance and motion modeling. With these two core designs, MAE-DFER largely reduces the computational cost while having comparable or even better performance.
- Extensive experiments on six DFER datasets show that our 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), which demonstrates that it can learn powerful dynamic facial representations for DFER via large-scale self-supervised pre-training.

2 RELATED WORK

2.1 Dynamic Facial Expression Recognition

The early studies on DFER primarily focus on designing various local descriptors and only several very small lab-controlled datasets are available for evaluation. With the emergence of deep learning and the proliferation of relatively larger datasets, the research paradigm has undergone a transformative shift towards training deep neural networks in an end-to-end fashion. In general, there are three trends. The first trend directly utilizes 3D CNNs (such as C3D [55], 3D ResNet [21], R(2+1)D [56], and P3D [46]) to extract joint spatiotemporal features from raw facial videos [15, 25, 27, 32, 60, 64]. The second trend uses the combination of 2D CNN (e.g., VGG [48] and ResNet [23]) and RNN (e.g., LSTM [24] and GRU [8]) [14, 25, 26, 32, 52, 64, 65]. Recently, with the rise of Transformer

[59], several studies exploit its global dependency modeling ability to augment CNN/RNN for better performance, which forms the third trend [29, 30, 32, 35, 37, 69]. For instance, Former-DFER [69] employs a Transformer-enhanced ResNet-18 for spatial feature extraction and another Transformer for temporal information aggregation. STT [37] improves Former-DFER by introducing factorized spatial and temporal attention for joint spatiotemporal feature learning. IAL [29] further introduces the global convolution-attention block and intensity-ware loss to deal with expressions with different intensities. However, all the above methods fall into the supervised learning paradigm, which is thus restricted by the limited training samples in existing DFER datasets. Unlike them, this paper proposes a self-supervised method that can learn powerful representations from massive unlabeled facial video data and achieve significant improvement over them.

2.2 Masked Autoencoders

Masked autoencoders (MAEs), as the representative of generative self-supervised learning, have recently achieved unprecedented success in many deep learning fields [67]. They are mainly inspired by the progress of masked language modeling (e.g., BERT [12] and GPT [47]) in natural language processing and typically adopt a mask-then-predict strategy to pre-train the vanilla ViT. Notably, iGPT [7] follows GPT to auto-regressively predict pixels and makes the first successful attempt. BEiT [2] follows BERT and adopts a two-stage training pipeline, i.e., first utilizing an off-the-shelf tokenizer to generate discrete visual tokens and then performing masked-then-predict training. MAE [22] improves BEiT by designing an asymmetric encoder-decoder architecture to enable efficient end-to-end pre-training. After that, many studies adopt the architecture of MAE to perform self-supervised pre-training on various tasks. For instance, VideoMAE [54] and its concurrent work MAE-ST [17] extends MAE to the video domain and achieve impressive results on lots of video benchmarks. Our proposed MAE-DFER is inspired by VideoMAE and it develops two core designs to facilitate effective and efficient representation learning for DFER.

3 METHOD

3.1 Revisiting VideoMAE

VideoMAE [54] is a simple extension of MAE [22] in the video domain. It basically follows the asymmetric encoder-decoder architecture of MAE for self-supervised video pre-training. The main difference is that a much higher masking ratio (i.e., 90% vs. 75%) and tube masking strategy (instead of random masking) are adopted, considering that large temporal redundancy and high temporal correlation in videos [54]. In specific, VideoMAE mainly consists of four modules: cube embedding, tube masking, a high-capacity encoder Φ_e (i.e., the vanilla ViT), and a lightweight decoder Φ_d. Given a raw video $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$, VideoMAE first utilizes cube embedding with a cube size of $2 \times 16 \times 16$ to transform $\mathbf{V}$ into a sequence of tokens $\mathbf{X} \in \mathbb{R}^{K \times C}$, where $K = \frac{T}{2} \cdot \frac{H}{16} \cdot \frac{W}{16}$ and C is the channel size. Then the tube masking module generates a mask $\mathbf{M} \in \{0, 1\}^K$ with a masking ratio of $\rho = 90\%$ and the high-capacity encoder Φ_e only takes the unmasked tokens $\mathbf{X} \odot \mathbf{M} \in \mathbb{R}^{L \times C}$ ($L = (1 - \rho)K$) as input and simply process them with global space-time self-attention. Subsequently, the lightweight decoder Φ_d combines the encoded visible tokens with the learnable mask tokens (with a size of ρK) to reconstruct the raw video data. Finally, the mean square error between the original and reconstructed video in the masked positions are calculated to optimize the whole model. The above process can be generally formulated as follows:

$$\mathcal{L}_{\text{VideoMAE}} = \text{MSE}(\Phi_d(\Phi_e(\mathbf{X} \odot \mathbf{M})), \mathbf{V} \odot \Psi(1 - \mathbf{M})) \tag{1}$$

where Ψ is a function used to obtain masked positions in the pixel space. In downstream tasks, the lightweight decoder Φ_d is discarded and only the high-capacity ViT encoder Φ_e will be fine-tuned.

3.2 MAE-DFER: Overview

Although VideoMAE has made great success in self-supervised video pre-training, it still faces two major challenges. First, it only focuses on reconstructing raw appearance contents in the video, which thus lacks explicit temporal motion modeling and might not be sufficient to model temporal facial motion information. Second, although it enjoys high efficiency during *pre-training* through an asymmetric encoder-decoder architecture (i.e., dropping a large proportion of masked tokens to save computation), the computational cost of global space-time self-attention in the vanilla ViT is still extremely expensive during downstream *fine-tuning* since it cannot drop input tokens at this stage. To tackle these issues, as shown in Fig. 1, we propose MAE-DFER, a new self-supervised framework for DFER. For the first issue, MAE-DFER introduces joint masked appearance and motion modeling to encourage the model to excavate both static appearance and dynamic motion information (Section 3.3). For the second issue, it employs a novel Local-Global Interaction Transformer (LGI-Former) as the encoder to largely reduce the computational cost of ViT during downstream fine-tuning (Section 3.4).

3.3 MAE-DFER: Joint Masked Appearance and Motion Modeling

Temporal motion information matters for DFER (e.g., the gradual appearance and disappearance of a smile may convey totally different emotions). To explicitly incorporate this information in self-supervised pre-training, our MAE-DFER adds an additional temporal motion reconstruction branch in parallel with the original appearance reconstruction branch in VideoMAE to achieve *joint* facial appearance and motion structure learning. Specifically, we simply calculate the frame difference signal as the temporal motion target given that its computation is very cheap and it has shown effectiveness in video action recognition [49, 62, 63]. To ensure that the computational cost during pre-training similar to VideoMAE, we share the decoder backbone for appearance and motion branches and only use two different linear heads to predict their targets. Besides, the decoder only outputs appearance predictions in the odd frames and motion predictions in the remaining even frames. Finally, the total loss is the weighted sum of mean square errors in two branches:

$$\begin{aligned}\mathcal{L}_{\text{MAE-DFER}} = \lambda \cdot \text{MSE}(\Phi_d(\Phi_e(\mathbf{X} \odot \mathbf{M})), \mathbf{V}_a \odot \Psi(1 - \mathbf{M})) + \\ (1 - \lambda) \cdot \text{MSE}(\Phi_d(\Phi_e(\mathbf{X} \odot \mathbf{M})), \mathbf{V}_m \odot \Psi(1 - \mathbf{M}))\end{aligned} \tag{2}$$

where $\mathbf{V}_a = \mathbf{V}[0:T:2]$ is the appearance target, $\mathbf{V}_m = \mathbf{V}[1:T:2] - \mathbf{V}[0:T:2]$ is the motion target, λ is a hyperparameter to balance the contribution of two branches and we empirically set it to 0.5.

