Micron-BERT: BERT-based Facial Micro-Expression Recognition

Xuan-Bac Nguyen[1], Chi Nhan Duong[2], Xin Li[3], Susan Gauch[1], Han-Seok Seo[4], Khoa Luu[1]
[1] CVIU Lab, University of Arkansas, USA [2] Concordia University, Canada
[3] West Virginia University, USA [4] Dep. of Food Science, University of Arkansas, USA
[1]{xnguyen, sgauch, hanseok, khoaluu}@uark.edu, [2]{dcnhan@ieee.org}, [3]xin.li@mail.wvu.edu

Abstract

Micro-expression recognition is one of the most challenging topics in affective computing. It aims to recognize tiny facial movements difficult for humans to perceive in a brief period, i.e., 0.25 to 0.5 seconds. Recent advances in pre-training deep Bidirectional Transformers (BERT) have significantly improved self-supervised learning tasks in computer vision. However, the standard BERT in vision problems is designed to learn only from full images or videos, and the architecture cannot accurately detect details of facial micro-expressions. This paper presents Micron-BERT (μ-BERT), a novel approach to facial micro-expression recognition. The proposed method can automatically capture these movements in an unsupervised manner based on two key ideas. First, we employ Diagonal Micro-Attention (DMA) to detect tiny differences between two frames. Second, we introduce a new Patch of Interest (PoI) module to localize and highlight micro-expression interest regions and simultaneously reduce noisy backgrounds and distractions. By incorporating these components into an end-to-end deep network, the proposed μ-BERT significantly outperforms all previous work in various micro-expression tasks. μ-BERT can be trained on a large-scale unlabeled dataset, i.e., up to 8 million images, and achieves high accuracy on new unseen facial micro-expression datasets. Empirical experiments show μ-BERT consistently outperforms state-of-the-art performance on four micro-expression benchmarks, including SAMM, CASME II, SMIC, and CASME3, by significant margins. Code will be available at https://github.com/uark-cviu/Micron-BERT

1. Introduction

Facial expressions are a complex mixture of conscious reactions directed toward given stimuli. They involve experiential, behavioral, and physiological elements. Because they are crucial to understanding human reactions, this topic has been widely studied in various application domains [5]. In general, facial expression problems can be classified into two main categories, macro-expression, and micro-expression. The main differences between the two are facial expression intensities, and duration [2]. In particular, macro-expressions happen spontaneously, cover large movement areas in a given face, e.g., mouth, eyes, cheeks, and typically last from 0.5 to 4 seconds. Humans can usually recognize these expressions. By contrast, micro-expressions are involuntary occurrences, have low intensity, and last between 5 milliseconds and half a second. Indeed, micro-expressions are challenging to identify and are mostly detectable only by experts. Micro-expression understanding is essential in numerous applications, primarily lie detection, which is crucial in criminal analysis.

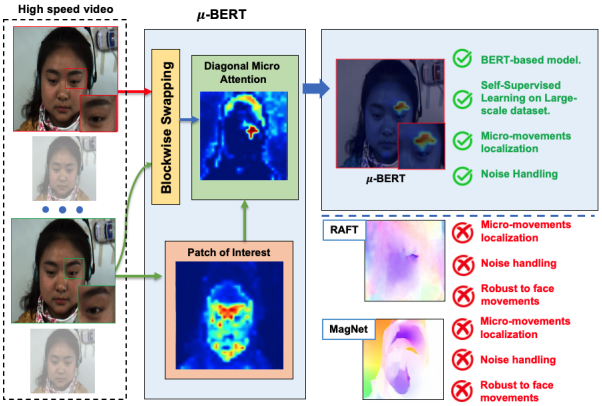

Figure 1. Given two frames from a high-speed video, the proposed μ-BERT method can localize and highlight the regions of micro-movements. **Best viewed in color.**

Micro-expression identification requires both semantics and micro-movement analysis. Since they are difficult to observe through human eyes, a high-speed camera, usually with 200 frames per second (FPS) [6, 15, 51], is typically used to capture the required video frames. Previous work [11] tried to understand this micro information using MagNet [29] to amplify small motions between two frames, e.g., onset and apex frames. However, these methods still have limitations in terms of accuracy and robustness. In summary, the contributions of this work are four-fold:

- A novel Facial Micro-expression Recognition (MER) via Pre-training of Deep Bidirectional Transformers approach (Micron-BERT or μ-BERT) is presented to

tackle the problem in a self-supervised learning manner. The proposed method aims to identify and localize micro-movements in faces accurately.

- As detecting the tiny moment changes in faces is an essential input to the MER module, a new *Diagonal Micro Attention* (DMA) mechanism is proposed to precisely identify small movements in faces between two consecutive video frames.
- A new *Patch of Interest* (POI) module is introduced to efficiently spot facial regions containing the micro-expressions. Far apart from prior methods, it is trained in an unsupervised manner without using any facial labels, such as facial bounding boxes or landmarks.
- The proposed μ-BERT framework is designed in a self-supervised learning manner and trained in an end-to-end deep network. Indeed, it consistently achieves State-of-the-Art (SOTA) results in various standard micro-expression benchmarks, including CASME II [50], CASME3 [14], SAMM [6] and SMIC [15]. It achieves high recognition accuracy on new unseen subjects of various gender, age, and ethnicity.

2. Related Work

Generally, prior studies in micro-expression can be divided into two categories, including micro-expression spotting (MES) and micro-expression recognition.

Micro-Expression Spotting (MES). The goal of MES is to determine the specific instant during which a micro-expression occurs. Li et al. [16] adopted a spatial-channel attention network to detect micro-expression action units. Tran et al. [39] attempted to standardize with the SMIC-E database and an evaluation protocol. MESNet [43] introduced a CNN-based approach with a (2+1)D convolutional network, a clip proposal, and a classifier.

Micro-Expression Recognition (MER). The goal of MER tasks is to classify the facial micro-expressions in a video. Ling et al. [11] present a new way of learning facial graph representations, allowing these small movements to be seen. Kumar and Bhanu [31] exploited connections between landmark points and their optical flow patch and achieved improvements over state-of-the-art (SOTA) methods for both the CASME II and SAMM. Liu et al. [21] presented a new method using transfer learning achieved an accuracy of 84.27% on a composite of three datasets. Wang et al. [45] presented an Eulerian-motion magnification-based approach that highlights these small movements.

