SelfME: Self-Supervised Motion Learning for Micro-Expression Recognition

Xinqi Fan[1], Xueli Chen[1], Mingjie Jiang[1], Ali Raza Shahid[1,2], Hong Yan[1]
[1]City University of Hong Kong [2]COMSATS University Islamabad
{xinqi.fan, xuelichen3-c, minjiang5-c, a.raza}@my.cityu.edu.hk, h.yan@cityu.edu.hk

Abstract

Facial micro-expressions (MEs) refer to brief spontaneous facial movements that can reveal a person's genuine emotion. They are valuable in lie detection, criminal analysis, and other areas. While deep learning-based ME recognition (MER) methods achieved impressive success, these methods typically require pre-processing using conventional optical flow-based methods to extract facial motions as inputs. To overcome this limitation, we proposed a novel MER framework using self-supervised learning to extract facial motion for ME (SelfME). To the best of our knowledge, this is the first work using an automatically self-learned motion technique for MER. However, the self-supervised motion learning method might suffer from ignoring symmetrical facial actions on the left and right sides of faces when extracting fine features. To address this issue, we developed a symmetric contrastive vision transformer (SCViT) to constrain the learning of similar facial action features for the left and right parts of faces. Experiments were conducted on two benchmark datasets showing that our method achieved state-of-the-art performance, and ablation studies demonstrated the effectiveness of our method.

1. Introduction

Personality and emotions are crucial aspects of human cognition and play a vital role in human understanding and human-computer interaction [20]. Facial expressions provide an important cue for understanding human emotions [3]. According to neuropsychological research, micro-expressions (MEs) are revealed when voluntary and involuntary expressions collide [8]. As a slight leakage of expression, MEs are subtle in terms of intensities, brief in duration (occur less than 0.5 seconds), and affect small facial areas [3]. Because MEs are hard to be controlled, they are more likely to reflect genuine human emotions, and thus have been implemented in various fields, such as national security, political psychology, and medical care [41]. Despite the fact that MEs are valuable, their unique characteristics bring multifarious challenges for ME analysis.

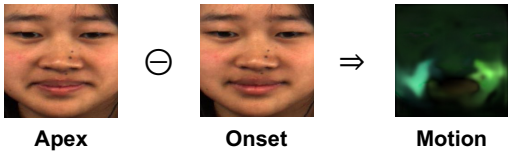


Figure 1. MEs may be imperceptible to the naked eye, but the motion between the onset (the moment when the facial action begins to grow stronger) and the apex (the moment when the facial action reaches its maximum intensity) makes them readily observable. SelfME learns this motion automatically.

The task of ME recognition (MER) is to classify MEs by the type of emotion [3]. Due to the small sample size of datasets, almost all methods for MER have used handcrafted features, which can be non-optical flow-based [2, 44, 50] or optical flow-based [25, 32, 42]. The extracted features were then fed into either traditional classification models [40, 43, 47], or deep learning-based classification models [12, 24, 51, 52]. Although recently proposed methods claimed they are deep learning-based methods, the ones with the highest performance often rely on traditional optical flow [10, 49] between the onset and apex frames as inputs. These optical flow methods are computed in a complex manner, and a number of researchers even proposed approaches for further processing the optical flow features prior to inputting them into the networks to improve their performance. This type of pipeline may hinder the development of MER in the deep learning era.

To overcome this limitation, we proposed a novel MER framework using self-supervised learning to extract facial motion for ME (SelfME) in this work. To the best of our knowledge, this is the first work with an automatically self-learned motion technique for MER. We visualized the learned motion by SelfME in Fig. 1. The symmetry of facial actions is important in MER, as spontaneous expressions are more symmetric than posed ones, or have intensity differences between left and right faces that are negligible [9,13]. However, the learned motion may suffer from ignoring symmetrical facial actions on the left and right sides of the face when extracting fine features. To address this issue, we developed a symmetric contrastive vision trans-

former (SCViT) to constrain the learning of similar facial action features for the left and right parts of the faces, mitigating asymmetry information irrelevant to MER. Experiments were conducted on two benchmark datasets, showing that our method achieved state-of-the-art performance. In addition, the ablation studies demonstrated the effectiveness of our method.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes our methodology. Section 4 presents the experiments with analysis and discussions. Limitations and ethical concerns are discussed in Section 5. Section 6 concludes the paper.