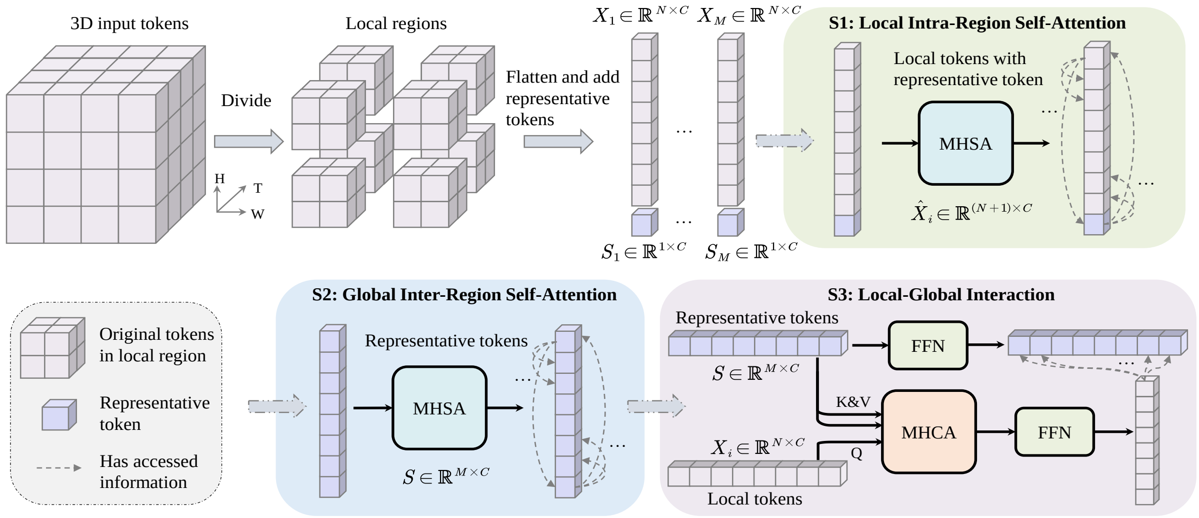
3D input tokens
Divide
H
T
W
Local regions
Flatten and add representative tokens
$X_1 \in \mathbb{R}^{N\times C}$
$X_M \in \mathbb{R}^{N\times C}$
$S_1 \in \mathbb{R}^{1\times C}$
$S_M \in \mathbb{R}^{1\times C}$
S1: Local Intra-Region Self-Attention
Local tokens with representative token
MHSA
$\hat{X}_i \in \mathbb{R}^{(N+1)\times C}$
Original tokens in local region
Representative token
Has accessed information
S2: Global Inter-Region Self-Attention
Representative tokens
MHSA
$S \in \mathbb{R}^{M\times C}$
S3: Local-Global Interaction
Representative tokens
$S \in \mathbb{R}^{M\times C}$
FFN
K&V
MHCA
FFN
$X_i \in \mathbb{R}^{N\times C}$
Q
Local tokens

Figure 2: The illustration of LGI-Former. For simplicity, we only present the information flow in one block, which mainly consists of three stages: 1) local intra-region self-attention, 2) global inter-region self-attention, and 3) local-global interaction.

3.4 MAE-DFER: Efficient LGI-Former

The architecture of LGI-Former is illustrated in Fig. 2. Unlike the global space-time self-attention adopted in the vanilla ViT, LGI-Former constrains self-attention in local spatiotemporal regions to save computation. However, simply stacking multiple local self-attention layers does not permit inter-region information exchange. Inspired by [16] and [51], the core idea of LGI-Former is to introduce a small set of *representative tokens* to local regions. On the one hand, these tokens take charge of summarizing critical information in local regions. On the other hand, they allow for long-range dependencies modeling between different regions and enable efficient local-global information exchange. Thanks to the introduction of representative tokens, the expensive global space-time self-attention can be decomposed into three stages with much cheaper computation: 1) local intra-region self-attention, 2) global inter-region self-attention, and 3) local-global interaction. In the following, for simplicity, we only describe the above three stages during fine-tuning. The process during pre-training is similar as MAE-DFER follows VideoMAE to adopt the tube masking strategy and applies the same masking ratio to each local region to ensure that all regions have an equal number of visible tokens.

Local Intra-Region Self-Attention. For convenience, we first reshape the input sequence $\mathbf{X} \in \mathbb{R}^{K\times C}$ (after cube embedding) to 3D tokens $\mathbf{X} \in \mathbb{R}^{\frac{T}{2}\times\frac{H}{16}\times\frac{W}{16}\times C}$ and divide it into non-overlapped local spatiotemporal regions with an equal size of $t \times h \times w$ as shown in Fig. 2. In each region, apart from the original tokens, we also add a learnable representative token. The local intra-region self-attention then operates on their concatenation to simultaneously promote fine-grained local feature learning and enable local information aggregation into the representative token. Assume that the original local tokens and the associated representative token in the ith region is $\mathbf{X}_i \in \mathbb{R}^{N\times C}$ and $\mathbf{S}_i \in \mathbb{R}^{1\times C}$ respectively ($N = thw$, $i \in \{1, 2, ..., M\}$, and $M = \frac{K}{N}$ is the number of representative tokens), the formulation of local intra-region self-attention is given as follows:

$$\hat{\mathbf{X}}_i = \mathrm{Concat}(\mathbf{S}_i, \mathbf{X}_i) \tag{3}$$

$$\hat{\mathbf{X}}_i = \mathrm{MHSA}(\mathrm{LN}(\hat{\mathbf{X}}_i)) + \hat{\mathbf{X}}_i \tag{4}$$

where $\hat{\mathbf{X}}_i \in \mathbb{R}^{(N+1)\times C}$, MHSA is the multi-head self-attention in the vanilla ViT, and LN stands for layer normalization. In particular, the calculation of MHSA is formulated as follows:

$$\mathrm{MHSA}(\mathbf{X}) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_h)\mathbf{W}^O \tag{5}$$

$$\mathrm{head}_j = \mathrm{Attention}(\mathbf{X}\mathbf{W}_j^Q, \mathbf{X}\mathbf{W}_j^K, \mathbf{X}\mathbf{W}_j^V) \tag{6}$$

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V} \tag{7}$$

where $\mathbf{W}_j^* \in \mathbb{R}^{C\times d}$ ($* \in \{Q, K, V\}$), $\mathbf{W}^O \in \mathbb{R}^{C\times C}$, h is the number of attention heads, $d = \frac{C}{h}$ is the feature dimension of each head.