Other Work. Other research, while not necessarily on MES or MER, is relevant to our approach. An advance in video motion magnification is shown in [29], outperforming the SOTA methods in multiple areas. This learning-based model can extract filters from data directly rather than rely on ones designed by hand, like the state-of-the-art method.

3. BERT Revisited

3.1. BERT in Vision Problems

Transformers and deep learning have significantly improved results for many tasks in computer vision [1, 7, 9, 22, 23, 26, 27, 30, 40, 41]. Worth mentioning is Vision Transformer (ViT) [7], one of the first research efforts at the intersection of Transformers and computer vision. Unlike the traditional CNN network, ViT splits an image into a sequence of patches and applies the Transformers-based framework directly. Inspired by the success of BERT in Natural Language Processing (NLP), Bidirectional Encoder representation from Image Transformers (BEiT) [1] is presented as a self-supervised learning framework in computer vision. In particular, image patches are tokenized using DALL-E [32] to the visual tokens. These tokens are then randomly masked before feeding into the transformer backbone. The training objective is to recover the original visual tokens from the corrupted patches. These methods [1, 38] have marked a remarkable improvement compared to supervised learning methods by leveraging large-scale unlabelled datasets, e.g., ImageNet-1K, ImagNet-21K [33], to discover semantic information.

3.2. Limitations of BERT in Vision Problems

One limitation of using BERT in vision problems is the tokenization step. In the NLP field, a token has precisely one word mapped into it. In vision problems, however, many possible images or patches can share the same token as long as they have the same content. Therefore, designing a BERT model to mask a token and train a prediction model in the missing contexts in computer vision is more challenging than NLP. In addition, the *tokenizer*, i.e., DALL-E [32], is not robust enough to map similar contexts to a token. It yields noise in the tokenization process and affects the overall training performance. He et al., [9] presented a Masked Auto Encoder (MAE) that utilizes the BERT framework. Instead of tokenizing images, it eliminates patches of an image via a random masking strategy and reconstructs the context of these masked patches to the original content. Although this method can avoid using the tokenizer, it only considers the context inside an image. Thus, it does not apply to micro-expression, which requires understanding semantic information from consecutive video frames. In this paper, μ-BERT is presented to address these limitations.

4. The Proposed μ-BERT Approach

μ-BERT is designed to model micro-changes of facial texture across temporal dimensions, which is hard to observe by unaided human eyes via a reconstruction process. The proposed μ-BERT architecture, shown in Figure 2, consists of five main blocks: a *μ-Encoder*, *Patch*

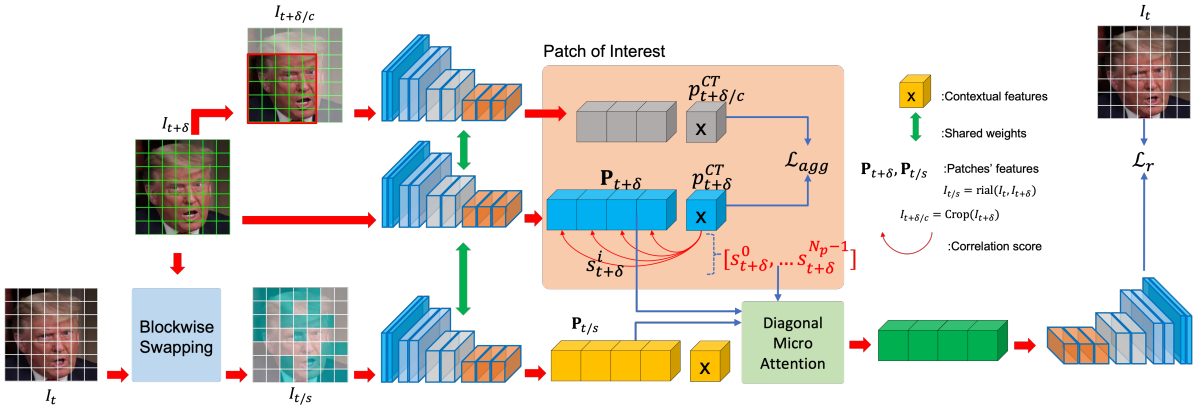

Figure 2. An overview of the proposed μ-BERT approach to facial micro-expression recognition.

of Interest (PoI), *Blockwise Swapping*, *Diagonal Micro Attention (DMA)*, and a *μ-Decoder*. Given input images I_t and $I_{t+\delta}$, the role of the μ-Encoder is to represent I_t and $I_{t+\delta}$ into latent vectors. Then, Patch of Interest (PoI) constrains μ-BERT to look into facial regions containing micro-expressions rather than unrelated regions such as the background. Blockwise Swapping and Diagonal Micro Attention (DMA) allow the model to focus on facial regions that primarily consist of micro differences between frames. Finally, μ-Decoder reconstructs the output signal back to the determined one. Compared to prior works, μ-BERT can adaptively focus on changes in facial regions while ignoring the ones in the background and effectively recognizes micro-expressions even when face movements occur. Moreover, μ-BERT can also alleviate the dependency on the accuracy of alignment approaches in pre-processing step.

4.1. Non-overlapping Patches Representation

In μ-BERT, an input frame $I_t \in \mathbb{R}^{H\times W\times C}$ is divided into a set of several non-overlapping patches $\mathcal{P}_t$ as Eqn. (1).

$$\mathcal{P}_t = \{p_t^i\}_{i=0}^{N_p-1} \qquad |\mathcal{P}_t| = HW/(ps^2) \tag{1}$$

where H, W, C are the height, width, and number of channels, respectively. Each patch p_t^i has a resolution of $ps \times ps$. In our experiments, $H = W = 224$, $C = 3$, and $ps = 8$.

