

# MMA-DFER: MultiModal Adaptation of unimodal models for Dynamic Facial Expression Recognition in-the-wild

Kateryna Chumachenko  
Tampere University  
Tampere, Finland

kateryna.chumachenko@tuni.fi

Alexandros Iosifidis  
Aarhus University  
Aarhus, Denmark

ai@ece.au.dk

Moncef Gabbouj  
Tampere University  
Tampere, Finland

moncef.gabbouj@tuni.fi

## Abstract

*Dynamic Facial Expression Recognition (DFER) has received significant interest in the recent years dictated by its pivotal role in enabling empathic and human-compatible technologies. Achieving robustness towards in-the-wild data in DFER is particularly important for real-world applications. One of the directions aimed at improving such models is multimodal emotion recognition based on audio and video data. Multimodal learning in DFER increases the model capabilities by leveraging richer, complementary data representations. Within the field of multimodal DFER, recent methods have focused on exploiting advances of self-supervised learning (SSL) for pre-training of strong multimodal encoders [40]. Another line of research has focused on adapting pre-trained static models for DFER [8]. In this work, we propose a different perspective on the problem and investigate the advancement of multimodal DFER performance by adapting SSL-pre-trained disjoint unimodal encoders. We identify main challenges associated with this task, namely, intra-modality adaptation, cross-modal alignment, and temporal adaptation, and propose solutions to each of them. As a result, we demonstrate improvement over current state-of-the-art on two popular DFER benchmarks, namely DFEW [19] and MFAW [29].*

## 1. Introduction

The ability to perceive non-verbal communication cues is essential for development of truly intelligent interactive technologies. Automatic understanding of human emotional states, whether in the form of facial expressions, voice characteristics, or language semantics, sees a rapid development in the recent years with adoption in a vast number of applications. These include, among others, collaborative robotics [32, 42], healthcare [3, 6] and human-computer interaction [9, 49].

The prevailing signal for this task has conventionally

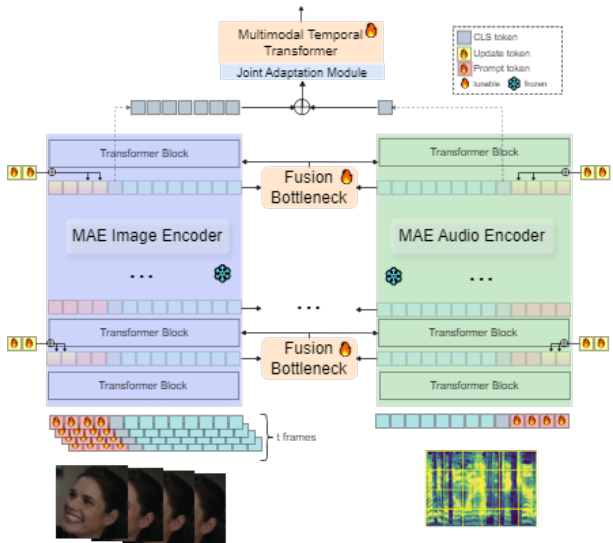


Figure 1. Schematic description of MMA-DFER: Two pre-trained frozen MAE encoders are joined by Fusion Bottleneck for modality alignment, followed by joint adaptation module and Multimodal Temporal Transformer. Learnable prompts in each modality independently handle intra-modality gaps between pre-training and downstream data.

been the vision modality in the form of facial expressions. The field of facial expression recognition has undergone an evolution from solving static expression recognition problems (SFER) on images captured in controlled environments, to dynamic facial expression recognition (DFER), and further the multimodal dynamic facial expression recognition (most commonly taking the form of audio-visual emotion recognition - AVER). Recently, the field has seen an increased interest towards in-the-wild analysis, facilitated by the introduction of large-scale datasets firstly in static domain [36], and further in dynamic and multimodal domain [19, 29]. Methods capable of in-the-wild analysis enable development of novel applications, and ability to

handle dynamic scenes increases robustness towards real-world scenarios.

The amount of data surrounding us is vast, and data relevant to facial expression recognition extends beyond the visual domain. Additional data inputs, such as audio or language can aid in the analysis by providing additional information about the signal. This idea has gained momentum also in the DFER field and inspired numerous multimodal facial expression recognition methods.

Within the multimodal DFER, a large portion of existing works has focused on designing elaborate fusion methods and relying on joint learning of multimodal features that utilizes each modality to its fullest. At the same time, closing the gap between DFER in constrained environments and in-the-wild DFER requires the ability to generalize to wider set of data distributions and to be robust to various challenges and variable factors associated with in-the-wild data. In the context of multimodal DFER, collection of large amounts of multimodal data representative of emotion labels poses a challenging task.