Global Inter-Region Self-Attention. After local intra-region self-attention, the representative token has extracted crucial information in each local region and can *represent* the original tokens to perform information exchange between different regions. Since the number of representative tokens is typically small (e.g., 8), the computational cost for inter-region communication can be negligible. Thus, we first aggregate all representative tokens and then simply utilize global inter-region self-attention on them to propagate information between different regions, i.e.,

$$\mathbf{S} = \mathrm{Concat}(\mathbf{S}_1, ..., \mathbf{S}_M) \tag{8}$$

$$\mathbf{S} = \mathrm{MHSA}(\mathrm{LN}(\mathbf{S})) + \mathbf{S} \tag{9}$$

where $\mathbf{S} \in \mathbb{R}^{M\times C}$ is the aggregated representative tokens.

Local-Global Interaction. After information propagation via global inter-region self-attention, the representative token in each local region has been consolidated by useful information from other regions, thus having a global view of the whole input tokens. To enable the original tokens in each local region to access the global information, we further employ cross-attention between local tokens and representative tokens to achieve that goal:

$$\mathbf{X}_i = \text{MHCA}(\text{LN}(\mathbf{X}_i), \text{LN}(\mathbf{S})) + \mathbf{X}_i \tag{10}$$

$$\mathbf{X}_i = \text{FFN}(\text{LN}(\mathbf{X}_i)) + \mathbf{X}_i \tag{11}$$

$$\mathbf{S} = \text{FFN}(\text{LN}(\mathbf{S})) + \mathbf{S} \tag{12}$$

where MHCA is multi-head cross-attention and FFN denotes feed-forward network. Specifically, MHCA has the similar implementation with MHSA except that its query and key/value come from different inputs, i.e.,

$$\text{MHCA}(\mathbf{X}, \mathbf{Y}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W}^O \tag{13}$$

$$\text{head}_h = \text{Attention}(\mathbf{X}\mathbf{W}_j^Q, \mathbf{Y}\mathbf{W}_j^K, \mathbf{Y}\mathbf{W}_j^V) \tag{14}$$

Complexity Analysis. We suppose that the flattened input is $\mathbf{X} \in \mathbb{R}^{K \times C}$, where $K = MN$ is the number of total input tokens, M is the number of local regions and N is the number of original tokens in each region. Since self-attention scales quadratically with the sequence length, the complexity of local intra-region self-attention is $O(M(N+1)^2) \approx O(MN^2) = O(\frac{K^2}{M})$. Similarly, the complexity of global inter-region self-attention is $O(M^2) = O(\frac{K^2}{N^2})$. Moreover, local-global interaction has a complexity of $O(MNM) = O(\frac{K^2}{N})$. Putting them together, the complexity of an LGI-Former block is $O((\frac{1}{M} + \frac{1}{N^2} + \frac{1}{N})K^2)$, while a standard Transformer block in the vanilla ViT has a complexity of $O(K^2)$. In practice, $M \ll K$ and $N \ll K$, thus the computational cost of LGI-Tranformer is largely reduced compared with the vanilla ViT.

4 RESULTS

4.1 Datasets

Pre-training Dataset. We perform self-supervised pre-training on VoxCeleb2 [9]. It has over 1 million video clips of more than 6,000 celebrities, extracted from around 150,000 interview videos on YouTube. It is divided into a development set and a test set. We only use the *development* set for pre-training, which contains 1,092,009 video clips from 145,569 videos.

DFER Datasets. We conduct experiments on 6 datasets, including 3 large *in-the-wild* datasets (i.e., DFEW [25], FERV39k [64], and MAFW [32]) and 3 small *lab-controlled* datasets (i.e., CREMA-D [5], RAVDESS [36], and eNTERFACE05 [38]). Their basic information is summarized in Table 1. Detailed introductions can be found in Appendix A. Following previous studies [25, 32, 64, 69], we report both unweighted average recall (UAR, i.e., the mean class accuracy) and weighted average recall (WAR, i.e., the overall accuracy). For those datasets using cross-validation, we combine the predictions and labels from all folds to calculate the final UAR and WAR.

4.2 Implementation Details

MAE-DFER. For the high-capacity encoder, we adopt the LGI-Former which has 16 blocks and a hidden size of 512. The total number of parameters is 84.9M, which is similar to that (86.2M) of ViT base model. The local region size is set to 2 × 5 × 10 by default. For the lightweight decoder, we follow VideoMAE to adopt four standard Transformer blocks with a hidden size of 384.

Table 1: Basic information of six DFER datasets used in this paper. CV: cross-validation. †: subject-independent setting.

Name	Wild	#Videos	#Classes	Evaluation
DFEW [25]	✓	11,697	7	Default 5-fold CV
FERV39k [64]	✓	38,935	7	Default train & test
MAFW [32]	✓	9,172	11	Default 5-fold CV
CREMA-D [5]	×	7,442	6	5-fold CV †
RAVDESS [36]	×	1,440	8	6-fold CV †
eNTERFACE05 [38]	×	1,287	6	5-fold CV †

Pre-training. The original videos provided in VoxCeleb2 have a resolution of 224×224. Given that the speaker's face generally does not fill the entire frame, we only used a 160 × 160 patch located in the upper center of each video frame to remove the irrelevant background information. During pre-training, we extract 16 frames from each video clip using a temporal stride of 4. This results in 8 × 10 × 10 input tokens after cube embedding, when using a cube size of 2 × 16 × 16. Regarding hyperparameters, we mainly follow VideoMAE. Specifically, we use an AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, an overall batch size of 128, a base learning rate of $3e-4$, and a weight decay of 0.05. We linearly scale the base learning rate according to the overall batch size, using the formula: $\text{lr} = \text{base learning rate} \times \frac{\text{batch size}}{256}$. In addition, we use a cosine decay learning rate scheduler. By default, we pre-train the model for 50 epochs, with 5 warmup epochs. When using 4 Nvidia Tesla V100 GPUs, the pre-training takes about 3-4 days.

Fine-tuning. Same as pre-training, the input clip size is 16 × 160 × 160 and the temporal stride is 4 for most datasets (except 1 for FERV39k). To optimize the model, we use an AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, with a base learning rate of $1e-3$ and an overall batch size of 96. The other hyperparameters remain the same as in pre-training, and more details can be found in [54]. We fine-tune the pre-trained model for 100 epochs, with 5 warmup epochs. During inference, we sample two clips uniformly along the temporal axis for each video and then calculate the average score as the final prediction.

