Video Representation Learning for Conversational Facial Expression Recognition Guided by Multiple View Reconstruction

Valeriya Strizhkova[1,2], Laura M. Ferrari[1,2], Hadi Kachmar[1,2],
Antitza Dantcheva[1,2], François Brémond[1,2]
[1] INRIA [2] Université Côte d'Azur

Abstract

Conversational facial expression recognition entails challenges such as handling of facial dynamics, small available datasets, low-intensity and fine-grained emotional expressions and extreme face angle. Towards addressing these challenges, we propose the Masking Action Units and Reconstructing multiple Angles (MAURA) pre-training. MAURA is an efficient self-supervised method that permits the use of small datasets, while preserving end-to-end conversational facial expression recognition with Vision Transformer. MAURA masks videos using the location with active Action Units and reconstructs synchronized multi-view videos, thus learning the dependencies between muscle movements and encoding information, which might only be visible in few frames and/or in certain views. Based on one view (e.g., frontal), the encoder reconstructs other views (e.g., top, down, laterals). Such masking and reconstructing strategy provides a powerful representation, beneficial in facial expression downstream tasks. Our experimental analysis shows that we consistently outperform the state-of-the-art in the challenging settings of low-intensity and fine-grained conversational facial expression recognition on four datasets including in-the-wild DFEW, CMU-MOSEI, MFA and multi-view MEAD. Our results suggest that MAURA is able to learn robust and generic video representations.

1. Introduction

Conversational Facial Expression Recognition (cFER) aims to categorize emotional expressions in videos, where emotional facial expressions occur jointly with talking-related facial expressions. Facial Expression Recognition (FER) aims at identifying and categorizing emotional *expressions* elicited by humans [29]. In this setting, the labels pertained to the emotional expressions are typically annotated by human evaluators. Often, emotional expressions are attributed according to the six basic emotions introduced by Ekman

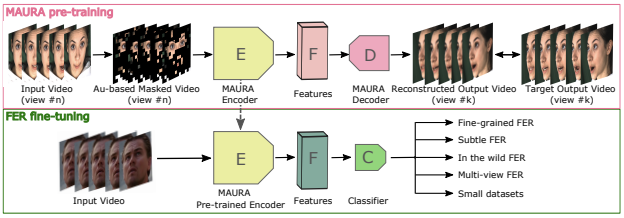

Figure 1. Overview of the proposed Masking Action Units and Reconstructing multiple Angles (MAURA) pre-training for various Facial Expression Recognition (FER) downstream tasks, including multi-view, fine-grained anger, low-intensity, and in-the-wild FER. MAURA learns a generic facial representation from available multi-view facial video data by utilizing dependencies between action units and reconstructing different views.

[8], namely 'fear', 'anger', 'joy', 'sadness', 'disgust', and 'surprise'. In the field of affective computing, the goal often is to identify *emotions*, where the associated labels have been self-annotated by the subjects [29]. Therefore, the goal of this work relates to FER. In the context of Conversational Facial Expression Recognition, open challenges include (1) distinguish facial expressions associated to emotions and speech, (2) recognise fine-grained emotional expressions, (3) insufficient visibility of the face due to pose-changes, and (4) small available training datasets. When humans talk, while exhibiting emotional expressions, muscle activation might be due to speech, as well as due to emotions, represented by challenge (1). The second challenge is the fine-grained discrete emotional expression recognition, where different shades of the same emotion (e.g. anger) have to be identified (2). In addition, pose variations might occlude part of the face hiding the emotional expression (3). Challenge (4) is related to the limited amount of annotated videos, fundamental in supervised learning of facial representations for affective computing.

Pre-training is a beneficial technique to exploit large datasets without annotations, in order to learn generic representations. *Pre-training* techniques, which do not utilize semantic information pertained to the data, show the

best performance on many downstream tasks including action recognition [24, 30, 33]. Recently, MARLIN [5] is introduced as a pre-training techniques for facial videos. It is based on Video Masked AutoEncoder (VideoMAE) [30] that learns video representations by randomly masking the input video and reconstructing it with an asymmetric encoder-decoder model. MARLIN proposes a masking strategy based on segments of the face: one part of the face is not masked (e.g., mouth), while other parts are masked and reconstructed based on the visible parts. While a masking method has been proposed, which utilizes facial information, we note that there is space for improving the masking and reconstructing strategies for facial expression recognition tasks. For example, the MARLIN masking method does not utilize the activations in facial expressions, i.e., a part of the face without movements might be unmasked and it might not contain enough information to reconstruct other parts of the face with muscle movements. The idea of reconstructing multi-view to encode powerful representations has never been exploited to overcome the issue of pose-changes and fine-grained FER. If the pre-training data contains a few videos taken from various angles, the transferability of the pre-trained encoder for the downstream tasks with different views might not be optimal.