These facts naturally invite the application of self-supervised learning (SSL) methods, that are able to learn meaningful features from data without requiring labels (followed by downstream fine-tuning for the task in hand). Indeed, previous works aiming at in-the-wild DFER have already shown benefits of such approach in both unimodal [38, 39] and multimodal [40] settings. Specifically in the multimodal space, HiCMAE [40] learns a joint audiovisual model by large-scale multimodal SSL pre-training, followed by full fine-tuning for DFER.

At the same time, an abundance of off-the-shelf unimodal foundational models are publicly available, and progress made in unimodal domains generally precedes their adoption in multimodal space by a significant margin, with multimodal extensions often being ad-hoc and not easily transferable. We find ourselves wishing for universal feature extractors that can be mix-and-matched for multimodal inference without the loss of performance on downstream tasks and without requiring additional multimodal pre-training. In this work, we aim to address the gap in multimodal adaptation of in-the-wild DFER from only publicly available unimodal foundational models, pre-trained independently, and on unrelated datasets. We show that with appropriate adaptation, we can obtain beyond state-of-the-art results on two popular DFER benchmarks.

Our contributions can be summarized as follows:

- We show that orthogonally to the works proposed in recent literature, state-of-the-art performance on multimodal dynamic face recognition can be achieved without a) large-scale paired multimodal pre-training; b) pre-training for static facial expression recognition
- We identify key challenges in adaptation of pre-trained models for multimodal DFER and propose solution to

each of them, showing their efficacy. Specifically, this includes progressive prompt tuning for bridging intra-modality gap, Fusion Bottleneck blocks for cross-modal alignment, and Multimodal Temporal Transformer for temporal alignment

- We set new SOTA on two popular in-the-wild DFER benchmarks, namely DFEW and MAFW.

## 2. Related Work

### 2.1. Dynamic Facial Expression Recognition in-the-wild

For a period of time, the field of emotion recognition or facial expression recognition has largely relied on datasets collected in controlled environments and methods designed for them, often leading to suboptimal performance in real-world applications. In the recent years, dynamic facial expression recognition in-the-wild has emerged as a separate task within the paradigm of affective behavior analysis, thanks to the appearance of a number of large-scale in-the-wild datasets [19, 22, 29]. Recently, a number of novel methods has emerged in this area. Among them, there is an ongoing trend of reliance on large-scale models performing large-scale pre-training, often in a self-supervised manner. Specifically, DFER-CLIP [54] relies on joint text-image space of CLIP for mapping dynamic videos to emotion labels; SVFAP [39] and MAE-DFER [38] explore Video-MAE-like masked reconstruction pre-training on dataset of face videos; HiCMAE [40] extends this idea to multimodal inputs. Our work differs from these approaches in a way that we explore adaptation of pre-trained unimodal models for multimodal DFER without multimodal pre-training.

Another line of work, most similar to ours in spirit, is S2D [8] that explores temporal expansion of pre-trained static facial expression recognition model for dynamic recognition, and employs temporal adaptation modules, and landmark-guided modules to achieve competitive results. We pose a similar question in the multimodal domain, and show that we can outperform S2D with 20% less trainable parameters (both ours and S2D vision encoder follows ViT-b architecture), while accommodating another modality and not requiring static facial expression recognition pre-training, only self-supervised pre-training in each modality independently.

### 2.2. Adaptation of pre-trained models

End-to-end fine-tuning, or fine-tuning with a frozen backbone has been a dominant approach in the task of adaptation of large-scale pre-trained models for downstream tasks. Recently, dictated by the adoption of Large Language Models, a different paradigm emerged, namely prompt tuning, primarily targeted at Transformer models. In the NLP domain, prompt tuning was introduced as a way to adapt the model

for a downstream task, while requiring substantially less trainable parameters by introduction of additional tokens in the input space, that are concatenated to the input, but are optimized via backpropagation [26–28]. Visual prompt tuning, introduced in [18] extends this idea to the vision domain, where learnable tokens are concatenated to the input sequence of patches of a pre-trained model. A few works aim at cross-modal adaptation via prompt tuning [20] (targeted at VLMs), where prompts are introduced in multiple input modalities, and interact throughout the model.

While tackling a multimodal task, our adoption of prompt tuning is aimed at each modality independently, with the goal of reducing the domain shifts in each of the unimodal foundation models. Moreover, we introduce the idea of progressive prompt adaptation, aimed at adapting the model at multiple levels of granularity independently, and therefore simplifying the adaptation task.