4.3 Ablation Studies

In this part, we conduct ablation experiments on DFEW and FERV39k to demonstrate the effects of several key factors in MAE-DFER. For simplicity, on DFEW, we only report results of fold 1 (fd1).

Pre-training Epochs. As shown in Table 2, we observe that longer pre-training is generally beneficial and the performance saturation occurs at around 50 epochs. Besides, we also find that the performance of training from scratch (i.e., #Epochs=0) is very poor (nearly random guessing). This is largely attributed to the limited training samples in current DFER datasets since large vision Transformers are data-hungry and training them typically requires more than million-level labeled data [13, 54]. This result also demonstrates the significance and superiority of large-scale self-supervised pre-training over traditional supervised learning.

Comparison of Different Model Architectures. We then investigate the effect of three key modules in LGI-Former by evaluating the performance of the following variants: 1) only local intra-region self-attention (i.e., no global inter-region self-attention

Table 2: Ablation study on the pre-training epochs.

Dataset	Metric	Pre-training Epochs				
		0	10	30	50	70
DFEW	UAR	15.08	58.05	60.46	**62.59**	61.93
	WAR	23.84	71.29	72.73	**74.88**	74.58
FERV39k	UAR	18.09	39.98	42.04	43.12	**43.15**
	WAR	28.33	50.21	51.62	**52.07**	52.01

Table 3: Ablation study on the model architecture. Intra: local intra-region self-attention. Inter: global inter-region self-attention. LGI: local-global interaction.

Dataset	Intra	Inter	LGI	#Params (M)	FLOPs (G)	UAR	WAR
DFEW	✓	×	×	51.2	42.7	59.66	72.00
	✓	✓	×	68.0	42.8	60.43	73.69
	✓	×	✓	68.1	49.6	60.98	74.58
	✓	✓	✓	84.9	49.8	62.59	74.88
	×	×	×	86.2	80.8	**62.85**	**74.93**
FERV39k	✓	×	×	51.2	42.7	40.94	50.88
	✓	✓	×	68.0	42.8	42.15	52.04
	✓	×	✓	68.1	49.6	42.25	52.01
	✓	✓	✓	84.9	49.8	43.12	52.07
	×	×	×	86.2	80.8	**43.72**	**52.52**

and local-global interaction), 2) no local-global interaction, 3) no global inter-region self-attention, and 4) using global space-time self-attention instead (i.e., ViT). The results are presented in Table 3. We have the following observations: 1) The first variant has the worst performance, which is as expected since only utilizing local intra-region self-attention does not allow local tokens to access global information. 2) Either global inter-region self-attention or local-global interaction contributes to better performance, demonstrating the effectiveness of these two modules in local-global information propagation. Besides, the latter is generally more effective than the former but at the cost of more computation. It also should be noted that global inter-region self-attention only introduces negligible computation (~0.1G FLOPs) thanks to the small number (i.e., 8) of representative tokens. 3) When combing the global inter-region self-attention with local-global interaction, LGI-Former achieves the best results. Besides, compared with the last variant which uses global space-time self-attention (i.e., ViT), we only observe slight performance drop (<0.6%) but large computation reduction (~38% FLOPs), thus demonstrating the efficiency of LGI-Former.

Effectiveness of Joint Masked Appearance and Motion Modeling. We study the effect of different loss weights in Equation 2, ranging from 1.0 (i.e., only the original appearance target) to 0.0 (i.e., only the motion target). As shown in Fig. 3, we find that the joint model outperforms the model with only one reconstruction target and it achieves the best performance when adopting a loss weight around 0.5. For instance, on DFEW fd1, the best joint model surpasses the standalone appearance model by 1.69% UAR and 0.92% WAR and its motion counterpart by 1.77% UAR and 1.02% WAR. These results indicate that joint masked appearance and motion modeling are indispensable to facilitate better spatiotemporal representation learning for DFER. In addition to our MAE-DFER, we apply it to VideoMAE (shown in Table 10 of Appendix), which can also bring further improvement (1.51% UAR with 0.30% WAR on DFEW fd1 and 0.39% UAR with 0.13% WAR on FERV39k).

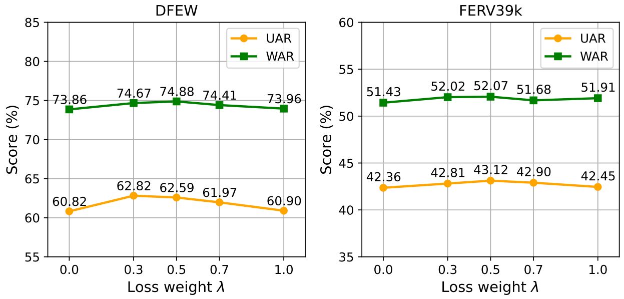

Figure 3: Ablation study on the loss weight.

Table 4: Ablation study on the local region size.

Dataset	Region size ($t \times h \times w$)	M	#Params (M)	FLOPs (G)	UAR	WAR
DFEW	$1 \times 5 \times 10$	16	84.9	49.8	62.36	74.33
	$2 \times 2 \times 10$	20	84.9	50.0	61.07	74.87
	$2 \times 5 \times 10$	8	84.9	49.8	**62.59**	**74.88**
	$2 \times 10 \times 10$	4	84.9	50.7	61.27	74.19
	$4 \times 5 \times 10$	4	84.9	50.7	62.36	74.67
FERV39k	$1 \times 5 \times 10$	16	84.9	49.8	42.71	52.26
	$2 \times 2 \times 10$	20	84.9	50.0	42.24	52.25
	$2 \times 5 \times 10$	8	84.9	49.8	**43.12**	52.07
	$2 \times 10 \times 10$	4	84.9	50.7	42.71	52.02
	$4 \times 5 \times 10$	4	84.9	50.7	43.09	**52.41**

Role of Local Region Size. We evaluate the effect of different local region sizes in LGI-Former and report the results in Table 4. We can find that the model performance is not very sensitive to the region size. Moreover, the model computation with different region sizes are similar to each other. These results indicate that, no matter how to divide the input into local regions, LGI-Former can achieve effective and efficient local-global information exchange via the introduced representative tokens and its specialized designs (i.e., the three key modules). Besides, when using the region size of $2 \times 5 \times 10$ (only using $M = 8$ representative tokens), the model achieves the best performance-computation trade-off.

4.4 Comparison with State-of-the-art Methods

Results on Large In-the-wild Datasets. We first compare MAE-DFER with previous state-of-the-art supervised methods on DFEW, FERV39k, and MAFW in Table 5, Table 6, and Table 7, respectively. On DFEW, MAE-DFER surpasses the previous best methods (i.e., DPC-Net [65] and M3DFEL [60]) with a significant margin, achieving a noteworthy **6.30%** UAR and **5.18%** WAR improvement. Besides, we also present fine-grained performance of each class in Table 12 of Appendix, MAE-DFER also achieves remarkable improvement across most facial expressions. Notably, for the *disgust* expression, which only accounts for 1.2% of the entire dataset and is very challenging for all baselines, MAE-DFER improves the best accuracy by

Table 5: Results on DFEW. †: pre-trained on VoxCeleb2. Underlined: the best supervised result. *Bold*: the best result.