4.2. μ-Encoder

Each patch $p_i \in \mathcal{P}_t$ is linearly projected into a latent vector of dimension d denoted as $\mathbf{z}_t^i \in \mathbb{R}^{1\times d}$, with additive fixed positional encoding [42]. Then, an image I_t can be represented as in Eqn. (2).

$$\begin{aligned}\mathbf{Z}_t &= concat\left[\mathbf{z}_t^0, \mathbf{z}_t^1, \dots \mathbf{z}_t^{N_p-1}\right] \in \mathbb{R}^{N_p\times d}\\ \mathbf{z}_t^i &= \alpha(p_t^i) + \mathbf{e}(i)\end{aligned} \tag{2}$$

where α and $\mathbf{e}$ are the projection embedding network and positional embedding, respectively. Let μ-Encoder, denoted as $\mathcal{E}$, be a stack of continuous blocks. Each block consists

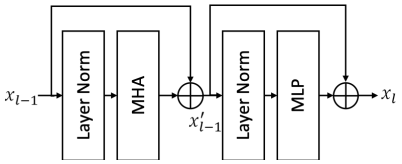

Figure 3. **Builing block of Encoder and Decoder.** Each block includes Multi-Head Attention (MHA) and Layer Normalization.

of alternating layers of Multi Head Attention (MHA) and Multi-Layer Perceptron (MLP), as illustrated in Figure 3. The Layer Norm (LN) is employed to the input signal before feeding to MHA and MLP layers, as in Eqn. (3).

$$\begin{aligned}\mathbf{x'}_l &= \mathbf{x}_{l-1} + \text{MHA}(\text{LN}(\mathbf{x}_{l-1}))\\ \mathbf{x}_l &= \mathbf{x'}_l + \text{MLP}(\text{LN}(\mathbf{x'}_l))\\ \mathbf{x}_0 &= \mathbf{Z}_t,\ 1 \le l \le L_e\end{aligned} \tag{3}$$

where L_e is the number of blocks in $\mathcal{E}$. Given $\mathbf{Z}_t$, The output latent vector $\mathbf{P_t}$ is represented as in Eqn. (4).

$$\mathbf{P}_t = \mathcal{E}(\mathbf{Z}_t) \qquad \mathbf{P}_t \in \mathbb{R}^{N_p\times d} \tag{4}$$

4.3. μ-Decoder

The proposed auto-encoder is designed symmetrically. It means that the decoder part denoted as $\mathcal{D}$, has a similar architecture to the encoder $\mathcal{E}$. Given a latent vector $\mathbf{P}_t$, the decoded signal $\mathbf{Q}_t$ is represented as in Eqn. (5).

$$\mathbf{Q}_t = \mathcal{D}(\mathbf{P}_t) \qquad \mathbf{Q}_t \in \mathbb{R}^{N_p\times d} \tag{5}$$

We add one more Linear layer to interpolate $\mathbf{Q}_t$ to an intermediate signal $\mathbf{y}_t$ before reshaping it into the image size.

$$\begin{aligned}\mathbf{Q}_t \in \mathbb{R}^{N_p\times d} &\xrightarrow{\text{linear}} \mathbf{y}_t \in \mathbb{R}^{N_p\times ps\times ps\times C}\\ \mathbf{y}_t \in \mathbb{R}^{N_p\times ps\times ps\times C} &\xrightarrow{\text{reshape}} \mathbf{y'}_t \in \mathbb{R}^{H\times W\times C}\end{aligned} \tag{6}$$

4.4. Blockwise Swapping

Given two frames I_t and $I_{t+\delta}$, we realize the fact that:

$$\lim_{\delta\to 0} s(p_t^i, p_{t+\delta}^i) = 1 \tag{7}$$

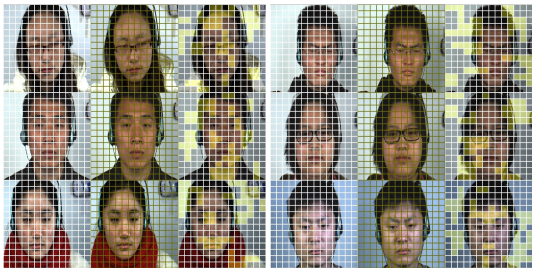

Figure 4. **Blockwise Swapping.** For each triplet, we present the I_t (left), the $I_{t+\delta}$ (middle) and the $I_{t/s}$ (right). The yellow blocks in $I_{t/s}$ represent swapped patches from $I_{t+\delta}$ that are randomly swapped into I_t. Best viewed in color.

where p_t^i is the i^{th}-patch at frame t. s denotes a function to measure the similarity between p_t^i and $p_{t+\delta}^i$ where a higher score indicates higher similarity and $0 \leq s(p_t^i, p_{t+\delta}^i) \leq 1$. Given a patch correlation as in Eqn. (7), we propose a *Blockwise Swapping mechanism* to (1) firstly *randomly swap two corresponding blocks* p_t^i and $p_{t+\delta}^i$ between two frames to create a swapped image $I_{t/s}$, and then (2) *enforce the model to spot these changes* and reconstruct I_t from $I_{t/s}$. By doing so, the model is further strengthened in recognizing and restoring the swapped patches. As a result, the learned model can be enhanced by the capability to notice small differences between frames. Moreover, as shown in Eqn. (7), shorter time δ causing larger similarity between I_t from $I_{t/s}$ can further help to enhance the robustness on spotting these differences. The detail of this strategy is described in Algorithm 1 and Figure 4.

4.5. Diagonal Micro Attention (DMA)

As a result of Blockwise Swapping, the image patches $\mathcal{P}_{t/s}$ from $I_{t/s}$ consists of two types, i.e. $p_{t/s}^j$ from $\mathcal{P}_t$ of I_t and $p_{t/s}^i$ from $\mathcal{P}_{t+\delta}$ of $I_{t+\delta}$. Then, the next stage is to learn how to reconstruct $\mathcal{P}_t$ from $\mathcal{P}_{t/s}$. Since $p_{t/s}^i$ includes all changes between I_t and $I_{t/s}$, more emphasis is placed on $p_{t/s}^i$ during reconstruction process. Theoretically, the *ground truth* of the index of $p_{t/s}^i$ in $\mathcal{P}_{t/s}$ can be utilized to enforce the model focusing on these swapped patches. However, adopting this information may reduce the learning capability to spot these microchanges. Therefore, a novel attention mechanism named Diagonal Micro-Attention (DMA) is presented to enforce the network automatically focusing on swapped patches $p_{t/s}^i$ and equip it with the ability to precisely spot and identify all changes between images. Notice that these changes may include patches in the background. The following section introduces a solution to constrain the learned network focusing on only meaningful facial regions. The details of DMA are presented in Figure 5. Formally, we construct an attention map $\hat{A}$ between $\mathcal{P}_{t+\delta}$ and $\mathcal{P}_{t/s}$ where the $diag(\hat{A})$ illustrates correlations between two corresponding patches $p_{t+\delta}^i$