We propose a pre-training method for conversational facial expression recognition that overcomes the four challenges aforementioned, by Masking Action Units and Reconstructing multiple Angles (MAURA). Firstly, MAURA chooses not just a random part of the face not to mask, but instead, it retains the part of the face comprising active action units. Therefore, it forces the network to learn not just a correlation between facial parts, but the correlation between facial muscle movements. Secondly, MAURA reconstructs not only the input video, but videos taken simultaneously from different views, forcing to learn a 4D spatio-temporal information about a face. We show that MAURA learns rich and transferable video representations by demonstrating both linear probing and end-to-end fine-tuning results on four emotional video-based datasets. We show that our pre-training is instrumental in achieving state-of-the-art results in several tasks including fine-grained, low-intensity, multi-view and in-the-wild facial expression recognition.

Our contributions are summarized as follows.

1. We propose a new masking strategy based on facial muscle movements and show its advantages for cFER.
2. We propose the first multi-view representation learning for cFER and demonstrate a detailed analysis of its quantitative and qualitative performance.
3. Our proposed MAURA pre-training method, based on Masking Action Units and Reconstructing multiple Angles, achieves the state-of-the-art results in fine-grained, low-intensity, multi-view and in-the-wild cFER tasks.

2. Related Work

2.1. Conversational Facial Expression Recognition

Facial expression recognition from static frames exhibits promising classification accuracy [15, 23]. However, it remains challenging to detect discrete emotional expressions in videos where people talk, while showcasing their feelings. Image-based methods are not applicable to conversational videos as facial muscle movements in a frame might be related to speech rather than to emotional expressions. Therefore, the dynamics of an input video are taken into account in the early cFER algorithms [19, 20] by utilizing RNNs and CNNs. More recent approaches [6, 35, 37] extract features with other pre-trained models and feed the frozen features to a shallow Transformer model to recognize emotional expressions in conversational videos. Delbrouck *et at.* [6] extracted visual features with R(2+1)D-152 [31] and Zhang *et al.* [35] used unsupervised MAE-based [12] and supervised IResNet-based [4] and DenseNet-based [14] feature extractors, whereas Zhang *et al.* [37] extracted features using DLN [36]. These feature extractors output task-specific rather than generic features, so important information may be lost. To learn more generic and robust facial video representations, a sufficient amount of data and adequate pre-training are needed.

MARLIN [5] is the current state-of-the-art in cFER on the CMU-MOSEI dataset [34]. MARLIN is based on the self-supervised Video Masked Autoencoder (VideoMAE) pre-training [30] and it aims to learn universal facial video representations. It proposes a masking strategy based on facial segmentation: one part (e.g, mouth) is unmasked while others are masked and reconstructed based on the visible part. The main limitation of the method is that a chosen visible segment might not contain enough active muscle movements for reconstructing the dynamics of the masked parts. We overcome these limitations by keeping a facial part visible that contains an active Action Unit.

2.2. Multi-View Facial Expression Recognition

Several works [26, 27] utilize multi-view information to recognize facial expressions. Romero *et al.* [26] detected AU from multi-view videos. However, one model for each view is trained and each model detects AUs per frame, so the method does not utilize time information. Roy and Etemad [27] minimized the distance between the images with the same emotion obtained from different angles, and maximized the distance between the images with different emotional expressions without using any temporal information. Similar to [26, 27], we show the benefits of using multi-view data for FER. In contrast to [26, 27], we utilize temporal information of multi-view videos to recognize emotional expressions.