Another challenge differentiating DFER from static facial expression recognition from images is the need to learn temporal dependencies in the data, and identify most relevant ones. To address this challenge, prior works have either employed temporally-aware pre-training on videos [38, 40] or performed temporal adaptation of pre-trained static models at the fine-tuning stage [8, 54]. A natural choice for the latter approach is often a form of temporal self-attention, introduced in different stages of feature extraction. Working with a static image encoder, we also follow a similar temporal self-attention-based approach and propose a Multimodal Temporal Transformer for temporal information extraction. We also evaluate its efficacy at different stages of the model.

In the field of multimodal facial expression recognition, prior works have to a large extent focused on joint cross-modal feature learning through complex fusion methods and/or multimodal pre-training [1, 10, 35, 37, 40, 41]. The fusion methods are often ad-hoc and range from simple concatenation, to multimodal fusion Transformers or other elaborate frameworks. We are instead focusing on adapting already pre-trained unimodal models via lightweight bottleneck fusion adaptors while preserving their unimodal feature extraction capabilities.

### 3. Methodology

Our method operates by employing two off-the-shelf SSL-pre-trained unimodal encoders trained independently on audio sequences and static images, and adapts them for audio-visual dynamic facial expression recognition in-the-wild. Specifically, we choose among the models pre-trained with Masked Autoencoder (MAE) reconstruction objective, dictated by outstanding performance of MAEs on tasks requiring fine-grained features [14]. Following this, as foundation models we employ two publicly available Vision Transformer-based encoders, namely AudioMAE [17] and MAE-Face [33] that both follow ViT-base [11] architec-

ture and therefore have the same depth. AudioMAE is pre-trained on AudioSet [12] and MAE-Face is pre-trained on a combination of static face image datasets [33].

While large-scale pre-training of unimodal encoders ensures rich and to an extent generalizable representation within each modality, it is associated with certain challenges when adapting to multimodal DFER. These challenges can be distinguished into three categories:

- **Within-modality domain gap:** naturally, although pre-trained on large-scale data, certain level of domain gap can be expected between the data distributions of the pre-trained models and the downstream domains, especially if the pre-training datasets have little relevance to DFER.
- **Modality alignment gap:** considering a model where two encoders are pre-trained without awareness of another modality, and potentially on different data sets, there exists a gap between the distributions of feature spaces of the two modalities, both on the decision-level, and on the intermediate feature level.
- **Temporal adaptation gap:** considering a pre-trained model that was trained on static images, adaptation to dynamic facial expression recognition could benefit from learning temporal dependencies between the frames.

In the following sections, we discuss how to address each of these gaps.

#### 3.1. Formulation

Let us start with formally defining a problem. Each data sample is represented by a pair of audio sequence and frame sequence and corresponds to a single emotion class label. Each frame is initially split into patches independently, following the Transformer embedding layer [11, 46]. Frames corresponding to the same video are concatenated batch-wise and we describe their temporal interaction within the batch further. Similarly, a spectrogram is extracted from audio sequence and split into patches, too, following the AudioMAE [17] procedure.

#### 3.2. Progressive prompt learning

First, we propose to address the domain gap within each modality encoder independently. To achieve this, we introduce a set of learnable prompts for each modality that are concatenated to the data sequences and are updated via backpropagation. As the tokens are processed by the model, learnable prompts interact with the data tokens and are able to divert their feature representation distribution so that it is closer to the initial distribution of the data on which the model was trained.

Specifically, learnable prompts are a set of  $M$  tokens  $\mathbf{p}^m$  in  $\mathcal{R}^d$ , where  $d$  is a hidden dimension of the encoder, and the learnable prompts are concatenated to the input sequence following the embedding layer.

We note that Transformer adaptation via learnable

prompts has been shown to be successful in few application areas. However, in the prior work, this idea is applied purely on the input space level. At the same time, discrepancies in feature distributions at different depths can have different nature, and it can be difficult to address all of them with the same prompt on the input level. Instead, applying dedicated prompts at different depths can aid the adaptation. Therefore, we propose *progressive prompt adaptation*, where we introduce a set of  $M^l$  tokens at different depths of the model, that complement the initial  $M$  learnable prompt tokens and are introduced to the network progressively. Formally, given  $M$  learnable prompt tokens  $\mathbf{P} \in \mathcal{R}^{M \times d}$ , and  $L$  intermediate layers, we introduce  $L$  sets of tokens  $\mathbf{P}^l \in \mathcal{R}^{M^l \times d}$ ,  $M^l = \frac{M}{L}$  and at  $l^{\text{th}}$  layer, the prompts are updated as:

$$\mathbf{P}[(l-1) * M^l : l * M^l, :] = \mathbf{P}[(l-1) * M^l : l * M^l, :] + \mathbf{P}^l. \quad (1)$$

This way, adaptation of the model to new data distribution happens progressively thereby simplifying the task.