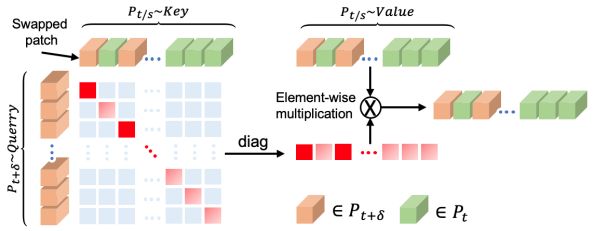

Figure 5. **Diagonal Micro-Attention (DMA) module.** Diagonal values from the attention map between $\mathbf{P}_{t/s}$ and $\mathbf{P}_{t+\delta}$ are used to rank the importance of each patch in the swapped image.

and $p_{t/s}^i$. From the observation that $\hat{A}(i,i) > \hat{A}(j,j)$ for all $p_{t/s}^i \in \mathcal{P}_{t+\delta}$ and $p_{t/s}^j \in \mathcal{P}_t$, $diag(\hat{A})$ can be effectively adopted as weights indicating important features. Full operations of DMA are presented in Eqn. (8) and Eqn. (9).

$$\hat{A} = \text{softmax}\left(Q(\mathbf{P}_{t+\delta}) \otimes K(\mathbf{P}_{t/s})^T\right), \sum_{j=0}^{N_p} \hat{A}(i,j) = 1 \quad (8)$$

$$\mathbf{P}_{dma} = \text{diag}(\hat{A}) \times V(\mathbf{P}_{t/s}) \quad (9)$$

where $\times$ denotes the Element-wise multiplication operator.

4.6. Patch of Interest (POI)

In Section 4.5, Diagonal Micro-Attention has been introduced to weigh the importance of swapped patches automatically. These swapped patches are randomly produced via Blockwise Swapping, as in Algorithm 1. In theory, the ideal case is when all swapped patches are located within the facial region only so that the deep network can learn the micro-movements from the facial parts solely and not be distracted by the background. In practice, however, we

Algorithm 1 Blockwise Swapping

Input: $\mathcal{P}_t, \mathcal{P}_{t+\delta}$ image patches ($N_p = h \times w$); r_s: swapping ratio (default: 0.5); min_bs: minimum block size (default: 16); min_ar: minimum aspect ratio (default: 0.3)
Output: Swapped image patches $\mathcal{P}_{t/s}$
$\mathcal{P}_{t/s} \leftarrow \mathcal{P}_t$; $N_p \leftarrow |\mathcal{P}_t|$
$c \leftarrow 0$
while $c \leq r_s \times N_p$ **do**
  $bs \leftarrow \text{rnd}(min\_bs, r_s \times N_p - c)$
  $ar \leftarrow \text{rnd}(min\_ar, 1/min\_ar)$
  $m, n \leftarrow \sqrt{bs \cdot ar}, \sqrt{bs/ar}$
  $p, q \leftarrow \text{rnd}(0, h - m), \text{rnd}(0, w - n)$
  $\forall i \in [p, p+m), j \in [q, q+n)$:
    $k \leftarrow i \times w + j$
    $\mathcal{P}_{t/s}(k) \leftarrow \mathcal{P}_{t+\delta}(k)$
  $c \leftarrow c + m \times n$
end while
return $\mathcal{P}_{t/s}$

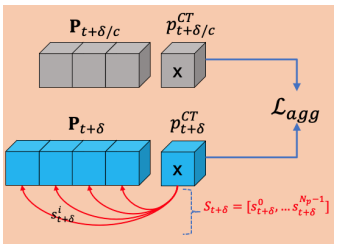

Figure 6. **Patch of Interest (POI) module.** The $\mathbf{P}_t$ and $\mathbf{P}_{t+\delta/c}$ are sequence of patch features of $I_{t+\delta}$ and its random cropped version. $p^{CT}_{t+\delta}$ and $p^{CT}_{t+\delta/c}$ are their corresponding contextual features.

can only identify which parts are selected in the Block-wise Swapping algorithm if the facial regions are available. Thus, the Patch of Interest (POI) is introduced to automatically explore the salient regions and ignore the background patches in an image. Apart from prior methods, the proposed POI leverages the characteristic of self-attention and can be achieved through self-learning without facial labels, such as facial bounding boxes or segmentation masks. The idea of the POI module is illustrated in Figure 6. Thanks to POI, a capability of *automatically focusing on facial regions* is further equipped to the learned model, making it *more robust against facial movements*.

The POI relies on the contextual agreement between the frame $I_{t+\delta}$ and Crop($I_{t+\delta}$). Motivated by the BERT framework, we add a Contextual Token z^{CT} to the beginning of the sequence of patches, as in Eqn. (2), to learn the contextual information in the image. The deeper this token passes through the Transformer blocks, the more information is accumulated from $z^i_t \in \mathcal{P}_t$. As a result, z^{CT} becomes a placeholder to store the information extracted from other patches in the sequence and present the contextual information of the image. Let $p^{CT}_{t+\delta}$ and $p^{CT}_{t+\delta/c}$ be the contextual features of frame $I_{t+\delta}$ and its cropped version Crop($I_{t+\delta}$) respectively. The agreement loss is then defined as in Eqn. (10).

$$\mathcal{L}_{agg} = H(p^{CT}_{t+\delta}, p^{CT}_{t+\delta/c}) \tag{10}$$

where H is the function that enforces $p^{CT}_{t+\delta}$ to be similar to $p^{CT}_{t+\delta/crop}$ so that the model can discover the salient patches. The POI can be extracted from the attention map A at the last attention layer of encoder $\mathcal{E}$. In particular, we measure:

$$\mathbf{S}_{t+\delta} = A\,[0,:] = \left[s^0_{t+\delta}, s^1_{t+\delta}, \ldots s^{N_p-1}_{t+\delta}\right] \tag{11}$$

where $\sum_{i=0}^{N_p-1} s^i_{t+\delta} = 1$. The higher the score $s^i_{t+\delta}$, the richer the patch contains contextual information. Now, Eqn. (9) can be reformulated as in Eqn. (12).

$$\begin{aligned} \mathrm{W} &= \mathrm{diag}(\hat{A}) \times S_{t+\delta} \\ \mathbf{P}_{dma} &= \mathrm{W} \times V(\mathbf{P}_{t/s}) \end{aligned} \tag{12}$$

4.7. Loss Functions

The proposed μ-BERT deep network is optimized using the proposed loss function as in Eqn. (13).

$$\mathcal{L} = \gamma \times \mathcal{L}_r + \beta \times \mathcal{L}_{agg} \tag{13}$$

where γ and β are the weights for each loss.