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

Table 6: Results on FERV39k. †: pre-trained on VoxCeleb2. Underlined: the best supervised result. *Bold*: the best result.

Method	#Params (M)	FLOPs (G)	UAR	WAR
Supervised methods				
C3D [55]	78	39	22.68	31.69
P3D [46]	-	-	30.48	40.81
R(2+1)D [56]	-	-	31.55	41.28
3D ResNet-18 [21]	33	8	26.67	37.57
ResNet-18+LSTM [64]	-	-	30.92	42.59
VGG-13+LSTM [64]	-	-	32.42	43.37
Two C3D [64]	-	-	30.72	41.77
Two ResNet-18+LSTM [64]	-	-	31.28	43.20
Two VGG-13+LSTM [64]	-	-	32.79	44.54
Former-DFER [69]	18	9	37.20	46.85
STT [37]	-	-	37.76	48.11
NR-DFERNet [30]	-	6	33.99	45.97
IAL [29]	19	10	35.82	48.54
M3DFEL [60]	-	2	35.94	47.67
Self-supervised methods				
VideoMAE [54]	86	81	38.50	49.61
VideoMAE [54] †	86	81	**43.33**	**52.39**
MAE-DFER (ours)	85	50	43.12	52.07

Table 7: Results on MAFW. †: pre-trained on VoxCeleb2. Underlined: the best supervised result. *Bold*: the best result.

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**

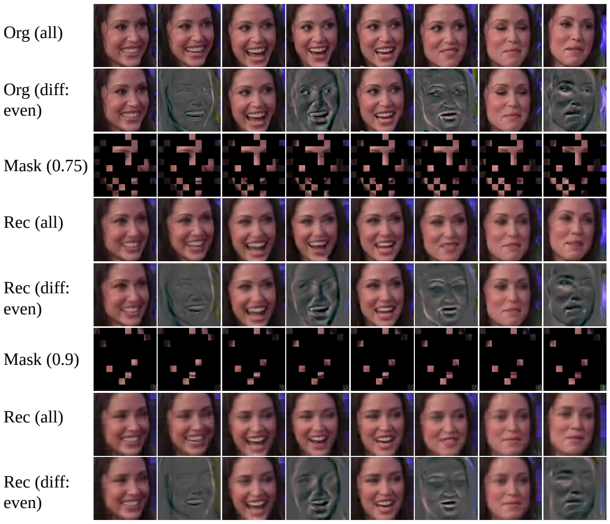

Figure 4: Reconstruction results of a VoxCeleb2 *test* video under masking ratios of 0.75 and 0.9. We only show 8 frames due to the space limitation.

over **10%**. This considerable improvement indicates that our method is capable of learning powerful representations for DFER via large-scale self-supervised pre-training. As for the other two datasets, we have similar observations. On the current largest DFER dataset, FERV39k, MAE-DFER achieves the new state-of-the-art performance, exceeding the previous best methods (i.e., STT [37] and IAL [29]) by **5.36%** UAR and **3.53%** WAR. On MAFW, MAE-DFER outperforms the best-performing T-ESFL [32] by a considerable margin of **8.34%** UAR and **6.13%** WAR. Besides, large performance improvement for several rare expressions are also observed on FERV39k and MAFW in Table 13 and Table 14 of Appendix. In summary, the promising results on three in-the-wild datasets demonstrate the strong generalization ability of MAE-DFER in practical scenarios.

Comparison with VideoMAE. To verify the effectiveness and efficiency of MAE-DFER, we also show the results of VideoMAE [54] on three in-the-wild datasets, including both the original model pre-trained on Kinetics-400 [6] for 1600 epochs and the model pre-trained on VoxCeleb2 under the same setting as MAE-DFER. From Table 5-7, we have the following observations: 1) The original VideoMAE model pre-trained on *general* videos (i.e., action recognition) is largely inferior to its counterpart pre-trained on *facial* videos, indicating that the large-domain gap between self-supervised pre-training and downstream fine-tuning will severely hurt the performance. 2) Compared with VideoMAE pre-trained on VoxCeleb2, our MAE-DFER largely reduces the computational

Table 8: Results on three lab-controlled datasets. Underlined: the best supervised result. *Bold*: the best result.

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**

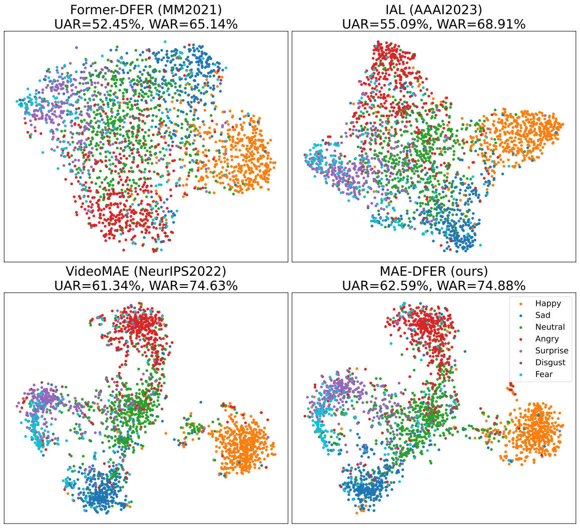

Figure 5: Embedding space visualization using t-SNE [58].

cost (~38% FLOPs) during fine-tuning, while achieving comparable performance on DFEW and FERV39k (only -0.11%~0.19% UAR and -0.17%~0.32% WAR), and even better performance on MAFW (+0.75% UAR and +0.80% WAR). Thus, these results demonstrate the effectiveness and efficiency of the proposed method.

Results on Small Lab-controlled Datasets. We show the comparison results on CREMA-D, RAVDESS, and eNTERFACE05 in Table 8. Compared with in-the-wild datasets, we observe *even larger* performance improvement on three lab-controlled datasets. On CREMA-D, our MAE-DFER outperforms the best unimodal methods by over **12%** UAR and **10%** WAR. More surprisingly, it also shows slightly better performance than the state-of-the-art multimodal method, thus amply demonstrating the superiority of MAE-DFER. On RAVDESS, MAE-DFER improves the previous best by more than **12%** WAR and also achieves comparable performance with the best audio-visual method. Finally, on eNTERFACE05, MAE-DFER surpasses the best-performing Graph-Tran [68] by about **7%** WAR.