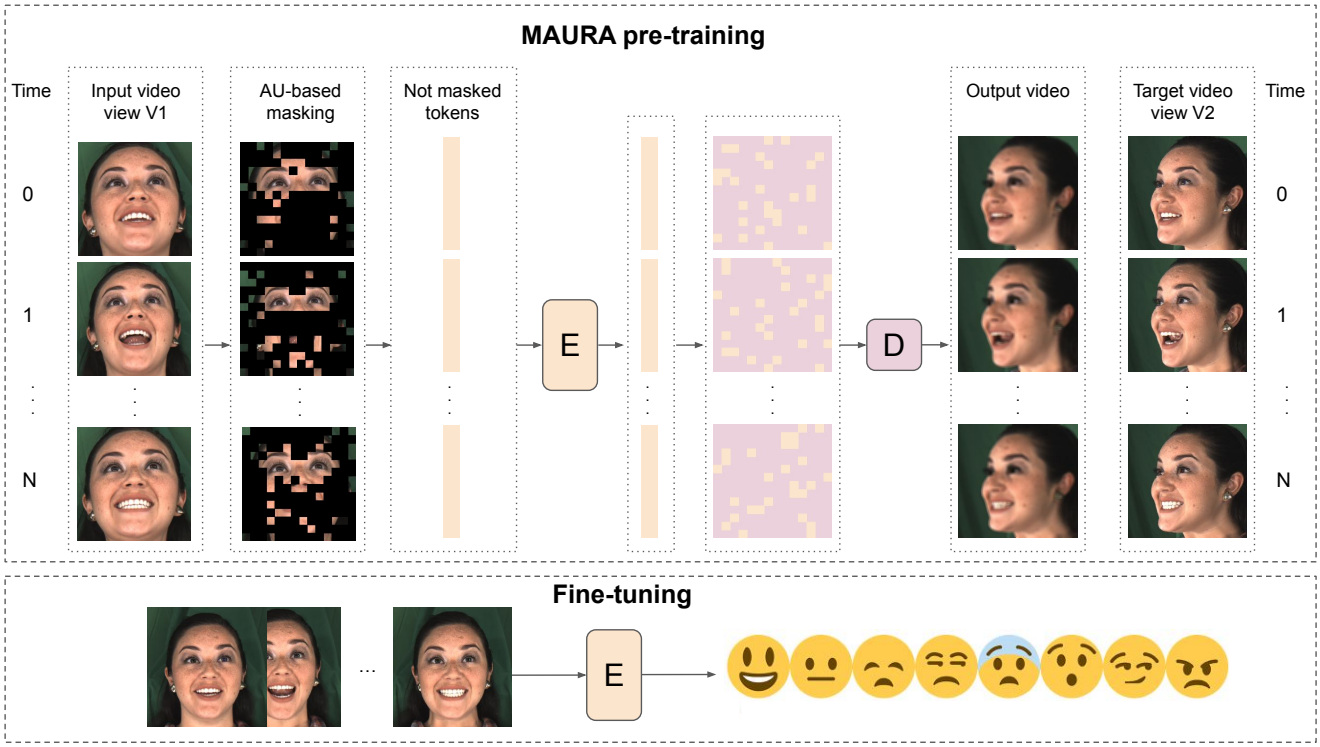

Figure 2. Overview of the Masking Action Units and Reconstructing multiple Angles (MAURA) autoencoder pre-training strategy. On top, the pre-training with masking AUs and reconstructing multiple views is represented. Below, the fine-tuning process is shown. The encoder, pre-trained with the MAURA method, takes a video as input and predicts one of the eight discrete emotional expressions: *happy*, *neutral*, *sad*, *disgust*, *fear*, *surprised*, *contempt*, *angry*.

3. Method

3.1. Revisiting Video Masked AutoEncoder

VideoMAE uses an asymmetric encoder-decoder architecture to reconstruct masked videos as a pre-training task for action recognition. The encoder and decoder are vanilla Vision Transformers (ViT) [7] with joint space-time attention [2, 10, 21, 22] so all pair tokens could interact with each other in the multi-head self-attention layer. The decoder is a narrower and shallower ViT than the encoder. VideoMAE represents an input video of size $T \times 3 \times W \times H$ as non-overlapping cube patches of size $t \times 3 \times w \times h$. VideoMAE applies cube embedding on the cube patches to produce the video tokens and masks an extremely high proportion (90%) of the tokens. Unmasked visible tokens are used along with corresponding positional space-time representations as input to the encoder. Then the decoder takes as input both encoded and learnable mask tokens with the joint space-time positional embeddings to reconstruct all normalized input cube patches. VideoMAE compares three masking strategies: (1) masking spacetime-agnostic patches, (2) masking temporally consistent tubes, and (3) masking spatially consistent frames.

3.2. MAURA

Figure 2 illustrates the overview of the Masking Action Units and Reconstructing multiple Angles (MAURA) pre-training. MAURA has an asymmetric encoder-decoder architecture similar to VideoMAE. The encoder and decoder are vanilla Vision Transformers (ViT) with joint space-time attention. The input and target videos are represented as cube patches, each patch is transformed to a token embedding. A high ratio of input tokens are masked with our proposed AU-based masking strategy and the unmasked visible tokens with the corresponding joint space-time positional embeddings enter the encoder. Unmasked tokens are mapped into latent features, which, along with joint spatio-temporal positional embeddings, are taken as input by the decoder to reconstruct the normalized target video cube patches. The input and the target videos are randomly sampled from seven synchronized videos from different views, so the input and output videos might be identical as in VideoMAE or from two different views. After the pre-training step, the encoder is used to fine-tune on the downstream tasks using the Cross Entropy loss for the cFER tasks. We evaluate the pre-training quality by end-to-end fine-tuning and linear probing the same as in [3, 11, 12].

Figure 3. Examples of the proposed mapping of 18 Action Units (AU) to patches.