2. Related Work

2.1. Micro-Expression Recognition

Numerous methods were proposed for tackling MER, but the majority of them rely on handcrafted features [2, 44, 50]. To capture the dynamic texture, Zhao *et al*. proposed local binary patterns from three orthogonal planes (LBP-TOP) [50], which combined the local binary pattern (LBP) histogram in spatial and time dimensions. To capture textures in all directions and remove the redundant features, the hot wheel pattern with three orthogonal planes [2], and LBP with six intersection points [44] were proposed, respectively. Another group of researchers discovered the significance of optical flow features for MER and developed several enhanced optical flow extraction methods [25, 32, 42]. Patel *et al*. estimated optical flow features in selected regions and integrated them into spatiotemporal features [32]. Wang *et al*. proposed a main directional maximal difference algorithm to extract optical flow features in the maximal difference direction [42]. Liong *et al*. presented a bi-weighted approach for weighting the histograms of directed optical flow utilizing both magnitude and optical strain values (Bi-WOOF) [25].

One of the earliest attempts at deep learning-based MER methods was proposed in [31], where the authors employed several pretrained networks to extract ME features, followed by an evolutionary feature selection. Following this success, numerous methods based on deep learning were presented. Gan *et al*. proposed optical flow features from apex frame network (OFF-ApexNet) by feeding the horizontal and vertical optical flow features as two streams into the networks [12]. A shallow triple stream three-dimensional convolutional neural network (STSTNet) further added an optical strain feature to form three optical flow input streams [24]. Dual-Inception enhanced the OFF-ApexNet network by utilizing two Inception networks to process optical flows in horizontal and vertical directions [52]. On the basis of [52], Zhou *et al*. proposed a feature refinement network with an expression-shared and an expression-specific module for MER [51]. Although deep learning-based MER methods achieved impressive success, they still suffer from sophisticated preprocessing when using traditional optical flow extraction methods to extract facial motions as inputs. To alleviate this problem, we proposed a novel framework, SelfME, to learn the motion representation of ME in a self-supervised manner, moving a further step for transforming the ME pipeline to a fully end-to-end manner.

2.2. Self-Supervised Learning

Self-supervised learning is a powerful technique that allows for feature representation learning without the need for labeled data. Solving pretext tasks, such as colorization [18, 19], Jigsaw puzzle [29, 45], and ordering shuffled image patches [6], requires high-level understanding, so they can be used in self-supervised learning for learning generalized representations that can help with downstream learning tasks. Contrastive learning (CL) solves the self-supervised learning problem by pulling positive pairs closer while pushing negative pairs far apart [4, 14, 46]. When labels are available, supervised contrastive learning can further improve the performance by pulling the representations of positive samples from the same class together [16]. To address the issue of the learned motion representations suffering from ignoring symmetrical facial actions, we introduce a novel way of using contrastive learning by pulling together the left and right parts of the faces to preserve facial symmetry information.

For videos, self-supervised signals can be provided by reconstructing a target frame given a source frame in the same video clip, as done in self-supervised optical flow learning [26, 27]. This approach has been used in facial attribute learning [17] and facial action unit detection [22, 23]. The facial attributes network (FAB-Net) learns facial attributes by encoding source and target frames into embeddings and combining them to estimate the optical flow for reconstructing the target frame [17]. Twin-cycle autoencoder (TCAE) [23] and its variant [22], proposed for AU detection, improved FAB-Net by disentangling facial actions and head movements. However, MEs are subtle movements, so these approaches with less accurate estimation of flow may not be effective for MER. Our SelfME can extract fine-grained subtle motions by leveraging intrinsic keypoint detectors and fusing local motions into dense motions.