4.5 Visualization Analysis

Reconstruction. We first visualize the reconstructed results of MAE-DFER in Fig. 4. The video is randomly selected from the VoxCeleb2 *test* set. For better visualization, we use a gray-style background for frame difference images shown in *even* frames and also show *all* the reconstructed video by adding the reconstructed frame difference images in *even* frames with the adjacent recovered *odd* frame images. From Fig. 4, we see that under such a high masking ratio (75% or 90%), MAE-DFER still can generate satisfactory reconstructed results for both the facial appearance content and temporal motion information. Notably, despite the change in identity information (as the model does not see this person during pre-training), the dynamic facial expression can be well restored by reasoning in limited visible contexts (e.g., the opening mouth). This imply that our model is able to learn meaningful dynamic facial representations that capture the global spatiotemporal structure.

Embedding Space. To further qualitatively show the superiority of MAE-DFER over traditional supervised methods, we visualize the learned embeddings using t-SNE [58] on DFEW fd1. As can be seen in Fig. 5, the embeddings of our method are more compact and separable than those of two state-of-the-art supervised methods (i.e., IAL [29] and Former-DFER [69]), which demonstrates that MAE-DFER can learn more discriminative representations for different dynamic facial expressions through large-scale self-supervised pre-training. Besides, VideoMAE has similar embedding space with our MAE-DFER but at the cost of much larger computational cost.

5 CONCLUSION

In this paper, we have presented an effective and efficient self-supervised framework, namely MAE-DFER, to exploit large amounts of unlabeled facial videos to address the dilemma of current supervised methods and promote the development of DFER. We believe MAE-DFER will serve as a strong baseline and foster relevant research in DFER. In the future, we plan to explore the scaling behavior of MAE-DFER (i.e., using more data and larger models). Beside, it is also interesting to apply it to other related tasks (e.g., dynamic micro-expression recognition and facial action unit detection).

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61831022, No.62276259, No.62201572, No.U21B2010), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (No.Z211100004821013), Open Research Projects of Zhejiang Lab (No.2021KH0AB06), and CCF-Baidu Open Fund (No.OF2022025).

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*. https://openreview.net/forum?id=p-BhZSz59o4

[3] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. 2022. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics* 18, 8 (2022), 5619–5627.

[4] Young-Hyen Byeon and Keun-Chang Kwak. 2014. Facial expression recognition using 3d convolutional neural network. *International journal of advanced computer science and applications* 5, 12 (2014).

[5] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

[6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International conference on machine learning*. PMLR, 1691–1703.

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[9] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. *Proc. Interspeech 2018* (2018), 1086–1090.

[10] Charles Darwin and Phillip Prodger. 1998. *The expression of the emotions in man and animals.* Oxford University Press, USA.

[11] Celso M de Melo, Jonathan Gratch, Stacy Marsella, and Catherine Pelachaud. 2023. Social Functions of Machine Emotional Expressions. *Proc. IEEE* (2023).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[14] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 467–474.

[15] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*. 445–450.

[16] Jiemin Fang, Lingxi Xie, Xinggang Wang, Xiaopeng Zhang, Wenyu Liu, and Qi Tian. 2022. Msg-transformer: Exchanging local spatial information by manipulating messenger tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12063–12072.

[17] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. 2022. Masked Autoencoders As Spatiotemporal Learners. *arXiv preprint arXiv:2205.09113* (2022).

[18] Ziwang Fu, Feng Liu, Hanyang Wang, Jiayin Qi, Xiangling Fu, Aimin Zhou, and Zhibin Li. 2021. A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. *arXiv preprint arXiv:2111.02172* (2021).

[19] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis. 2019. Multimodal and temporal perception of audio-visual cues for emotion recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 552–558.

[20] Lucas Goncalves and Carlos Busso. 2022. Robust Audiovisual Emotion Recognition: Aligning Modalities, Capturing Temporal Information, and Handling Missing Features. *IEEE Transactions on Affective Computing* 13, 04 (2022), 2156–2170.

[21] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16000–16009.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[24] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[25] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2881–2889.

[26] Dimitrios Kollias and Stefanos Zafeiriou. 2020. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing* 12, 3 (2020), 595–606.

[27] Jean Kossaifi, Antoine Toisoul, Adrian Bulat, Yannis Panagakis, Timothy M Hospedales, and Maja Pantic. 2020. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6060–6069.

[28] Yuanyuan Lei and Houwei Cao. 2023. Audio-Visual Emotion Recognition With Preference Learning Based on Intended and Multi-Modal Perceived Labels. *IEEE Transactions on Affective Computing* (2023), 1–16. https://doi.org/10.1109/TAFFC.2023.3234777

[29] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. 2023. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 67–75.

[30] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. 2022. NR-DFERNet: Noise-Robust Network for Dynamic Facial Expression Recognition. *arXiv preprint arXiv:2206.04975* (2022).

[31] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing* 13, 3 (2020), 1195–1215.

[32] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia*. 24–32.

[33] Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhongwen Luo. 2022. Clip-aware expressive feature learning for video-based facial expression recognition. *Information Sciences* 598 (2022), 182–195.

[34] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. 2021. Expression Snippet Transformer for Robust Video-based Facial Expression Recognition. *arXiv preprint arXiv:2109.08409* (2021).

[35] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. 2023. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition* 138 (2023), 109368.

[36] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.

[37] Fuyan Ma, Bin Sun, and Shutao Li. 2022. Spatio-Temporal Transformer for Dynamic Facial Expression Recognition in the Wild. *arXiv preprint arXiv:2205.04749* (2022).

[38] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. 2006. The eNTERFACE'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 8–8.

[39] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. 2019. Frame attention networks for facial expression recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*. IEEE, 3866–3870.

[40] Ryo Miyoshi, Noriko Nagata, and Manabu Hashimoto. 2019. Facial-expression recognition from video using enhanced convolutional lstm. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 1–6.

[41] Ryo Miyoshi, Noriko Nagata, and Manabu Hashimoto. 2021. Enhanced convolutional LSTM with spatial and temporal skip connections and temporal gates for facial expression recognition from video. *Neural Computing and Applications* 33, 13 (2021), 7381–7392.

[42] Xianzhang Pan, Wenping Guo, Xiaoying Guo, Wenshu Li, Junjie Xu, and Jinzhao Wu. 2019. Deep temporal–spatial aggregation for video-based facial expression recognition. *Symmetry* 11, 1 (2019), 52.

[43] Xianzhang Pan, Guoliang Ying, Guodong Chen, Hongming Li, and Wenshu Li. 2019. A deep spatial and temporal aggregation framework for video-based facial expression recognition. *IEEE Access* 7 (2019), 48807–48815.

[44] Maja Pantic and Leon J. M. Rothkrantz. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence* 22, 12 (2000), 1424–1445.

[45] Rosalind W Picard. 2000. *Affective computing.* MIT press.

[46] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.

[47] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[48] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[49] Yuxin Song, Min Yang, Wenhao Wu, Dongliang He, Fu Li, and Jingdong Wang. 2022. It Takes Two: Masked Appearance-Motion Modeling for Self-supervised Video Transformer Pre-training. *arXiv preprint arXiv:2210.05234* (2022).