Reconstruction Loss. The output of the decoder $\mathbf{y'}_t$ is reconstructed to the original image I_t using the Mean Square Error (MSE) function.

$$\mathcal{L}_r = \mathrm{MSE}(\mathbf{y'}_t, I_t) \tag{14}$$

Contextual Agreement Loss. MSE is also used to enforce the similarity of contextual features of $I_{t+\delta/crop}$ and $I_{t+\delta}$

$$\mathcal{L}_{agg} = \mathrm{MSE}(p^{CT}_{t+\delta}, p^{CT}_{t+\delta/crop}) \tag{15}$$

5. Experimental Results

5.1. Datasets and Protocols

CASME II [50]. With a 200 fps sampling rate and a facial resolution of 280 × 340, CASME II provides 247 micro-expression samples from 26 subjects of the same ethnicity. Labels include apex frames, action units, and emotions.
SAMM [6]. Also, using a 200 fps frame rate and a facial resolution of 400 × 400, SAMM consists of 159 samples from 32 participants and 13 ethnicities. The samples all have emotions, apex frames, and action unit labels.
SMIC [15]. SIMC is made up of 164 samples. Lacking apex frame and action unit labels, the samples span 16 participants of 3 ethnicities. The recordings are taken with a resolution of 640 × 480 at 100 fps.
CASME3 [14]. Officially known as CAS(ME)^3 provides 1,109 labeled micro-expressions and 3,490 labeled macro-expressions. This dataset has roughly 80 hours of footage with a resolution of 1280 × 720.

5.2. Micro-Expression Self-Training

We use all raw frames from CASME3 for self-training except frames of test set. It is important to note that we do not use labels or meta information such as onset, offset, and apex index frames nor labeled emotions. In total, we construct an unlabelled dataset of 8M frames. The images are resized to 224 × 224. Then, each image is divided into patches of 8 × 8, yielding $N_p = 784$ patches. The temporal index δ is selected randomly between a lower bound of 5 and an upper bound of 11, experimentally. The swapping ratio r_s is selected as 50% of the number of patches being swapped from $I_{t+\delta}$ to I_t. Each patch is projected to a latent space of $d = 512$ dimensions before being fed into the encoder and decoder. For the encoder and decoder, we keep the same d for all vectors and similar configurations, i.e., $L_e = L_d = 4$. μ-BERT is implemented easily in

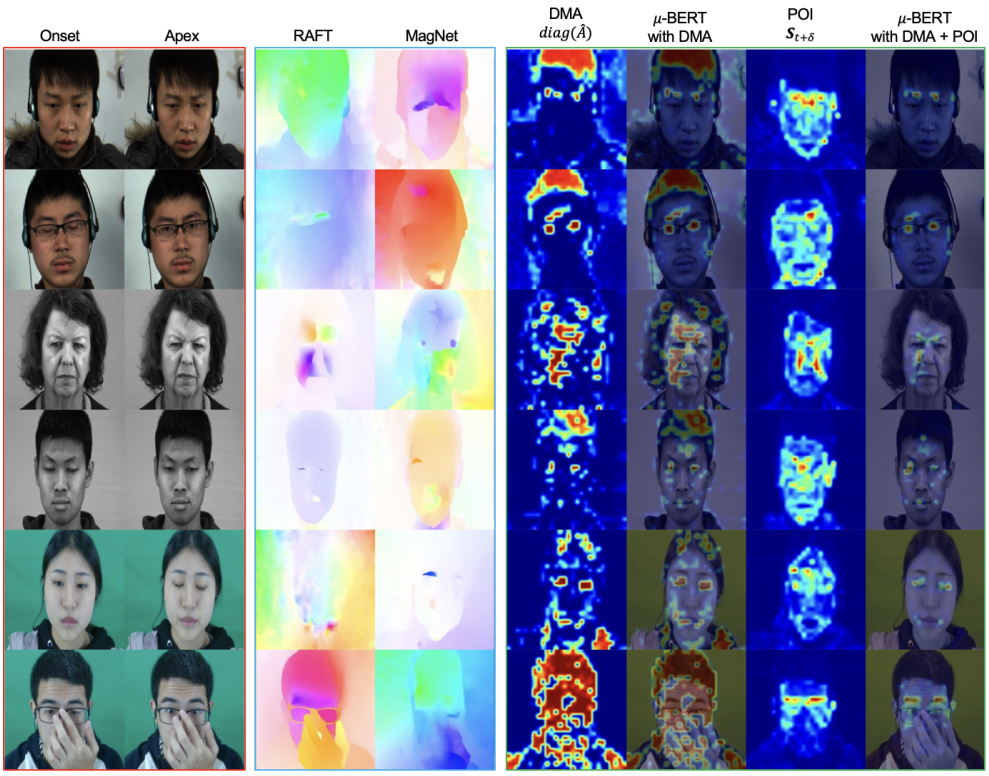

Figure 7. We demonstrate how μ-BERT perceives the tiny differences between two frames. The first two rows are *onset* and *apex* input frames. The third and fourth rows are the results of RAFT and MagNet, respectively. The rest of the rows are our μ-BERT results.

Pytorch framework and trained by 32 $\times$ A100 GPUs (40G each). The learning rate is set to 0.0001 initially and then reduced to zero gradually under ConsineLinear [24] policy. The batch size is set to 64/GPU. The model is optimized by AdamW [25] for 100 epochs. The training is completed within three days.