3. Methodology

3.1. SelfME Framework

In this work, we proposed a novel MER framework (SelfME) using self-supervised learning to extract facial motion representation, as depicted in Fig. 2. SelfME consists of two stages: a motion learning stage and a classification stage. In the motion learning stage, the self-supervised

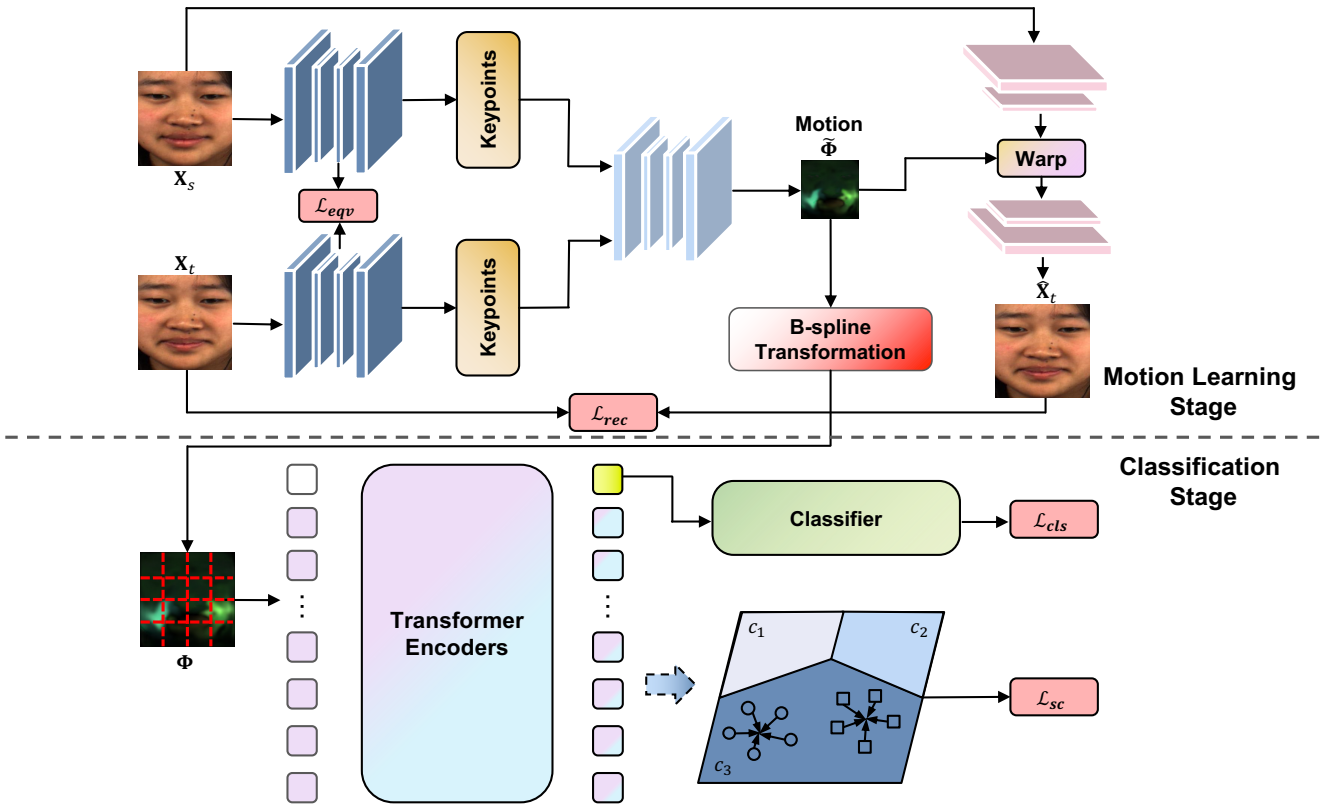


Figure 2. Framework of SelfME. SelfME consists of two stages: a motion learning stage and a classification stage. In the **motion learning stage**, the self-supervised motion learner was trained on a source frame $\mathbf{X}_s$ and a target frame $\mathbf{X}_t$ in ME sequences with a reconstruction objective, to warp the source frame $\mathbf{X}_s$ into the estimated target frame $\hat{\mathbf{X}}_t$. A B-spline transformation expands the size of the motion field to a similar size as the original image input, producing the improved motions $\mathbf{\Phi}$ that are fed into the classification stage. In the **classification stage**, an SCViT was used to constrain the learning of similar facial action features for the left and right parts of faces.

motion learner was trained on ME sequences using a reconstruction objective, so that the learned motion is capable of warping the source image into the estimated target image. However, the learned motions may be nonsmooth, folding, and of low resolution, which would affect the accuracy of MER. Thus, we applied a scaling-and-squaring cubic B-spline transformation to ensure the diffeomorphism of the motion field, and expand the size of the motion field. These motions were then fed into the classification stage. In the classification stage, we adopted an SCViT to let the network focus on small and subtle ME features. As learned motions may ignore symmetrical facial actions when extracting fine features, SCViT can also constrain the learning of similar facial action features for the left and right parts of faces.