### 3.3. Modality fusion bottlenecks

To exploit cross-modal dependencies for knowledge extraction, we introduce fusion of modalities at multiple intermediate stages. Our Modality Fusion blocks are based on creating a low-dimensional bottleneck where multimodal features are obtained by fusing compressed and normalized representations of each modality, and routing them back to the corresponding unimodal branches via a gating mechanism. This approach relates to principles of learning to control the amount of transferred information, as well as information compression followed by its expansion for highlighting the most relevant features, that have been successfully applied as building blocks of different methods in the field [2, 4, 16, 21]. Our formalization in modality fusion allows to flexibly adapt pre-trained unimodal models. A schematic representation is provided in the Figure 2.

Given a video representation  $\mathbf{V} = \{\mathbf{v}\}_t$  corresponding to  $t$  frames, and an audio sequence  $\mathbf{A}$ , we first project each of them to a low-dimensional latent space to obtain corresponding low-dimensional representations  $\hat{\mathbf{V}}$  and  $\hat{\mathbf{A}}$ :

$$\begin{aligned} \hat{\mathbf{V}} &= \{\mathcal{G}_V(\mathbf{v})\}_t, \\ \hat{\mathbf{A}} &= \mathcal{G}_A(\mathbf{A}), \end{aligned} \quad (2)$$

where  $\mathcal{G}_V$  and  $\mathcal{G}_A$  consist of a Linear layer projecting from high- to low-dimensional space, and a Layer Normalization layer. Further, each low-dimensional representation is followed by a corresponding aggregation function to obtain global sequence representation in each modality. For audio modality, this is represented by a Global Average Pooling over tokens of audio sequence, and for vision modality, by Global Average Pooling over tokens of all image sequences

in all frames corresponding to one video:

$$\mathbf{L}_v = \frac{\sum_{i=1}^t \sum_{j=1}^{N_v} \hat{\mathbf{V}}_i^j}{t * N_v} \quad \text{and} \quad \mathbf{L}_a = \frac{\sum_{j=1}^{N_a} \hat{\mathbf{A}}^j}{N_a}. \quad (3)$$

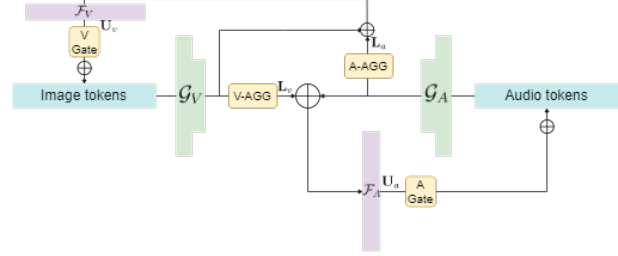


Figure 2. Fusion bottleneck. The features of each modality are first compressed to a low-dimensional representation, aggregated to obtain global representation per sequence, then fused with the complementary modality, and the joint representation is expanded back to the original space, and added to the initial features via a gating mechanism.

Having obtained global low-dimensional representations for each modality, we fuse them via addition to the opposite modality (unaggregated), and employ upsampling functions  $\mathcal{F}_A$  and  $\mathcal{F}_V$  to expand the joint representations back to the original dimensionality spaces. In our implementation,  $\mathcal{F}_A$  and  $\mathcal{F}_V$  are represented by a Linear layer, followed by a GELU activation:

$$\begin{aligned} \mathbf{U}_a &= \mathcal{F}_A(\hat{\mathbf{A}} + \mathbf{L}_v), \\ \mathbf{U}_v &= \mathcal{F}_V(\hat{\mathbf{V}} + \mathbf{L}_a). \end{aligned} \quad (4)$$

Finally, the obtained fused representations are added to the original ones via gated skip-connections, where we employ a learnable parameter  $\alpha$  that controls the strength of multimodal representations:

$$\begin{aligned} \tilde{\mathbf{A}} &= \mathbf{A} + \tanh(\alpha) * \mathbf{U}_a, \\ \tilde{\mathbf{V}} &= \mathbf{V} + \tanh(\alpha) * \mathbf{U}_v. \end{aligned} \quad (5)$$

We initialize  $\alpha$  to zero such that the modality encoders receive the original (unimodal) feature representations in the first iterations of the training, and the appropriate magnitude of multimodal features is progressively learnt by the model, hence making adaptation to multimodal features easier.

As a result, compression of the features of each modality to a low-dimensional latent space helps highlight the most relevant information in each modality, and further expand it back to the corresponding branch, while keeping this process adaptive to the strength of each modality via gating.

### 3.4. Temporal fusion

The task of dynamic facial expression recognition can benefit from knowledge about between-frame temporal dependencies. However, a model that was pre-trained on static

images lacks such capability. We address this challenge by employing a Multimodal Temporal Transformer.