5.3. Micro-Expression Recognition

We leverage the pretrained μ-BERT as an initial weight and take the encoder $\mathcal{E}$ and DMA module of μ-BERT as the MER backbone. The input of MER is the *onset* and *apex* frames which correspond to I_t and $I_{t+\delta}$ respectively. In Eqn. (8), $\mathbf{P}_{dma}$ are the features representing the micro changes and movements between onset and apex frames. They can be effectively adopted for recognizing micro-expressions. We adopt the standard metrics and protocols of MER2019 challenge [34] with the unweighted F1 score $UF1 = \frac{1}{C}\sum_{i=0}^{C-1}\frac{2\times TP_i}{TP_i+FP_i+FN_i}$ and accuracy $UAR = \frac{1}{C}\sum_{i=0}^{C-1}\frac{TP_i}{N_i}$, where C is the number of MEs, N_i is the total number of i^{th} ME in the dataset. Leave-one-out cross-validation (LOOCV) scheme is used for evaluation.

5.4. Results

Our proposed μ-BERT shows a significant improvement over prior methods and baselines, as shown in Table 1 on the CASME3. Tested using 3, 4, and 7 emotion classes, μ-BERT achieves double-digit gains over the compared methods in each category. In the case of 3 emotion classes, μ-BERT achieved a 56.04% UF1 score and 61.25% UAR, compared to RCN-A's [49] 39.28% UF1 and 38.93% UAR. For 4 emotion classes, μ-BERT outperforms Baseline (+Depth) [14] 47.18% to 30.01% for UF1 and 49.13% to 29.82% for UAR. Large gains over Baseline (+Depth) [14] are seen in the case of 7 emotion classes, where μ-BERT attains UF1 and UAR scores of 32.64% and 32.54% respectively, compared to 17.73% and 18.29% for the baseline.

Table 2 details results for CASMEII. μ-BERT shows improvements over all other methods. For three categories, it achieves a UF1 of 90.34% and UAR of 89.14%, representing 3.37% and 0.86% increases over the prior leading method (OFF-ApexNet [8]), respectively. Similar improvement is seen in five categories: a 4.83% over TSCNN [35] in terms of UF1 and a 0.89% increase over SMA-STN [19] for UAR. Similarly, μ-BERT performs competitively with other methods on the SAMM as seen in Table 3. Using 5 emotion classes, μ-BERT outperforms MinMaNet [47] by a large margin in terms of UF1 (83.86% vs 76.40%) and UAR (84.75% vs 76.70%), respectively. The performance of μ-BERT on SMIC is compared against several others in Table 4. μ-BERT outperforms others with a 7.5% increase in UF1 to 85.5% and a 3.97% boost in UAR to 83.84%.

On the composite dataset, μ-BERT again outperforms

Table 1. MER on the CASME3 dataset.

Method	# Classes	UF1 (%)	UAR(%)
FR [54]	3	34.93	34.13
STSTNet [18]	3	37.95	37.92
RCN-A [49]	3	39.28	38.93
μ-BERT (ours)	3	**56.04**	**61.25**
Baseline [14]	4	29.15	29.10
Baseline (+Depth) [14]	4	30.01	29.82
μ-BERT (ours)	4	**47.18**	**49.13**
Baseline [14]	7	17.59	18.01
Baseline(+Depth) [14]	7	17.73	18.29
μ-BERT (ours)	7	**32.64**	**32.54**

Table 2. MER on CASME II dataset.

Method	# Classes	UF1 (%)	UAR (%)
LR-GACNN [10]	5	70.90	81.30
AMAN [46]	5	71.00	75.40
Graph-TCN [13]	5	72.46	73.98
DSTAN [44]	5	73.00	75.00
GEME [28]	5	73.54	75.20
MiMaNet [47]	5	75.90	79.90
SMA-STN [19]	5	79.46	82.59
TSCNN [35]	5	80.70	80.97
μ-BERT (ours)	5	**85.53**	**83.48**
STSTNet [17]	3	83.82	86.86
OFF-ApexNet [8]	3	86.97	88.28
MAE [9]	3	88.03	87.28
μ-BERT (ours)	3	**90.34**	**89.14**

Table 3. MER on SAMM dataset.

Method	# Classes	UF1 (%)	UAR (%)
AMAN [46]	5	67.00	68.85
SMA-STN [19]	5	70.33	77.20
GRAPH-AU [12]	5	70.45	74.26
MTMNet [48]	5	73.60	74.10
MiMaNet [47]	5	76.40	76.70
MAE [9]	5	80.40	88.98
μ-BERT (ours)	5	**83.86**	**84.75**

other methods (Table 5). Attaining a UF1 score of 89.03% and UAR of 88.42%, μ-BERT realizes 0.73%, and 0.82% gains over previous best MiMaNet [47], respectively. Table 6 shows the impact of DMA and POI on CASME3. Our method gives more modest gains of approximately 2% in both metrics. A greater improvement is seen with DMA, where UF1 and UAR increase by another 2-4%. Significant improvement from μ-BERT is seen when adopting both modules, with a UF1 of 32.64% and UAR of 32.54%, representing roughly 10% gains over previous methods.

Table 4. MER on SMIC dataset.

Method	# Classes	UF1 (%)	UAR (%)
DIKD [36]	3	71.00	76.06
TSCNN [35]	3	72.36	72.74
MTMNet [48]	3	74.40	76.00
AMAN [46]	3	77.00	79.87
MiMaNet [47]	3	77.80	78.60
DSTAN [44]	3	78.00	77.00
MAE [9]	3	81.86	80.82
μ-BERT (ours)	3	**85.50**	**83.84**

5.5. How μ-BERT perceives micro-movements

To understand the micro-movements between two frames, the onset and apex frames are inputs for μ-BERT. These frames represent the moments that the micro-expression starts and is observed. We measure diag($\hat{A}$) (Subsection 4.5) and $\mathbf{S}_{t+\delta}$ (Eqn (11)) values to identify which regions contain small movements between two frames. Comparisons of μ-BERT with RAFT [37], i.e., optical flow-based method and MagNet [29] are also conducted as in Fig. 7. The third and fourth columns in Fig 7 show the results of RAFT [37], and MagNet [29] on spotting the micro-movements, respectively. While RAFT is an optical flow-based method, MagNet amplifies small differences between the two frames. These methods are sensitive to the environment (e.g., lighting, illuminations). Thus, noises in the background still exist in their outputs. In addition, neither RAFT nor MagNet understand semantic information in the frame and distinguish changes inside facial or background regions. Meanwhile, μ-BERT shows its advantages in perceiving micro-movements via distinguishing the facial regions and spotting the micro-expressions. In particular, the attention map in the fifth column, in Fig. 7 illustrates the micro-differences between onset and apex frames. The higher contrast represents the higher chance of small movements in these regions. With the POI module, μ-BERT can automatically figure out the informative patches and ignore the background ones. Then, with DMA module, μ-BERT, can detect and localize which corresponding patches/regions contain tiny movements. As shown in the seventh column, attention maps represent the most salient regions in the image. By empowering DMA and POI, μ-BERT effectively identifies micro-movements within facial

Table 5. MER on the Composite dataset (MECG2019).