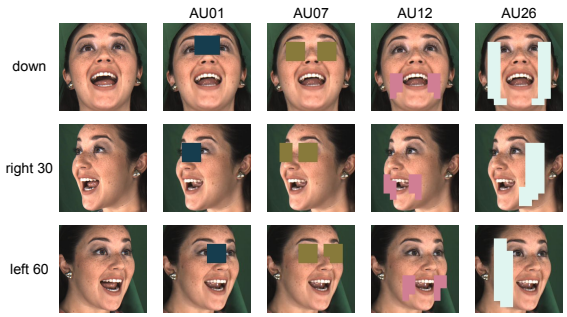

Figure 4. Masking Action Units (AUs) in the images from down, right 30° and left 60° views. The same AUs allocate different patch positions depending on the view, e.g. AU26 is masked on both sides of the face in the down view, whereas it is masked on only one side of the face in the right 30° and left 60° views.

3.3. Masking Action Units

Our proposed masking strategy is based on not masking a randomly selected active AU and masking other parts of the video. AUs are fine-grained facial muscle movements [9], each AU relates to a subset of extracted facial landmarks [25]. During each iteration, we randomly select one AU among all active AUs detected in a video by the OpenFace library [1, 28]. Then we do not mask patches where the chosen active AU is located. We use the rules from [25] to find patches corresponding to AUs, that is, each AU is placed on a patch with facial landmarks associated with that AU. E.g. AU26, associated with Jaw Drop, includes landmarks 51, 53, 57, and 59, so its corresponding patches are located at these landmarks. We track patches for unmasking in each frame of the video similar to Motion Guided Masking for VideoMAE (MGMAE) [13]. If a person turns their head and the selected AU is located in different patches of two frames of one video, then the locations of patches for unmasking are different in these frames. In an ablation study, we show that random masking of patches not associated with the selected active AU produces better results than tube masking. Therefore, our masking combines tube unmasking of a random active AU and random masking of other parts of the video.

We use the OpenFace library to detect the following 18 AUs: AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, and AU45. We use all 18 AUs detected by OpenPose. Figures 3 and 4 show how AUs correspond to the patches in the image depending on the facial expression and view. Some AUs allocate less space, e.g. AU12 takes only 10 patches. Some AUs are more visible from frontal/down/top views and can be overlapped by other parts of the face in lateral views.

3.4. Masking Ratio

In MAURA, we identify the best masking percentage to be 70%, as detailed in Table 5, in the ablation study. This is different from the VideoMAE and MGMAE methods where 90% and 85% masking ratios are reported as the most effective. The relatively low masking ratio of MAURA is attributed to shorter duration of facial expressions, as well as to related dense spatial localization. When facial expressions are present, they are usually visible in few frames and in specific locations of the face, which can be around the eyes or the mouth or the forehead. Also reconstructing a video from another view is a more difficult task that requires more visible unmasked tokens.

3.5. Reconstructing Multiple Views

The pipeline, as shown in Figure 2, is implemented in two steps, first the autoencoder is pre-trained and then the encoder is fine-tuned to classify emotions. During the pre-training step, the autoencoder performs a masking and reconstruction task (Figure 6.) Figure 2 shows that the input and target videos are from two different views. This is a more difficult pre-training that encodes information available from other angles. We use the MEAD dataset with seven simultaneously captured angles (i.e. frontal, top, down, left 30°, left 60°, right 30°, right 60°). During MAURA pre-training, the target video is randomly selected from seven possible views. Reconstructing multiple views enables the learning of an augmented representation, where details, visible in few frames from some angle, are encoded.

3.6. Loss

The loss function is the mean squared error (MSE) loss between the normalized masked tokens and reconstructed ones in the pixel space:

$$L = \frac{1}{\Omega} \sum_{p \in \Omega} |V(p) - \hat{V}(p)|^2, \quad (1)$$

where p is the token index, Ω is the set of all tokens, V is the input video, and $\hat{V}$ is the reconstructed one.

Table 1. The main characteristics of the four conversational emotion datasets used in this study.

Dataset	# Videos	Source	# Emotions
MEAD [32]	221K	48 actors	8
MFA [16]	200	YouTube	5
DFEW [17]	16K	1500 movies	7
CMU-MOSEI [34]	23K	1000 speakers	6

4. Experiments

4.1. Datasets and Preprocessing

We use four conversational datasets (Table 1), namely MEAD [32], MFA [16], DFEW [17], and CMU-MOSEI [34].