3.2. Self-supervised Motion Learning

Given a sequence $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\} \in \mathbb{R}^{H \times W \times 3 \times T}$, our purpose is to learn the motion pattern between two frames. However, we do not have the ground truth motion for transforming one frame to another frame. Therefore, rather than directly estimating the motion between two frames in a fully-supervised manner, we formed an alternative self-supervised learning task by reconstructing the target frame with a given source frame, in which the motion was implicitly learned to achieve the motion learning task. The pipeline of self-supervised motion learning is shown as the motion learning stage in Fig. 2.

MEs are subtle in terms of intensities and cover limited face regions [3]. Typically, small movements occur around the eyes, eyebrows, and lips. To reduce the influence of noise and illumination variants on subtle movement extraction, we learned the motion pattern by focusing on key facial moving parts. In particular, we addressed the problem of learning the motion pattern through two subtasks: 1) detecting motion-related keypoints and computing their local motions; 2) obtaining a dense motion field by a weighted combination of the local motions.

Similar to [35], a source frame $\mathbf{X}_s$ and a target frame $\mathbf{X}_t$ were randomly selected from the sequence $\mathbf{X}$. K keypoints of the two frames are estimated respectively by a shared network $\mathcal{F}_{kp}$ as

$$\{\mathbf{z}_i^k\}_{k=1}^K = \mathcal{F}_{kp}(\mathbf{X}_i),\ i \in \{s, t\}, \tag{1}$$

where $\mathbf{z}_s^k$ and $\mathbf{z}_t^k$ are the k-th keypoint for the source and

target frames, respectively. Then, the local motions were computed as the location displacements of corresponding keypoints in $\mathbf{X}_s$ and $\mathbf{X}_t$:

$$\mathbf{\Phi}^k(\mathbf{u}) = \gamma(\mathbf{z}_s^k - \mathbf{z}_t^k),\ \forall \mathbf{u} \in U, \tag{2}$$

where γ is a motion amplification (MA) factor, and U is a lattice of size $\widetilde{H} \times \widetilde{W}$. $\widetilde{H}$ and $\widetilde{W}$ are one-fourth of the original image size H and W as the motion can be learned better in a smaller and compact space.

To compute the dense motion field $\widetilde{\mathbf{\Phi}}$ by a weighted combination of the local motions, we estimated weights $\psi^k(\mathbf{u})$ for each motion $\mathbf{\Phi}^k(\mathbf{u})$. The weights were estimated from aligned frames $\{\chi^k\}_{k=1}^K$ and heatmaps $\{\mathbf{H}^k\}_{k=1}^K$. The aligned frames $\{\chi^k\}_{k=1}^K$ are obtained by warping the source frame $\mathbf{X}_s$ according to the local motion fields in Eq. 2. The heatmaps $\{\mathbf{H}^k\}_{k=1}^K$ preserve the keypoint information in the form of a keypoint location confidence map, which is computed using the following equation:

$$\mathbf{H}^k(\mathbf{u}) = \exp(-(\mathbf{u} - \mathbf{\Phi}^k(\mathbf{u}))^\top \mathbf{\Sigma}^{-1}(\mathbf{u} - \mathbf{\Phi}^k(\mathbf{u}))), \tag{3}$$

where $\mathbf{\Sigma}$ is the covariance matrix. All aligned frames $\{\chi^k\}_{k=1}^K$ and heatmaps $\{\mathbf{H}^k\}_{k=1}^K$ are concatenated and passed to a network $\mathcal{F}_{dense}$ to generate the predicted weights ψ^k for each local motion as

$$\{\psi^k\}_{k=1}^K = \mathcal{F}_{dense}(\{\chi^k\}_{k=1}^K, \{\mathbf{H}^k\}_{k=1}^K). \tag{4}$$