Method	# Classes	UF1 (%)	UAR(%)
Dual-Inception [55]	3	73.22	72.78
FR [53]	3	78.38	78.32
NMER [20]	3	78.85	78.24
GRAPH-AU [12]	3	79.14	79.33
ICE-GAN [52]	3	84.50	84.10
BDCNN [3]	3	85.09	85.00
moment [48]	3	86.40	85.70
MiMaNet [47]	3	88.30	87.60
MAE [9]	3	88.50	87.40
μ-BERT (ours)	3	**89.03**	**88.42**

regions, as demonstrated in the last column.

5.6. Ablation studies

This section compares μ-BERT against other self-supervised learning (SSL) methods on the MER task. CASME3 is used for experiments since it has many unlabelled images to demonstrate the power of SSL methods. We also analyze the essential contributions of Diagonal Micro-Attention (DMA) and Patch of Interest (POI) modules. Finally, we illustrate the robustness of μ-BERT pretrained on CASME3 on unseen datasets and domains.

Comparisons with self-supervised learning methods. We utilize the encoder and decoder parts of μ-BERT (**without** DMA and POI) to train previous SSL methods (MoCo V3 [4], BEIT [1], MAE [9]) and then continue learning the MER task on the large-scale database CASME3 [14]. Overall results are shown in Table 6. It is expected that ViT-S achieves the lowest performance for UF1 and UAR as ImageNet and Micro-Expression are two different domains. Three self-supervised methods (MoCo V3, BEIT, and MAE) got better results when they were pretrained on CASME before fine-tuning to the recognition task. Compared to ViT-S, these SSL methods gain remarkable performance. Especially, MAE [9] achieves 3.5% and 2% up on UF1 and UAR compared to ViT-S, respectively.

The role of Blockwise Swapping. Our basic setup of μ-BERT (denoted as MB1) is employed to train in an SSL manner. It is noted that only Blockwise Swapping is involved, and it does not contain either DMA or POI. Compared to MAE, MB1 outperforms MAE by 2% in both UF1 and UAR, approximately. The reasons are: (1) Blockwise Swapping enforces the model to learn local context features inside an image, i.e., I_t, and (2) It helps the network to figure out micro-disparities between two frames I_t and $I_{t+\delta}$.

The role of DMA. This module is the guide to tell the network where to look and which patches to focus. By doing so, the μ-BERT gets more robust knowledge of micro-movements between two frames. For this reason, the network (denoted as MB2) achieves 2% on UF1 and a significant 4% gain on UAR compared to MB1.

Table 6. MER performance on CASME3 by different self-supervised methods and various settings of μ-BERT

Method	Pre-train	DMA	POI	UF1	UAR
ViT-S [7]	ImageNet	✗	✗	20.34	18.76
MoCo V3 - R50 [4]	CASME3	✗	✗	19.12	17.36
MoCo V3 - R101 [4]	CASME3	✗	✗	20.14	18.52
MoCo V3 [4]	CASME3	✗	✗	22.13	19.34
BEIT [1]	CASME3	✗	✗	23.54	19.89
MAE [9]	CASME3	✗	✗	23.86	20.87
μ-BERT (MB1)	CASME3	✗	✗	25.27	22.96
μ-BERT (MB2)	CASME3	✓	✗	27.35	26.18
μ-BERT (MB3)	CASME3	✓	✓	**32.64**	**32.54**

The role of POI. Since MB1 are sensitive to background noise, the micro-disparities features $\mathbf{P}_{dma}$ might contain unwanted features coming from the background. The POI is designed as a filter that only lets the typical interesting patches belonging to the subject go through and preserves the micro-movement features only. The improvements of up to 6% compared to MB2 demonstrate the important role of POI in μ-BERT for micro-expression tasks. Qualitative results demonstrated in Supplementary Material can further emphasize the advantages of POI in assisting the network to be robust against facial movements.

6. Conclusions and Discussions

Unlike a few concurrent research on micro-expression, we move forward and study how to explore BERT pretraining for this problem. In our proposed μ-BERT, we presented a novel Diagonal Micro Attention (DMA) to learn the micro-movements of the subject across frames. The Patch of Interest (POI) module is proposed to guide the network, focusing on the most salient parts, i.e., facial regions, and ignoring the noisy sensitivities from the background. Empowered by the simple design of μ-BERT, SOTA performance on micro-expression recognition tasks is achieved in four benchmark datasets. Our perspective will inspire more future study efforts in this direction.

Limitations. We demonstrated the efficiency of the POI module in removing noise in the background, which is sensitive to lighting and illumination. However, suppose any facial parts, e.g., the forehead, are affected by lighting conditions while there are no movements. In that case, these noisy factors might also be included as micro-difference features. The robustness with different lighting conditions will be left as our future works.