Recall that in the image encoder, given a sequence of  $t$  frames of a video, we concatenate them in a batch manner and process independently in a parallel manner. Following this stage, we extract the  $[CLS]$  token of every frame corresponding to the same video and concatenate them to form a temporal sequence. Therefore, the input tensor in the vision branch is reshaped from  $B * t \times N \times D$  to  $B \times t \times D$ , where  $B$  is the batch size,  $t$  is the temporal length (number of frames),  $N$  is the sequence length of each frame (number of patches), and  $D$  is the dimensionality. We additionally fuse the corresponding  $[CLS]$  token of audio branch to the video sequence via addition, and process the new multimodal sequence with a Joint Adaptation Module, which in our implementation is formalized as a single Linear layer (with slightly reduced dimensionality than prior encoders). Further, we add learnable temporal embeddings to the multimodal temporal sequence and concatenate a new  $[CLS]$  token. We process the new sequence with a Transformer block, which is now operating on temporal level on joint multimodal sequence, hence we refer to it as Multimodal Temporal Transformer. The  $[CLS]$  token is further used as input to the classifier.

We also note that some works have suggested adoption of temporal adaptation modules on intermediate feature level [8], with the aim of enriching image features with temporal information already at the feature extraction stage. We perform ablation studies on the placement of our Multimodal Temporal Transformer with intermediate temporal modules in experimental section, and find our approach to be beneficial. This can be partially attributed to fusion modules that already implicitly embed a certain level of temporal information at intermediate stages by fusing an audio sequence to image patches.

## 4. Experiments

### 4.1. Datasets

We benchmark our method on two popular multimodal dynamic facial expression recognition in-the-wild datasets, namely DFEW [19] and MAFW [29]. DFEW consists of 16,000 audiovisual clips split into 5 folds, posing a 7-class classification task, with classes being emotion labels of ‘happy’, ‘sad’, ‘angry’, ‘neutral’, ‘surprise’, ‘disgust’, and ‘fear’. MAFW contains 10,045 clips and follows a 5-fold split as well. In this dataset set of emotions follows 11 classes: ‘anger’, ‘disgust’, ‘fear’, ‘happiness’, ‘sadness’, ‘surprise’, ‘contempt’, ‘anxiety’, ‘helplessness’, ‘disappointment’, and ‘neutral’.

### 4.2. Experimental setup

We follow the traditional setup of extracting 16 frames to form a video sequence during training [8, 40]. We notice the discrepancies in image sizes in the experimental setup of existing methods, therefore report results on 112x112 images [29], 160x160 images [40], and on 224x224 images [8]. For MAFW, we extract faces with MTCNN [50]. We use learning rate of 1e-4 that is annealed via cosine schedule to 0 over 25 epochs. We use batch size of 8 and weight decay of 1e-2, and AdamW optimizer with default parameters. We fix the random seed to 1. In unimodal encoders, we interpolate the positional embeddings to new sequence lengths and keep them tunable to adapt to the new resolutions, the rest of unimodal encoder parameters remain frozen. Therefore, tunable parameters include the Fusion Bottleneck blocks, Multimodal Temporal Transformer, learnable prompts, classifier and positional embeddings, totaling 7.5M parameters. For comparison, S2D that follows a similar framework of adapting pre-trained (for static emotion recognition in their case) network, includes 9M tunable parameters.

Multimodal Temporal Transformer is a 1-layer Transformer with hidden dimension of 512 and 8 heads. Fusion Bottleneck latent dimensionality is set to 128, and 6 learnable prompts are introduced in each modality, and progressive updates are introduced twice in the network, after 1st and 7th layers, with 3 tokens each.

We follow the 5-fold experimental protocol in each dataset with the provided splits. We train the models on the train set and report the result of final checkpoint, i.e., at 25th epoch, on the test set. Training is done on a single Nvidia Tesla V-100-32 GPU, and single training (of single fold) takes approximately 8 hours on resolution 160. We report results both with following the 16-frame uniform sampling at inference time, as well as widely adopted 2-clip average results [8, 40], where two clips are sampled uniformly from a single video and predictions are averaged. We note that we do not observe a significant difference between these two approaches. As prior works, we report Unweighted Average Recall (UAR) and Weighted Average Recall (WAR).