Finally, we can obtain the dense motion $\widetilde{\mathbf{\Phi}}$ by a weighted summation of local motions as

$$\widetilde{\mathbf{\Phi}} = \sum_{k=1}^{K} \psi_k \mathbf{\Phi}^k. \tag{5}$$

To predict an estimated target frame $\hat{\mathbf{X}}_t$, we utilized a generator network that warps the feature of the source frame $\mathbf{X}_s$ as shown in the upper right part of Fig. 2. To learn the motion, we use a feature reconstruction loss and an equivariance loss. The feature reconstruction loss is similar to the perceptual loss [15] as

$$\mathcal{L}_{rec}^{\iota,m} = \frac{1}{n_m^\iota} \left| \mathcal{F}_{VGG}^{\iota,m}(\hat{\mathbf{X}}_t) - \mathcal{F}_{VGG}^{\iota,m}(\mathbf{X}_t) \right|, \tag{6}$$

where $\mathcal{F}_{VGG}^{\iota,m}$ refers to the m-th feature map of VGG-19 with ι-th scale inputs. n_m^ι is the corresponding number of pixels. The equivariance loss is used to constrain the keypoint detection task as

$$\mathcal{L}_{eqv} = |\mathcal{F}_{kp}(\mathcal{T}_{rand}(\mathbf{X}_s)) - \mathcal{T}_{rand}(\mathcal{F}_{kp}(\mathbf{X}_s))|, \tag{7}$$

where $\mathcal{T}_{rand}$ is a known random thin-plate spline transformation. Finally, we can obtain the final loss for the motion learning stage through a weighted summation as

$$\mathcal{L}_{ml} = \sum_{\iota} \sum_{m} (\lambda_{rec}^{\iota,m} \mathcal{L}_{rec}^{\iota,m}) + \lambda_{eqv} \mathcal{L}_{eqv}, \tag{8}$$

where $\{\lambda_{rec}^{m,\iota}\}$ and λ_{eqv} are the weights.

To prepare inputs for the classification stage, only the motion between the onset and apex frames needs to be extracted from each video sequence, while the self-supervised learning of the motion learning stage leverages all frames of the video. One problem is that $\widetilde{\mathbf{\Phi}}$ may be nonsmooth and folding, which would affect the accuracy of MER. In addition, the generated dense motion field $\widetilde{\mathbf{\Phi}}$ is smaller than the original image size. To address these problems, we adopt a scaling-and-squaring cubic B-spline transformation strategy to ensure diffeomorphism of the motion field and expand the motion field's size to be similar to the original image. Diffeomorphism preserves topology and guarantees invertibility, which would be beneficial for MER. The expanded motion field $\mathbf{\Phi}$ is obtained using a weighted combination of cubic B-spline basis functions $\beta(\cdot)$ [33]:

$$\mathbf{\Phi}(\mathbf{v}) = \sum_{\mathbf{u} \in G} \widetilde{\mathbf{\Phi}}(\mathbf{u}) \prod_{\delta=1}^{2} \beta_{\mathbf{u}}^{\delta}(\mathbf{v}^\delta - \mathbf{u}^\delta), \tag{9}$$

where $\mathbf{v}^\delta$ is the δ-th coordinate of the point $\mathbf{v}$ in the expanded motion field $\mathbf{\Phi}$, and $\mathbf{u}^\delta$ is the δ-th coordinate of the point $\mathbf{u}$ in the control grid G. Then, the scaling and squaring algorithm [1] is adopted to generate the group exponential of the stationary velocity field for $\mathbf{\Phi}$ to ensure diffeomorphism. The diffeomorphic motion field generated after the scaling and squaring algorithm is also denoted as $\mathbf{\Phi}$.

3.3. Symmetric Contrastive Vision Transformer

Spontaneous expressions are more symmetric than posed ones, or have intensity differences between left and right faces that are negligible [9, 13]. However, the learned motion does not show such symmetry. To tackle this problem, we developed an SCViT to preserve the facial symmetric information. Our motivation for imposing the symmetric constraint is to improve the discrimination of ViT by preserving the facial geometry destroyed by using patches as input, thus learning similar representations for left and right faces, and mitigating asymmetry information irrelevant to MER. A graphical illustration of symmetric contrastive (SC) is shown in Fig. 3.