The **MEAD** dataset is the only one including multi-view and multi-intensity emotion samples. We use it for pre-training and cFER fine-tuning. It is a talking-face video corpus, where 48 performers are recorded while reproducing eight different emotional expressions at three intensities. The eight emotional expressions expand the six Ekman's basic emotions (i.e., anger, disgust, fear, sadness, neutral, contempt, surprise, and happiness) [8] with the neutral and the contempt states. The three intensities relate to low, normal, and high. The participants are simultaneously recorded from seven different views (Figure 5a): front, top, down, left 30°, left 60°, right 30°, right 60°. It is a large-scale and high-quality dataset with more than 220k videos and 1920x1080 pixel resolution.

MFA is a multicultural video dataset of negative emotional expressions in-the-wild [16] (Figure 5b). The MFA dataset expresses two major challenges in cFER. Firstly, it is small (around 200 videos). Secondly, it contains a multitude of emotional nuances. The majority of emotional datasets collect data under broad label (as "anger"), while we typically experience a wider range of emotional expressions such as annoyed, contemptuous and more. The fine-grained five labels used are: contempt, annoyed, anger, hatred, and furious. Notably, in MFA the subjects are not recorded only in frontal view, as happening in many in-the-wild datasets, they are moving freely in the scene.

Dynamic Facial Expression in-the-Wild (**DFEW**) is a large-scale facial expression database with 16,372 videos derived from movies. Videos in DFEW have challenging interferences, such as extreme illumination, occlusions and sudden pose changes. DFEW is annotated with seven discrete emotional expressions: happiness, sadness, neutral, anger, surprise, disgust, and fear.

CMU-MOSEI [34] is an in-the-wild conversational dataset with 23,453 annotated videos from 1000 distinct speakers. Each video is annotated with 6 classes: happiness, sadness, anger, fear, disgust, and surprise.

For all the datasets the input videos are cropped with the OpenFace library and resized to 240×240 pixels. We select OpenFace as it is a stable and largely adopted library in the community. We apply a random crop of the input frames to 224×224 pixels and random flip during training.

Figure 5. The MEAD and the MFA datasets. a) The MEAD dataset with the 8 emotional expressions, illustrated in high and low-intensity, and 7 views. b) The MFA dataset with the 5 emotional expressionsand examples of the encountered views with extreme face angle.

4.2. Experimental setup

In the first step, we pre-train the autoencoder on the MEAD dataset using the MAURA pre-training. In the second step, all weights of the pre-trained encoder are fine-tuned with fine-tuning (FT) or only the last linear layers of the pre-trained encoder are fine-tuned with linear probing (LP). For linear probing and fine-tuning we use multiple datasets: MEAD, MFA, DFEW, and CMU-MOSEI. For the MFA, DFEW and CMU-MOSEI datasets, we use the train/validation/test splits provided by the authors. In the case of MEAD, we use all views as input and target data, and also use 5-fold cross-validation, as we are the first to apply this dataset for FER.

4.3. Implementation Details

MAURA has an asymmetric encoder-decoder architecture where both the encoder and the decoder are ViT-B with 12 and 4 blocks, respectively. Except for learnable joint space-time positional embeddings, neither the encoder, the decoder, nor the masking strategy, has any spatio-temporal inductive bias. The input and target videos are temporally downsampled with a stride two and transformed to patches. Each patch has a size of $2 \times 3 \times 16 \times 16$, where 2 is the temporal size, 3 is the number of channels, and 16×16 is the spatial size. For a $16 \times 3 \times 224 \times 224$ video, this patch size produces $8 \times 14 \times 14 = 1568$ tokens.

Table 2. Comparison with state-of-the-art FER methods on the MEAD, DFEW, MFA and CMU-MOSI datasets. We compare Linear Probing (LP) and Fine-Tuning (FT) results. * denotes supervised methods.

Method	F1-score				Accuracy
	MEAD (low)	MEAD (high)	MFA	DFEW	CMU-MOSEI
ViT-B (pt on MEAD)*	41.2	42.2	43.4	43.6	-
3D Resnet18* [18]	-	-	-	41.1	-
MLKNN* [16]	-	-	42.0	-	-
UMONS* [6]	43.4	52.7	-	-	**80.7**
MARLIN [5]	-	-	-	-	80.6
VideoMAE [30]	43.3	46.1	52.7	43.6	80.4
MAURA (LP)	50.6	51.2	53.1	45.2	80.5
MAURA (FT)	**51.9**	**54.6**	**55.6**	**47.5**	**80.7**

5. Results

5.1. Comparison with SOTA

We adopt linear probing and fine-tuning for downstream adaptation. We show the results on three affective downstream tasks: low-intensity, fine-grained, and in-the-wild cFER.