### 4.3. Results

The comparison to state-of-the-art methods is provided in Table 1. As can be seen, MMA-DFER outperforms the competing methods. Specifically, MMA-DFER outperforms current state-of-the-art of S2D by 1.5% UAR and WAR on DFEW dataset and 1% on MAFW. We also note that for S2D, the best UAR and WAR are obtained from different models/training strategies (with and without oversampling of underrepresented classes), while in MMA-DFER this is achieved by a single model. Compared to the best multimodal model - HiCMAE, we achieve stronger results as well, both on 224 and 160 resolution. With the same image resolution, we obtain 2-3% improvement on DFEW

Method	DFEW		MAFW		M	Res.
	UAR	WAR	UAR	WAR		
Wav2Vec2.0 [5]	36.15	43.05	21.59	29.69	A	-
HuBERT [15]	35.98	43.24	25.00	32.60	A	-
WavLM-Plus [7]	37.78	44.64	26.33	34.07	A	-
C3D+LSTM [29]	53.77	65.17	30.47	44.15	AV	224
T-ESFL [29]	-	-	33.35	48.70	AV	224
C3D [43]	42.74	53.54	31.17	42.25	V	112
R(2+1)D-18 [44]	42.79	53.22	-	-	V	112
3D ResNet-18 [13]	46.52	58.27	-	-	V	112
Former-DFER [52]	53.69	65.70	-	-	V	112
CEFLNet [30]	51.14	65.35	-	-	V	224
T-ESFL [29]	-	-	33.28	48.18	V	224
EST [31]	53.43	65.85	-	-	V	224
STT [34]	54.58	66.65	-	-	V	112
NR-DFERNet [23]	54.21	68.19	-	-	V	112
AMH [48]	54.48	66.51	32.98	48.83	AV	224
IAL [24]	55.71	69.24	-	-	V	112
M3DFEL [47]	56.10	69.25	-	-	V	-
CLIPER [25]	57.56	70.84	-	-	V	224
TMEP [51]	57.16	68.85	37.17	51.15	AV	112
DFER-CLIP [53]	59.61	71.25	38.89	52.55	V	224
SVFAP [39]	62.83	74.27	41.19	54.28	V	160
MAE-DFER [38]	63.41	74.43	41.62	54.31	V	160
HiCMAE [40]	63.76	75.01	42.65	56.17	AV	160
S2D [8]	65.45	76.03	43.40	57.37	V	224
MMA (ours).112	64.24	75.40	43.34	56.46	AV	112
MMA (ours).112*	64.35	75.48	43.29	56.60	AV	112
MMA (ours).160	66.61	77.15	44.19	57.90	AV	160
MMA (ours).160*	66.51	77.10	44.07	57.85	AV	160
MMA (ours).224	<b>67.01</b>	<b>77.51</b>	44.11	<b>58.52</b>	AV	224
MMA (ours).224*	66.85	77.43	<b>44.25</b>	58.45	AV	224

Table 1. Comparison to SOTA methods. \* denotes mean prediction over two uniformly sampled video clips following [8, 40]. M denotes the modality, and Res denotes the image resolution.

and 1.5% on MAFW.

#### 4.4. Comparison of temporal adaptation approaches

In MMA, the Multimodal Temporal Transformer is placed following the unimodal branches. However, some works have suggested a different approach, where temporal adaptation happens already in the intermediate features of the model [8]. We perform an experiment where we add such temporal adaptors in the intermediate steps of the network, following each Bottleneck Fusion block. Such temporal adaptors are similar to TMA described in [8] and in practice we apply temporal self-attention over  $[CLS]$  tokens of every frame and fuse the obtained representation back to the vision branch. We evaluate intermediate temporal adaptors (ITA), Multimodal Temporal Transformer (MTM), and the combination of the two. In the case where only intermediate temporal adaptors are applied, we control for the total num-

Temporal adaptation	UAR	WAR	Params (M)
ITA, d=64	63.27	76.59	1.3
ITA, d=128	64.44	76.80	0.8
ITA, d=256	63.35	76.59	4.3
MTM + ITA-128	64.66	77.36	2.9
MTM (ours)	<b>66.52</b>	<b>77.92</b>	2.2

Table 2. Comparison of Multimodal Temporal Transformer vs intermediate temporal blocks

ber of parameters and therefore each intermediate block is smaller than our Multimodal Temporal Transformer. We also provide results with different dimensionalities of intermediate temporal adaptors. We compare the results on 1st fold of DFEW on 160x160 resolution.

The results can be seen in Table 2, where we report UAR, WAR, and number of parameters that correspond to temporal processing blocks. We note that the variant with ITA of dimensionality 128 has less parameters than ITA-64 due to internal dimensionality of Bottleneck Fusion already being 128, therefore no additional layers need to be introduced to this variant to project to new dimensionality space, unlike all other variants. We can see that the best result is achieved by MTM and the second-best by MTM combined with ITA. Among ITA, we observe the variant with d=128 to be outperforming the competing ones. We also note that the performance is not directly correlated with the number of parameters, but more with the placement of the temporal module.