Given the need of local attention capability and tiny details in the ME images, we adopted ViT [7] to focus on the small crucial regions for extracting subtle ME features. Transformers received 1D sequences as inputs. To handle 2D inputs, a split operation is used to embed the image (motion) $\mathbf{\Phi} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\mathbf{\Phi}_p = \{\mathbf{\Phi}_p^i\}_{i=1}^N \in \mathbb{R}^{N \times (H_p W_p C)}$, where $H \times W$ is the original image resolution, C is the number of channels, $H_p \times W_p$ is the patch resolution, and $N = HW/H_pW_p$ is the number of generated patches. Each flattened patch $\mathbf{\Phi}_p^i$ is then linearly projected with trainable weights $\mathbf{E} \in$

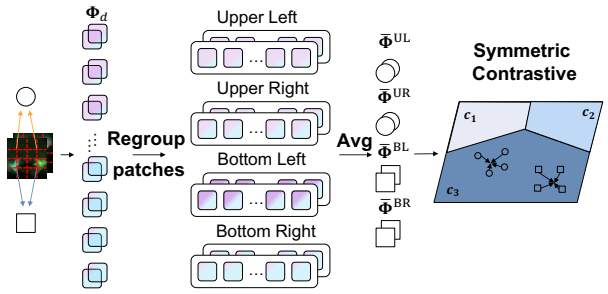


Figure 3. The graphical illustration of SC with a batch of two samples. To mimic the facial action coding process, the patches $\mathbf{\Phi}_d$ are regrouped according to the upper left, upper right, bottom left, and bottom right regions. Then, the patches are averaged within each group, generating 4 regional embeddings $\bar{\mathbf{\Phi}}^{\mathrm{UL}}$, $\bar{\mathbf{\Phi}}^{\mathrm{UR}}$, $\bar{\mathbf{\Phi}}^{\mathrm{BL}}$, and $\bar{\mathbf{\Phi}}^{\mathrm{BR}}$. The SC loss is then applied to symmetrical regions of the same image and across the samples within the batch.

$\mathbb{R}^{(H_p W_p C)\times D}$ to a D-dimensional feature vector $\mathbf{\Phi}_d^i$:

$$\mathbf{\Phi}_d = [\mathbf{\Phi}_p^1\mathbf{E};\, \mathbf{\Phi}_p^2\mathbf{E}; \cdots ;\, \mathbf{\Phi}_p^N\mathbf{E}], \tag{10}$$

$$= [\mathbf{\Phi}_d^1;\, \mathbf{\Phi}_d^2; \cdots ;\, \mathbf{\Phi}_d^N]. \tag{11}$$

However, these patches are directly fed to the transformer blocks in ViT, which do not preserve facial geometry features and may violate the prior knowledge that most ME are symmetrical. To overcome this limitation, we explicitly regularized the learning of ME feature representations by introducing an SC loss. The SC loss is designed to mimic the facial action coding process [11]. The patches $\mathbf{\Phi}_d$ are regrouped according to the upper left (UL), upper right (UR), bottom left (BL), and bottom right (BR) regions of the original input images. Then, the patches of each group are averaged to form a single embedding $\bar{\mathbf{\Phi}}^{\alpha}$ which represents the region α as

$$\mathbf{\Phi}^{\alpha} = [\mathbf{\Phi}_d^{\alpha_1};\, \mathbf{\Phi}_d^{\alpha_2}; \cdots ;\, \mathbf{\Phi}_d^{\alpha_{N_0}}], \tag{12}$$

$$\bar{\mathbf{\Phi}}^{\alpha} = \frac{1}{N_0}\sum_{i=1}^{N_0} \mathbf{\Phi}_d^{\alpha_i}, \tag{13}$$

where $\alpha \in R = \{\mathrm{UL}, \mathrm{UR}, \mathrm{BL}, \mathrm{BR}\}$ represents the index set for the corresponding regions, and $N_0 = N/4$ is the number of patches for each region.

To increase the similarity between the left and right facial action features, we constrained the region symmetry of ME and the samples in a batch. Let $i \in I$ be the index of an arbitrary sample. The proposed SC loss can be expressed as

$$\mathcal{L}_{sc} = \sum_{i\in I} \frac{-1}{|P(i,\alpha,\beta)|} \sum_{\alpha,\beta\in R} \sum_{p\in P(i,\alpha,\beta)} \log \frac{\exp\left(\bar{\mathbf{\Phi}}_i^{\alpha}\cdot\bar{\mathbf{\Phi}}_p^{\beta}/\tau\right)}{\sum_{a\in A(i,\alpha,\beta)} \exp\left(\bar{\mathbf{\Phi}}_i^{\alpha}\cdot\bar{\mathbf{\Phi}}_a^{\beta}/\tau\right)}, \tag{14}$$

where $A(i,\alpha,\beta) \equiv \{a \in I : a \neq i \text{ if } \alpha = \beta\}$. $P(i,\alpha,\beta) \equiv \{p \in I : (\mathbf{y}_p = \mathbf{y}_i) \cap (\alpha,\beta \text{ identical or symmetrical})\}$ is the set of indices of all positive regrouped embeddings. The symmetrical pairs are UL-UR and BL-BR. $|P(i,\alpha,\beta)|$ is the cardinality of $P(i,\alpha,\beta)$. τ is the temperature parameter. This loss is inspired by supervised contrastive learning loss [16], which could also push away the samples with different classes.