5.1.1 Low-intensity cFER

To test low-intensity cFER we use the MEAD dataset. We firstly compare MAURA with the self-supervised VideoMAE and MARLIN pre-training methods. All methods use the ViT-B encoder and the same 70% masking percentage. We then compare our approach with supervised UMONS [6] that uses fixed Action Units features as input to a shallow Transformer encoder. The first two columns of Table 2 show that we obtain the best results with MAURA in both low and high-intensity cFER. In the case of low-intensity, we achieve the result of 51.9% F1-score using fine-tuning. Both VideoMAE pre-training and UMONS show a little (2%) increase over ViT-B without pre-training, reaching around 43%. This shows how difficult it is to classify low-intensity emotions using the available methods. For high-intensity FER using ViT-B without pre-training, F1-score is around 42%, which is increased with UMONS and VideoMAE by 10% and 4%, respectively. In the high-intensity FER, MAURA FT achieves the best F1-score of 54.6%.

5.1.2 Fine-Grained Anger cFER

To test the fine-grained anger cFER, we use the MFA dataset. MFA is quite challenging as it expresses multiple nuances of anger and the videos contain multiple views, representing free-moving humans. Moreover, it is small as it contains only 200 videos. The MAURA encoder is pre-trained on the MEAD dataset for both high and low-intensity emotions. To show the generalizability of our method, we fine-tune the pre-trained ViT-B on the MFA dataset to show transferable learning. We compare the obtained results with (1) the MLKNN method, which is the state-of-the-art on MFA, (2) ViT-B trained on MFA from scratch, (3) ViT-B pre-trained with VideoMAE and (4) ViT-B pre-trained with MAURA. We use V-F1-score from [16] where V means assigning the label to the whole video by taking the majority of predicted labels on the frames. Table 2 shows that our MAURA pre-training achieves the highest F1-score on MFA.

5.1.3 In-the-wild cFER

To further evaluate generalizability on datasets, we compare the LP and FT adaptation performance of MAURA with the current state-of-the-art methods on the DFEW and CMU-MOSEI datasets.

These are large datasets with difficulties related to occlusions and sudden pose changes. We compare the self-supervised MAURA pre-training with a fully supervised FER pre-training: ViT-B (pt on MEAD) in Table 2. We use ViT-B pre-trained with MAURA on MEAD for the unsupervised pre-training and ViT-B pre-trained on DFEW for the supervised pre-training. We fine-tune both pre-trained models on the MFA dataset using the same protocol. Table 2 shows that MAURA achieves a higher F1-score, being able to learn more generalizable representations.

We also show the results on the CMU-MOSEI dataset. UMONS is trained on video, audio, and text modalities, MARLIN and MAURA use only video modality. MAURA and UMONS show similar performance outperforming MARLIN with 0.1 accuracy. This small difference is explained by the fact that the visual modality is not considered the most relevant for this dataset, since the annotations are made primarily based on the linguistic modality.[6].

5.2. Ablation Study

We have performed extensive ablation studies to provide justification for our design choices. We provide the ablation results on the MEAD, DFEW and MFA datasets.

Table 3. Multi-view study. F1-score for ViT-B pre-trained with VideoMAE, VideoMAE+MAU, VideoMAE+RA, and MAURA on the different views of the MEAD dataset with low and high intensities.

View	MEAD low				MEAD high			
	VideoMAE	+ MAU	+ RA	+ Both (MAURA)	VideoMAE	+ MAU	+ RA	+ Both (MAURA)
front	54.2	54.9	55.3	56.0	57.1	57.2	57.4	57.9
left 30°	43.1	45.3	49.7	51.2	46.0	47.0	54.9	56.5
left 60°	34.0	38.1	47.8	48.5	36.6	38.2	50.2	52.3
right 30°	47.5	47.9	53.1	53.7	48.2	49.0	55.5	55.8
right 60°	36.9	38.0	51.3	49.5	38.0	42.1	51.2	52.5
top	45.2	47.6	51.8	52.3	49.8	50.0	53.8	54.7
down	42.6	44.4	49.5	50.0	47.1	48.3	50.7	52.6
average	43.3	45.2	51.2	**51.9**	46.1	47.4	53.4	**54.6**

Table 4. Contribution of Masking Action Units and Reconstructing multiple Angles on the MEAD (low-intensity), DFEW and MFA datasets. F1-score is reported.

Modules	MEAD low	DFEW	MFA
VideoMAE	43.3	43.6	52.7
+ MAU	45.2	44.9	53.4
+ RA	51.2	47.0	55.1
+ Both (MAURA)	**51.9**	**47.5**	**55.6**

Table 5. Ablation study on masking strategy and masking ratio on the MEAD (high-intensity) dataset. The reported ratio of masking for MAURA is related to the total amount of masking percentage, which combines the AU-masked + random masking. F1-score is reported.