#### 4.5. Comparison of fusion approaches

We additionally compare our Fusion Bottleneck blocks with other popular multimodal fusion approaches. Specifically, we compare with:

- Addition in original space, i.e., without compression and expansion (ADD);
- Multimodal Transformer following [45], where we introduce 2 multi-head self-attention blocks, audio-to-vision, and vision-to-audio, outputs of which are added to corresponding branches (MULT);
- Single multimodal Transformer on concatenated audio and image features, with global average pooled representation added back to each branch (MULT-concat).
- No modality fusion;

For MULT, we utilize 2 heads with total dimensionality of 128, to match the Fusion Bottleneck dimensionality. We compare the results on 1st fold of DFEW on 160x160 resolution.

The results can be seen in Table 3. As can be seen, our approach outperforms the competing methods by a significant margin, indicating the effectiveness of the proposed Fusion Bottlenecks. Poor performance of MULT and MULT-concat can be associated with difficulty of drawing

fusion methods	UAR	WAR
None	59.27	70.56
MULT	59.89	71.12
MULT-concat	54.96	65.66
ADD	59.97	71.20
Fusion Bottleneck (ours)	<b>66.52</b>	<b>77.92</b>

Table 3. Comparison of modality fusion approaches.

Pr	Pr.Pr.	MTT	FB	UAR	WAR
-	-	-	-	56.59	68.43
+	-	-	-	57.45	69.33
+	+	-	-	58.21	69.34
+	+	+	-	59.27	70.56
+	+	-	+	61.96	74.58
-	-	+	+	64.62	77.40
+	+	+	+	<b>66.52</b>	<b>77.92</b>

Table 4. Ablation studies

dependencies between individual frames and whole audio spectrogram.

#### 4.6. Ablation studies

Further, we ablate each component of our model independently, with the results shown in Table 4. Here, ‘Pr.’ corresponds to learnable prompts that are not updated with depth but the initialization is kept; ‘Pr.Pr.’ corresponds to the alternative with the progressive prompts, i.e., new learnable prompts are introduced at different depths; ‘MTT’ corresponds to Multimodal Temporal Transformer, and ‘FB’ corresponds to Fusion Bottleneck. When no MTT is employed, prediction is done on averaged features following Joint Adaptation Module. Joint Adaptation Module and a classifier are present and unfrozen in all variants. We report the results on first fold of DFEW.

As can be seen, performance of plain pre-trained models is rather poor, and each of the components progressively improves the performance. The biggest effect is brought by Fusion Bottleneck which improves WAR by 5% and UAR by 3.7% if comparing variants without MTT. We also note how performance increases by using MTT by 1.5%, but when used in conjunction with Fusion Bottlenecks, MTT improves the performance by 3.5%. This shows that fusion of features on intermediate level helps late-stage adaptation that precedes temporal modeling.

We further study the effect of each modality independently. Here, in the unimodal cases each branch is frozen and only the classifier is updated. The results are shown in Table 5. As can be seen, performance of each individual modality is significantly lower than the combination.

We additionally study the dimensionality of the bottleneck space. Recall that dimensionality of both audio and

	UAR	WAR
Audio only	26.43	33.53
Vision only	54.34	67.15
MMA-DFER	<b>66.52</b>	<b>77.92</b>

Table 5. Comparison of each modality encoder to the multimodal MMA-DFER

latent space d	UAR	WAR	Params (M)
64	64.57	77.44	5.9
128	66.52	77.92	7.5
256	66.72	<b>78.17</b>	12.3
512	<b>66.77</b>	78.13	21.7

Table 6. Dimensionality of the latent space

Pr.Pr. Num.	UAR	WAR
12	65.14	77.87
6	65.44	77.83
4	64.41	77.62
2	<b>66.52</b>	<b>77.92</b>
0	64.62	77.40

Table 7. Ablation on the frequency of progressive prompts

image encoders in ViT-base are 768 [11], and we set our bottleneck space to 128. Here, we evaluate the effect of larger and smaller dimensionalities of the latent space. The results are shown in Table 6. As can be seen, larger values do increase the performance of the model, but require significantly more parameters. We also find that increasing dimensionality beyond d-128 brings rather diminishing returns in terms of UAR and WAR.

Finally, we also study the frequency of prompt updates in the case of progressive prompting. The results are shown in Table 7. Here, 12 corresponds to introducing new learnable prompts every layer, 6 every 2nd layer, 2 every 3rd layer, 2 every 6th layer (our final model case), and 0 not using learnable prompts at all. As can be seen, any number of updates results in better accuracy than no prompts, with the best result achieved when introducing them twice.