To perform the classification task, a learnable class representation $\mathbf{\Phi}_{cls}$ is prepended to the embedded patches obtained by Eq. 11. In addition, another learnable position embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$ are added to the patch embeddings to retain positional information. These embeddings are then fed into a transformer encoder to learn patch-level relationships [39]. The transformer encoder comprises multiple alternating layers of a multi-head self-attention (MSA) module and a feed-forward (FF) module. The final output of the corresponding class representation $\mathbf{z}_{cls}$ is passed through a layer norm (LN) and a multi-layer perceptron (MLP) with one hidden layer for mapping the output dimension to the desired number of classes as

$$\hat{\mathbf{y}} = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}_{cls})). \tag{15}$$

We then used a standard multi-class cross-entropy loss $\mathcal{L}_{cls}$ computed between the predicted $\hat{\mathbf{y}}$ and the ground truth $\mathbf{y}$. The final loss for the classification stage is

$$\mathcal{L}_c = \mathcal{L}_{cls} + \omega\mathcal{L}_{sc}, \tag{16}$$

where ω is a trade-off weight for balancing the classification and the SC objectives.

4. Experiment

4.1. Dataset

CASME II. CASME II [48] has 35 participants and 247 ME sequences. Videos are at 200 frames per second (FPS) and resolutions of 640 × 480 with 280 × 340 for the face.

SMIC-HS. SMIC-HS [21] has 164 spontaneous ME clips at 100 FPS from 16 individuals at the same resolution as CASME II, but with 190 × 230 for the face.

4.2. Evaluation Metrics

Experiments were conducted using a leave-one-subject-out cross-validation setting. Each time, samples from one subject were set aside for testing, while all remaining samples were utilized for training. The experiment was repeated S times, where S is the total number of subjects. The unweighted F1-score (UF1) and unweighted average recall (UAR) were used to measure the performance. UF1 is calculated as the average F1 across all classes n_c as

$$\mathrm{UF1} = \frac{1}{n_c}\sum_j \mathrm{F1}_j, \tag{17}$$

Table 1. Ablation study of key components. SC plays a crucial role by pulling together the left and right facial action features, while B-spline transformation and MA also bring some improvements.

B-spline	SC	MA	UF1	UAR
-	-	-	0.8468	0.8849
✓	-	-	0.8629	0.8903
-	✓	-	0.8903	0.9028
-	-	✓	0.8718	0.8851
✓	✓	-	0.8923	0.8984
-	✓	✓	0.8951	0.9109
✓	-	✓	0.8784	0.8960
✓	✓	✓	**0.9078**	**0.9290**

where $\text{F1}_j = \frac{2\sum_{s=1}^{S}\text{TP}_j^s}{2\sum_{s=1}^{S}\text{TP}_j^s+\sum_{s=1}^{S}\text{FP}_j^s+\sum_{s=1}^{S}\text{FN}_j^s}$ is the F1 for the j-th class. UAR is defined as the average recall as

$$\text{UAR} = \frac{1}{n_c}\sum_j \text{Recall}_j, \tag{18}$$

where $\text{Recall}_j = \frac{\sum_{s=1}^{S}\text{TP}_j^s}{\sum_{s=1}^{S}\text{FP}_j^s+\sum_{s=1}^{S}\text{FN}_j^s}$. TP_j^s, FP_j^s, and FN_j^s are true positives, false positives, and false negatives for the j-th class of the s-th subject.