[50] Lang Su, Chuqing Hu, Guofa Li, and Dongpu Cao. 2020. Msaf: Multimodal split attention fusion. *arXiv preprint arXiv:2012.07175* (2020).

[51] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing* (2023).

[52] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. 2020. Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*. 27–34.

[53] Jianhua Tao and Tieniu Tan. 2005. Affective computing: A review. In *Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1*. Springer, 981–995.

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

[55] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[56] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

[57] Minh Tran and Mohammad Soleymani. 2022. A Pre-Trained Audio-Visual Transformer for Emotion Recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4698–4702.

[58] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[60] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. 2023. Rethinking the Learning Paradigm for Dynamic Facial Expression Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17958–17968.

[61] Kexin Wang, Zheng Lian, Licai Sun, Bin Liu, Jianhua Tao, and Yin Fan. 2022. Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 75–80.

[62] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. 2021. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1895–1904.

[63] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[64] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. 2022. FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20922–20931.

[65] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. 2022. DPCNet: Dual Path Multi-Excitation Collaborative Network for Facial Expression Representation Learning in Videos. In *Proceedings of the 30th ACM International Conference on Multimedia*. 101–110.

[66] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[67] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. 2022. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173* (2022).

[68] Rui Zhao, Tianshan Liu, Zixun Huang, Daniel PK Lun, and Kin-Man Lam. 2022. Spatial-Temporal Graphs Plus Transformers for Geometry-Guided Facial Expression Recognition. *IEEE Transactions on Affective Computing* (2022).

[69] Zengqun Zhao and Qingshan Liu. 2021. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1553–1561.

In Appendix, we provide more information about six DFER dataset, additional ablation studies, and fine-grained results on three in-the-wild datasets.

A DATASETS

DFEW comprises 16,372 video clips extracted from over 1,500 high-definition movies. Each video clip is annotated with seven basic emotions (i.e., happy, sad, neutral, anger, surprise, disgust, and fear). We only use 11,697 single-labeled clips in this paper.

FERV39k is currently the largest real-world dynamic facial expression dataset. It has 38,935 video clips with an average length of 1.5 seconds and is annotated with seven basic emotions.

MAFW is a multimodal compound in-the-wild affective dataset, consisting of 10,045 video clips annotated with 11 compound emotions (including contempt, anxiety, helplessness, disappointment, and seven basic emotions). In this paper, we conduct experiments on 9,172 single-labeled video clips.

CREMA-D is a high-quality audio-visual dataset with 7,442 video clips. Each of them is labeled with six emotions, including happy, sad, anger, fear, disgust, and neutral.

RAVDESS is an audio-visual dataset that includes emotional speech and song. It consists of 2,880 video clips, each labeled with 8 emotions (i.e., seven basic emotions and calm). In this paper, we only use the speech part consisting of 1,440 video clips.

eNTERFACE05 is an audio-visual emotion recognition dataset that contains approximately 1,200 video clips, each simulating six emotions, including anger, disgust, fear, happy, sad, and surprise.

B MORE ABLATION STUDIES

Model Size. We investigate the effect of different sizes of LGI-Former to downstream performance. In addition to the default *base* version (512-dim), we also design two smaller versions, i.e., *small* (384-dim) and *tiny* (256-dim). The *small* version has roughly half parameters and FLOPs of the *base* version and it is similar for *tiny* and *small*. As shown in Table 9, we find that the performance only degrades moderately when the model size becomes smaller, especially for FERV39k. It is worth noting that even the *tiny* version still largely outperforms the state-of-the-art supervised methods (such as DPCNet [65] and IAL [29] in Table 5 and Table 6), despite that they has similar parameters and computational cost, which thus further demonstrates the superiority of our proposed method.

Table 9: Ablation study on the model size.

Dataset	Size	Dim	#Params (M)	FLOPs (G)	UAR	WAR
DFEW	Tiny	256	21.5	13.0	59.90	73.30
	Small	384	47.9	28.4	61.09	74.03
	Base	512	84.9	49.8	**62.59**	**74.88**
FERV39k	Tiny	256	21.5	13.0	41.20	51.55
	Small	384	47.9	28.4	42.04	**52.24**
	Base	512	84.9	49.8	**43.12**	52.07

VideoMAE with Joint Masked Appearance and Motion Modeling. Besides our MAE-DFER, we further introduce explicit temporal facial motion modeling to VideoMAE. The results are presented in Table 10. Similar to our MAE-DFER, we observe that joint masked appearance and motion modeling can further boost the performance of VideoMAE, although standalone motion modeling performs slightly worse than standalone appearance modeling in the original VideoMAE.

Table 10: Ablation study on VideoMAE with additional temporal facial motion modeling.

Dataset	Appearance	Motion	#Params (M)	FLOPs (G)	UAR	WAR
DFEW	×	✓	86.2	80.8	60.86	74.02
	✓	×	86.2	80.8	61.34	74.63
	✓	✓	86.2	80.8	**62.85**	**74.93**
FERV39k	×	✓	86.2	80.8	42.17	51.96
	✓	×	86.2	80.8	43.33	52.39
	✓	✓	86.2	80.8	**43.72**	**52.52**

Role of Classification Token Type. We finally explore the effect of two different classification tokens (i.e., original tokens and representative tokens) for downstream fine-tuning. As shown in Table 11, we find that performing mean pooling on the representative tokens for final classification slightly outperforms that on the original tokens. We speculate that this is because the representative tokens are more compact and high-level than the original tokens.

Table 11: Ablation study on the classification token type.

Token type	DFEW		FERV39k	
	UAR	WAR	UAR	WAR
Original tokens	62.16	74.51	42.89	51.91
Representative tokens	**62.59**	**74.88**	**43.12**	**52.07**

C DETAILED RESULTS

In this section, we first present more fine-grained results (i.e., accuracy of each class) on DFEW, FERV39k, and MAFW in Table 12, Table 13, and Table 14, respectively. From three tables, we observe that MAE-DFER significantly outperforms the state-of-the-art supervised methods on most facial expressions, especially on some *rare* facial expressions (such as *disgust*, *contempt*, and *disappointment*). For instance, on DFEW, our MAE-DFER surpasses the previous best supervised results by about 9% on *sad*, 13% on *disgust*, and 8% on *fear*. On MAFW, it improves the best-performing supervised methods by over 5% on *anger*, 7% on *disgust*, 8% on *contempt*, 8% on *anxiety*, 6% on *helplessness*, and 7% on *disappointment*. Moreover, compared with VideoMAE pre-trained under the same setting, MAE-DFER has comparable or even better fine-grained performance while largely reduces the computational cost during fine-tuning. We also note that the original VideoMAE pre-trained on Kinetics-400 does not perform well on some rare expressions (e.g., *disgust* on FERV39k), although it could achieve the best results on some dominated expressions (e.g., *neutral* on FERV39k). These results indicate that our MAE-DFER can effectively and efficiently learn more robust and general representations for DFER via large-scale self-supervised training on abundant unlabeled facial videos, thus mitigating the unbalanced learning issue and achieving superior fine-grained performance.