Method	random		tube	
	70%	90%	70%	90%
VideoMAE	46.1	45.1	46.0	45.7
MAURA	**54.6**	46.1	48.2	46.3

5.2.1 Multi-View

Table 3 reports the study on multiple input views of the MEAD dataset. When studying multiple input views, MAURA is compared with VideoMAE, VideoMAE + MAU, and VideoMAE + RA. During pre-training, MAURA and VideoMAE + RA take a video from any of the seven views as input and reconstruct another video, randomly selected from the remaining six views. The same view is reconstructed in VideoMAE and VideoMAE + MAU. The results show that VideoMAE + RA significantly improves the recognition on lateral views while VideoMAE + MAU mainly improves the performance on the frontal view. Not big performance improvement using masking AUs might be ascribed to a more precise AU localization on the frontal view, while there are some mistakes in creating AU-based mask on other views. There are two main reasons for this. Firstly, OpenFace was trained primarily on front-view videos, so predicting AUs on other views may introduce errors and thus lead to degraded performance when pre-training with MAURA. Secondly, the rules for locating AUs on the face are created for frontal view and, thus perform worse on other views. Table 3 demonstrates that MAURA achieves the best result on average. While the frontal view gives the highest score in absolute, it is more relevant to compare the average results, as in real applications faces are presented in different angles.

5.2.2 Modules

Tables 4 and 5 show the ablation on Masking Action Units and Reconstructing Multiple Views strategies. The results show that there is a continuous improvement with each module added. Reconstructing multiple Angles gives the largest improvement on the MEAD, DFEW, and MFA datasets. Masking AUs improves F1-score more on MEAD and DFEW and less on MFA. This might be ascribed to the fact that there are more extreme views in MFA than in MEAD and DFEW and MFA has worse video quality.

5.2.3 Masking Strategy and Masking Ratio

Table 5 shows the study on masking strategy and masking percentage. We compare 70% and 90% percentages of video masking using random and tube masking strategies. In MAURA, the visible unmasked patches are detected with the tube strategy and other patches are masked using a random or tube masking strategy which correspond to *random* and *tube* in Table 5. Random and tube masking strategies are applied to both masked and unmasked patches in VideoMAE. The study is done on the MEAD dataset with high-intensity emotions. The best result is achieved using MAURA with the random masking strategy and 70% ratio.

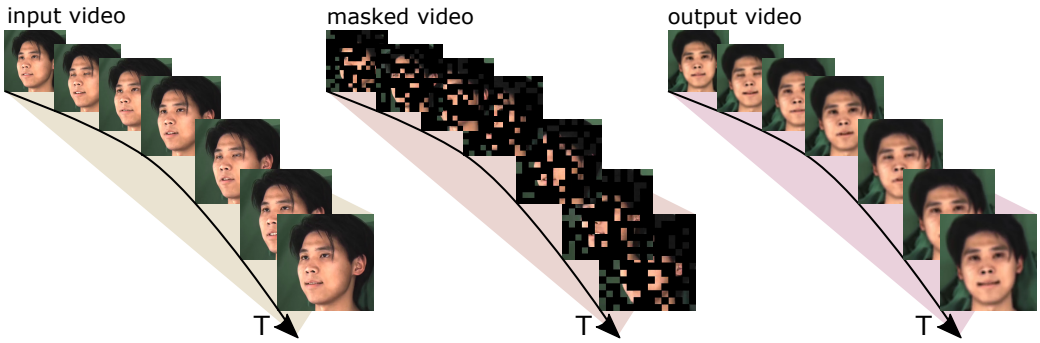

Figure 6. The input, masked and reconstructed output videos of the MAURA method with the MEAD dataset.

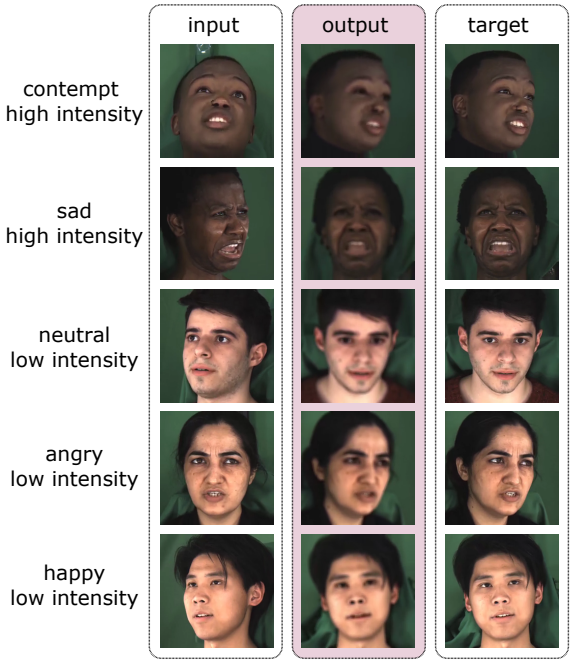

Figure 7. Qualitative analysis. In the central column the reconstructed videos of the different views for the high and low level intensity and five different emotions.