Acknowledgement This work is supported by Arkansas Biosciences Institute (ABI) Grant, NSF WVAR-CRESH and NSF Data Science, Data Analytics that are Robust and Trusted (DART). We also acknowledge the Arkansas High-Performance Computing Center for providing GPUs.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. 2021. 2, 8

[2] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1

[3] Bin Chen, Kun-Hong Liu, Yong Xu, Qing-Qiang Wu, and Jun-Feng Yao. Block division convolutional network with implicit deep features augmentation for micro-expression recognition. *IEEE Transactions on Multimedia*, 2022. 8

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 8

[5] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F. Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, 2016. 1

[6] Adrian K. Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, 9(1):116–129, 2018. 1, 2, 5

[7] 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. 2, 8

[8] Y.S. Gan, S.T. Liong, W.C. Yau, Y.C. Huang, and L.K. Tan. Off-apexnet on micro-expression recognition system. *Signal Processing: Image Communication*, 74:129–139, 2019. 6, 7

[9] 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. 2, 7, 8

[10] Ankith Jain Rakesh Kumar and Bir Bhanu. Micro-expression classification based on landmark relations with graph attention convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1511–1520, June 2021. 7

[11] Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1571–1580, 2021. 1, 2

[12] Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1571–1580, June 2021. 7, 8

[13] Ling Lei, Jianfeng Li, Tong Chen, and Shigang Li. A novel Graph-TCN with a graph structured representation for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2237–2245, 2020. 7

[14] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. Cas(me)¡sup¿3¡/sup¿: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022. 2, 5, 6, 7, 8

[15] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013. 1, 2, 5

[16] Yante Li, Xiaohua Huang, and Guoying Zhao. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing*, 436:221–231, 2021. 2

[17] S.T. Liong, Y.S. Gan, J. See, H. Khor, and Y. Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5. IEEE, 2019. 7

[18] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–5. IEEE, 2019. 7

[19] Jiateng Liu, Wenming Zheng, and Yuan Zong. SMA-STN: Segmented movement-attending spatiotemporal network for micro-expression recognition. *arXiv preprint arXiv:2010.09342*, 2020. 6, 7

[20] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–4. IEEE, 2019. 8

[21] Yanju Liu, Yange Li, Xinhai Yi, Zuojin Hu, Huiyu Zhang, and Yanzhong Liu. Lightweight vit model for micro-expression recognition enhanced by transfer learning. *Frontiers in Neurorobotics*, 16, 2022. 2

[22] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 2

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] Xuan-Bac Nguyen, Duc Toan Bui, Chi Nhan Duong, Tien D Bui, and Khoa Luu. Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition. 2

[27] Xuan-Bac Nguyen, Guee Sang Lee, Soo Hyung Kim, and Hyung Jeong Yang. Self-supervised learning based on spatial awareness for medical image analysis. *IEEE Access*, 8:162973–162981, 2020. 2

[28] Xuan Nie, Madhumita A Takalkar, Mengyang Duan, Haimin Zhang, and Min Xu. GEME: Dual-stream multi-task gender-based micro-expression recognition. *Neurocomputing*, 427:13–28, 2021. 7

[29] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Frédo Durand, William T. Freeman, and Wojciech Matusik. Learning-based video motion magnification, 2018. 1, 2, 7

[30] Kha Gia Quach, Ngan Le, Chi Nhan Duong, Ibsa Jalata, Kaushik Roy, and Khoa Luu. Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. *Pattern Recognition*, 128:108646, 2022. 2

[31] Ankith Jain Rakesh Kumar and Bir Bhanu. Micro-expression classification based on landmark relations with graph attention convolutional network. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1511–1520, 2021. 2

[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[33] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 2

[34] John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. Megc 2019 – the second facial micro-expressions grand challenge. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–5, 2019. 6

[35] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao. Recognizing spontaneous micro-expression using a three-stream convolutional neural network. *IEEE Access*, 7:184537–184551, 2019. 6, 7

[36] Bo Sun, Siming Cao, Dongliang Li, Jun He, and Lejun Yu. Dynamic micro-expression recognition using knowledge distillation. *IEEE Transactions on Affective Computing*, 2020. 7

[37] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 7

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 2

[39] Thuong-Khanh Tran, Quang-Nhat Vo, Xiaopeng Hong, Xiaobai Li, and Guoying Zhao. Micro-expression spotting: A new benchmark. *Neurocomputing*, 443:356–368, 2021. 2

[40] Thanh-Dat Truong, Quoc-Huy Bui, Chi Nhan Duong, Han-Seok Seo, Son Lam Phung, Xin Li, and Khoa Luu. Direcformer: A directed attention in transformer approach to robust action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20030–20040, June 2022. 2

[41] Thanh-Dat Truong, Chi Nhan Duong, The De Vu, Hoang Anh Pham, Bhiksha Raj, Ngan Le, and Khoa Luu. The right to talk: An audio-visual transformer approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1105–1114, October 2021. 2

[42] 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

[43] Su-Jing Wang, Ying He, Jingting Li, and Xiaolan Fu. Mesnet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos. *IEEE Transactions on Image Processing*, 30:3956–3969, 2021. 2

[44] Yan Wang, Yikun Huang, Can Liu, Xiaoying Gu, Dandan Yang, Shuopeng Wang, and Bo Zhang. Micro expression recognition via dual-stream spatiotemporal attention network. *Journal of Healthcare Engineering*, 2021, 2021. 7

[45] Yandan Wang, John See, Yee-Hui Oh, Raphael C.-W. Phan, Yogachandran Rahulamathavan, Huo-Chong Ling, Su-Wei Tan, and Xujie Li. Effective recognition of facial micro-expressions with video motion magnification. *Multimedia Tools and Applications*, 76(20):21665–21690, 2016. 2

[46] Mengting Wei, Wenming Zheng, Yuan Zong, Xingxun Jiang, Cheng Lu, and Jiateng Liu. A novel micro-expression recognition approach using attention-based magnification-adaptive networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2420–2424. IEEE, 2022. 7

[47] Bin Xia and Shangfei Wang. Micro-expression recognition enhanced by macro-expression from spatial-temporal domain. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1186–1193, 2021. 6, 7, 8

[48] Bin Xia, Weikang Wang, Shangfei Wang, and Enhong Chen. Learning from macro-expression: a micro-expression recognition framework. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2936–2944, 2020. 7, 8

[49] Zhaoqiang Xia, Wei Peng, Huai-Qian Khor, Xiaoyi Feng, and Guoying Zhao. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29:8590–8605, 2020. 6, 7

[50] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PLOS ONE*, 9(1):1–8, 01 2014. 2, 5

[51] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2013. 1

[52] Jianhui Yu, Chaoyi Zhang, Yang Song, and Weidong Cai. ICE-GAN: Identity-aware and capsule-enhanced gan for micro-expression recognition and synthesis. *arXiv preprint arXiv:2005.04370*, 2020. 8

[53] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *arXiv preprint arXiv:2101.04838*, 2021. 8

[54] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition*, 122:108275, 2022. 7

[55] L. Zhou, Q. Mao, and L. Xue. Dual-inception network for cross-database micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5. IEEE, 2019. 8