Table 12: Results on DFEW. †: pre-trained on VoxCeleb2. Underlined: the best supervised result. *Bold*: the best result.

Method	#Params (M)	FLOPs (G)	Accuracy of Each Emotion (%)							Metric (%)	
			Happy	Sad	Neutral	Anger	Surprise	Disgust	Fear	UAR	WAR
Supervised methods											
C3D [55]	78	39	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
R(2+1)D-18 [56]	33	42	79.67	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22
3D ResNet-18 [21]	33	8	76.32	50.21	64.18	62.85	47.52	0.00	24.56	46.52	58.27
EC-STFL [25]	-	8	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
ResNet-18+LSTM [69]	-	8	83.56	61.56	68.27	65.29	51.26	0.00	29.34	51.32	63.85
ResNet-18+GRU [69]	-	8	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02
Former-DFER [69]	18	9	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
CEFLNet [33]	13	-	84.00	68.00	67.00	70.00	52.00	0.00	17.00	51.14	65.35
EST [34]	43	-	86.87	66.58	67.18	71.84	47.53	5.52	28.49	53.43	65.85
STT [37]	-	-	87.36	67.90	64.97	71.24	53.10	3.49	34.04	54.58	66.65
NR-DFERNet [30]	-	6	88.47	64.84	70.03	75.09	61.60	0.00	19.43	54.21	68.19
DPCNet [65]	51	10	-	-	-	-	-	-	-	57.11	66.32
IAL [29]	19	10	87.95	67.21	70.10	76.06	62.22	0.00	26.44	55.71	69.24
M3DFEL [60]	-	2	89.59	68.38	67.88	74.24	59.69	0.00	31.63	56.10	69.25
Self-supervised methods											
VideoMAE [54]	86	81	92.23	67.81	70.97	74.02	62.59	10.34	31.49	58.49	70.61
VideoMAE [54] †	86	81	**93.09**	**78.78**	71.75	**78.74**	**63.44**	17.93	41.46	**63.60**	**74.60**
MAE-DFER (ours)	85	50	92.92	77.46	**74.56**	76.94	60.99	**18.62**	**42.35**	63.41	74.43

Table 13: Results on FERV39k. †: pre-trained on VoxCeleb2. Underlined: the best supervised result. *Bold*: the best result.

Method	#Params (M)	FLOPs (G)	Accuracy of Each Emotion (%)							Metric (%)	
			Happy	Sad	Neutral	Anger	Surprise	Disgust	Fear	UAR	WAR
Supervised methods											
C3D [55]	78	39	48.20	35.53	52.71	13.72	3.45	4.93	0.23	22.68	31.69
P3D [46]	-	-	61.85	42.21	49.80	42.57	10.50	0.86	5.57	30.48	40.81
R(2+1)D [56]	-	-	59.33	42.43	50.82	42.57	16.30	4.50	4.87	31.55	41.28
3D ResNet-18 [21]	33	8	57.64	28.21	59.60	33.29	4.70	0.21	3.02	26.67	37.57
ResNet-18+LSTM [64]	-	-	61.91	31.95	61.70	45.93	14.26	0.00	0.70	30.92	42.59
VGG-13+LSTM [64]	-	-	66.26	51.26	53.22	37.93	13.64	0.43	4.18	32.42	43.37
Two C3D [64]	-	-	54.85	52.91	60.67	31.34	5.96	2.36	6.96	30.72	41.77
Two ResNet-18+LSTM [64]	-	-	59.00	45.87	61.90	40.15	9.87	1.71	0.46	31.28	43.20
Two VGG-13+LSTM [64]	-	-	69.65	47.31	52.55	47.88	7.68	1.93	2.55	32.79	44.54
Former-DFER [69]	18	9	65.65	51.33	56.74	43.64	21.94	8.57	12.53	37.20	46.85
STT [37]	-	-	69.77	47.81	59.14	47.41	20.22	10.49	9.51	37.76	48.11
NR-DFERNet [30]	-	6	69.18	**54.77**	51.12	49.70	13.17	0.00	0.23	33.99	45.97
IAL [29]	19	10	-	-	-	-	-	-	-	35.82	48.54
M3DFEL [60]	-	2	-	-	-	-	-	-	-	35.94	47.67
Self-supervised methods											
VideoMAE [54]	86	81	71.28	48.60	**63.99**	47.28	20.69	5.35	12.30	38.50	49.61
VideoMAE [54] †	86	81	72.91	54.34	59.50	**51.65**	29.47	17.77	**17.63**	**43.33**	**52.39**
MAE-DFER (ours)	85	50	**73.05**	53.98	59.14	50.44	**30.09**	**17.99**	17.17	43.12	52.07

Table 14: Results on MAFW. †: pre-trained on VoxCeleb2. Underlined: the best supervised result. *Bold*: the best result.

Method	#Params (M)	FLOPs (G)	Accuracy of Each Emotion (%)											Metric (%)	
			AN	DI	FE	HA	NE	SA	SU	CO	AX	HL	DS	UAR	WAR
ResNet-18 [23]	11	-	45.02	9.25	22.51	70.69	35.94	52.25	39.04	0.00	6.67	0.00	0.00	25.58	36.65
ViT [13]	-	-	46.03	18.18	27.49	76.89	50.70	68.19	45.13	1.27	18.93	1.53	1.65	32.36	45.04
C3D [55]	78	39	51.47	10.66	24.66	70.64	43.81	55.04	46.61	1.68	24.34	5.73	4.93	31.17	42.25
ResNet-18+LSTM [32]	-	-	46.25	4.70	25.56	68.92	44.99	51.91	45.88	1.69	15.75	1.53	1.65	28.08	39.38
ViT+LSTM [32]	-	-	42.42	14.58	**35.69**	76.25	54.48	68.87	41.01	0.00	24.40	0.00	1.65	32.67	45.56
C3D+LSTM [32]	-	-	54.91	0.47	9.00	73.43	41.39	64.92	**58.43**	0.00	24.62	0.00	0.00	29.75	43.76
Former-DFER [69]	18	9	-	-	-	-	-	-	-	-	-	-	-	31.16	43.27
T-ESFL [32]	-	-	62.70	2.51	29.90	**83.82**	**61.16**	67.98	48.50	0.00	9.52	0.00	0.00	33.28	48.18
Self-supervised methods															
VideoMAE [54]	86	81	62.23	23.32	32.64	78.18	60.28	66.60	56.81	0.41	27.62	5.34	8.24	38.34	51.74
VideoMAE [54] †	86	81	65.90	23.63	34.88	76.73	55.62	**73.47**	54.57	**9.75**	32.75	10.69	11.54	40.87	53.51
MAE-DFER (ours)	85	50	**67.77**	**25.35**	34.88	77.13	58.26	71.09	57.46	8.90	**33.08**	**11.83**	**12.09**	**41.62**	**54.31**