## 5. Conclusion

We have investigated adaptation of pre-trained unimodal models for multimodal dynamic facial expression recognition in-the-wild. We identified key limitations associated with adapting pre-trained models for this task, namely intra-modality adaptation, cross-modal alignment, and temporal adaptation, and proposed solutions to address them. Our proposed model, MMA-DFER sets a new state-of-the-art on two popular DFER benchmarks DFEW and MAFW. Future work could include experimentation with additional unimodal backbones and exploitation of further modali-

ties/sensors, such as landmarks or vision-language latent spaces.

## 6. Acknowledgement

This work was partially funded by the NSF CBL and Business Finland project AMALIA, and by the Horizon Europe programme PANDORA (GA 101135775).

## References

- [1] Naveed Ahmed, Zaher Al Aghbari, and Shini Giriya. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171, 2023. [3](#)
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. [4](#)
- [3] Değer Ayata, Yusuf Yaslan, and Mustafa E Kamasak. Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. *Journal of Medical and Biological Engineering*, 40:149–157, 2020. [1](#)
- [4] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, pages 1352–1361. PMLR, 2021. [4](#)
- [5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. [6](#)
- [6] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics*, 18(8):5619–5627, 2022. [1](#)
- [7] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. [6](#)
- [8] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *arXiv preprint arXiv:2312.05447*, 2023. [1](#), [2](#), [3](#), [5](#), [6](#)
- [9] M Kalpana Chowdary, Tu N Nguyen, and D Jude Hemanth. Deep learning-based facial emotion recognition for human-computer interaction applications. *Neural Computing and Applications*, 35(32):23311–23328, 2023. [1](#)
- [10] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Self-attention fusion for audiovisual emotion recognition with incomplete data. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2822–2828. IEEE, 2022. [3](#)
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [3](#), [7](#)
- [12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. [3](#)
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [6](#)
- [14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. [3](#)
- [15] WN Hsu, B Bolte, YHH Tsai, K Lakhota, R Salakhutdinov, and AH Mohamed. Self-supervised speech representation learning by masked prediction of hidden units. *IEEE Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021. [6](#)
- [16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. [4](#)
- [17] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022. [3](#)
- [18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. [3](#)
- [19] 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. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889, 2020. [1](#), [2](#), [5](#)
- [20] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. [3](#)
- [21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. [4](#)
- [22] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection and emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023. [2](#)
- [23] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet:



- Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022. 6
- [24] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 67–75, 2023. 6
- [25] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. *arXiv preprint arXiv:2303.00193*, 2023. 6
- [26] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3
- [27] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35, 2023.
- [28] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 3
- [29] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 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 (MM'22)*. 1, 2, 5, 6
- [30] Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhongwen Luo. Clip-aware expressive feature learning for video-based facial expression recognition. *Information Sciences*, 598:182–195, 2022. 6
- [31] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023. 6
- [32] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE CAA J. Autom. Sinica*, 4(4):668–676, 2017. 1
- [33] Bowen Ma, Wei Zhang, Feng Qiu, and Yu Ding. A unified approach to facial affect analysis: the mae-face visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5923–5932, 2023. 3
- [34] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022. 6
- [35] Bogdan Mocanu, Ruxandra Tapu, and Titus Zaharia. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing*, 133:104676, 2023. 3
- [36] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 1
- [37] Bei Pan, Kaoru Hirota, Zhiyang Jia, and Yaping Dai. A review of multimodal emotion recognition from datasets, pre-processing, features, and fusion methods. *Neurocomputing*, page 126866, 2023. 3
- [38] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023. 2, 3, 6
- [39] Licai Sun, Zheng Lian, Kexin Wang, Yu He, Mingyu Xu, Haiyang Sun, Bin Liu, and Jianhua Tao. Svfp: Self-supervised video facial affect perceiver. *arXiv preprint arXiv:2401.00416*, 2023. 2, 6
- [40] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Hic-mae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *arXiv preprint arXiv:2401.05698*, 2024. 1, 2, 3, 5, 6
- [41] Mani Kumar Tellamekala, Shahin Amiriparian, Björn W Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3
- [42] Aitor Toichoa Eyam, Wael M Mohammed, and Jose L Martinez Lastra. Emotion-driven analysis and control of human-robot interactions in collaborative applications. *Sensors*, 21(14):4626, 2021. 1
- [43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 6
- [44] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6
- [45] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, page 6558. NIH Public Access, 2019. 6
- [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3
- [47] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17958–17968, 2023. 6
- [48] Seunghyun Yoon, Subhadeep Dey, Hwanhee Lee, and Kyomin Jung. Attentive modality hopping mechanism for speech emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3362–3366. IEEE, 2020. 6

- [49] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W Picard. Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–30, 2020. [1](#)
- [50] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. [5](#)
- [51] Xiaoqin Zhang, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. [6](#)
- [52] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. [6](#)
- [53] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*, 2023. [6](#)
- [54] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*, 2023. [2](#), [3](#)