6. Discussion

The best results obtained with the MAURA pre-training relate to the reconstruction of other views being more beneficial than reconstructing the same video. MAURA outperforms VideoMAE by around 8% in both, high and low-intensity emotional expressions, thanks to encoding information from multiple views and learning the dependencies between muscle activations. Some examples of the reconstructed videos with the MAURA approach are reported in Figure 7. The relevance of using multi-view is supported by the results reported in Table 3 where we can see that different views have different recognition capabilities. With the MAURA method, the left 60° view is the one that encodes less information, both in high and low-intensity cases, however, the left 60° with MAURA is better than the one pre-trained with VideoMAE. The excellent F1-score achieved in the case of low-intensity emotions, when compared with VideoMAE and UMONS, can be ascribable to multiple reasons. First, small muscle movements, visible only from some views, are well encoded through the MAURA strategy. In Figure 7 we can see how the reconstructed facial expressions show different details, helping the model to discriminate better. Second, frozen features do not capture well low-intensity actions in all views. This is why raw data is needed in a more challenging task. Third, we hypothesize that low-intensity emotions are visible in fewer frames, meaning that the low-intensity is a reduced dataset. The ability to well discriminate fine-grained emotions is tested on the MFA dataset. The results confirm that our pre-training is a key step for such challenging and small datasets. The capability of MAURA in terms of transferable learning is further evaluated on the DFEW and CMU-MOSEI datasets. MAURA achieves very good F1-scores on DFEW showing that this method is able to learn powerful and transferable representations.

7. Conclusions and future work

The proposed MAURA approach is a highly efficient pre-training method for cFER, increasing the expressive power of video representation encoding of low-intensity facial movements. MAURA endows the ViT-B network with the ability to be successfully applied to small datasets. We show significant improvement over state-of-the-art *w.r.t.* classification accuracy in the challenging setting of in-the-wild, low-intensity and fine-grained cFER. Our future work will aim at adding additional modalities, including audio and language in the recognition step, and will also aim at exploring additional downstream tasks such as lip synchronization, AU detection, and DeepFake detection.

References

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 2016. 4

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *IEEE/CVF International Conference on Computer Vision*, 2021. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *International Conference on Learning Representations (ICLR)*, 2022. 3

[4] Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jorn-Henrik Jacobsen. Invertible residual networks. *International Conference on Machine Learning (ICML)*, 2019. 2

[5] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6

[6] Jean-Benoit Delbrouck, Noe Tits, Mathilde Brousmiche, and Stephane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2, 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. 3

[8] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 1992. 1, 5

[9] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 4

[10] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[11] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arxiv.org*, 2022. 3

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[13] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 4

[14] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv:1404.1869*, 2014. 2

[15] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[16] Roya Javadi and Angelica Lim. The many faces of anger: A multicultural video dataset of negative emotions in the wild (mfa-wild). *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2021. 5, 6

[17] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. *ACM Multimedia*. 5

[18] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. *ACM Multimedia*. 6

[19] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. *Proceedings of the ACM on international conference on multimodal interaction (ICMI)*, 2015. 2

[20] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli, and Thomas S. Huang. How deep neural networks can improve emotion recognition on video data. *IEEE International Conference on Image Processing (ICIP)*, 2016. 2

[21] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[23] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. 2

[24] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. 2

[25] Nazil Perveen and Chalavadi Mohan. Configural representation of facial action units for spontaneous facial expression recognition in the wild. *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2020. 4

[26] Andres Romero, Juan Leon, and Pablo Arbelaez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, 2017. 2

[27] Shuvendu Roy and Ali Etemad. Self-supervised contrastive

learning of multi-view facial expressions. *ACM International Conference on Multimodal Interaction (ICMI)*, 2021. 2

[28] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. *Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020. 4

[29] Abhinav Shukla, Stavros Petridis, and Maja Pantic. Does visual self-supervision improve learning of speech representations for emotion recognition? *IEEE Transactions on Affective Computing*, 2020. 1

[30] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *journal = Advances in Neural Information Processing Systems (NeurIPS),*, 2022. 2, 6

[31] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann Lecun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[32] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2021. 5

[33] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[34] AmirAli Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 2, 5

[35] Tenggan Zhang, Chuanhe Liu, Xiaolong Liu, Yuchen Liu, Liyu Meng, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Jinming Zhao, and Qin Jin. Multi-task learning framework for emotion recognition in-the-wild. *European Conference on Computer Vision Workshop (ECCVW)*, 2022. 2

[36] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a facial expression embedding disentangled from identity. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[37] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 2