4.3. Implementation Detail

The PyTorch framework [30] was used. The model for motion learning was pre-trained on the VoxCeleb dataset [28] and fine-tuned on the ME datasets by randomly selecting two frames each time. SCViT was initialized with pre-training from ImageNet and trained with motions extracted between the onset and apex frames. All encoder-decoder-like networks are Hourglass networks. The ViT is ViT-S/16. They were optimized by an adaptive moment (Adam) optimizer with an initial learning rate of 2×10^{-6} and 1×10^{-4}, respectively. A multistep learning rate scheduler was utilized to dynamically reduce the learning rate by 10 at epochs 20 and 30 for motion learning. The SCViT was trained with a batch size of 32 and exponential learning rate decay with a factor of 0.9 for 60 epochs per experiment. The MA factor γ was always set as 1 during training, while it was empirically set as 2 during inference (See Section 4.7 for more discussions). A high-performance computer with 4 CPU cores, 1 NVIDIA V100 GPU card, and 32 GB memory was utilized for training the model.

4.4. Ablation Study

Experiments were conducted on CASME II in order to evaluate the contribution of B-spline transformation, SC, and MA. As indicated by Table 1, the following observations can be made:

(1) SC plays the most important role. When only one component is employed, it outperformed the b-spline by 2.74% in UF1 and 1.25% in UAR, and outperformed MA by 1.85% in UF1 and 1.77% in UAR. This suggests that pulling the left and right facial action features together in a contrastive manner is beneficial for MER, given ME contains relatively symmetric expressions, but learned motions may not show this.

(2) When combining the learned motion with B-spline transformation, performance improved by 0.20% in UF1 over only using the SC loss. This finding demonstrates that improving the motion's diffeomorphism by B-spline transformation when expanding the motion's size is advantageous for MER.

(3) MA also brings some improvements. Although it amplifies the motions to make the ME-associated motions more distinct, it may also amplify other motions that are not related to ME. Further discussion on MA is in Section 4.7.

4.5. Comparison with the State-of-the-art

We compared SelfME with existing state-of-the-art methods on two popular ME benchmarks in Table 2. SelfME stands out as the only and first method to utilize self-supervised motion representation, while most of the other methods rely on dense optical flow estimation using total variation regularization with $L1$-norm (TV-L1) regularization [49]. LBP-TOP [50] and Bi-WOOF [25], which are traditional methods without any deep learning techniques, produced comparatively lower performance than deep learning-based models. CapsuleNet, which utilized a single apex frame as input rather than motion, performed significantly worse than other concurrent methods. These results demonstrate the importance of motion-based approaches in MER and show the potential of self-supervised learning techniques.

SelfME achieved the highest performance on CASME II, surpassing the second-best method by 4.17% in UAR and 1.63% in UF1, with scores exceeding 90% in both metrics. On SMIC-HS, FeatRef [51] took the lead, while SelfME delivered a similar performance. It is worth noting that the comparison with fully-supervised methods on SMIC-HS might not be entirely fair. When considering the average results on the two benchmarks, SelfME achieved the best overall performance among all methods, showing its effectiveness.

4.6. Impact of the Learned Motion

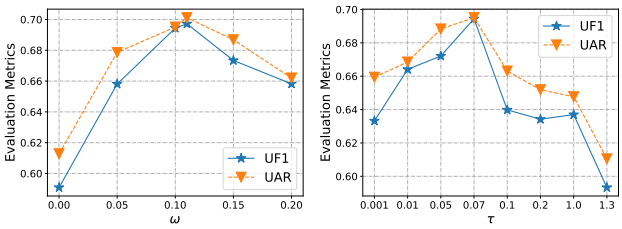
To analyze the impact of the learned motion by SelfME, SelfME was separated into SelfME's motion and SCViT for experiments on CASME II (Table 3). In the first part, SelfME's motion was compared to TV-L1 motion with either ViT or SCViT as the classifier, showing SelfME's motion was superior. In the second part, we showed that highly accurate motion estimation is more essential for ME than other facial tasks, such as AU recognition, by adopt-

Table 2. Comparison with the state-of-the-art methods. Experiments were conducted on CASME II and SMIC-HS datasets with three classes: Negative, Positive, and Surprise. Among all methods, our SelfME is the only and first method using self-supervised learned motion as inputs, whereas the majority of methods use a conventional optical flow extractor. LBP: local binary pattern; Apex: apex frame only; TV-L1: optical flow generated by the total variation regularization with $L1$-norm approach; Learned: self-supervised learned motion.

Method	Input	CASME II		SMIC-HS		Average	
		UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [50]	LBP	0.7026	0.7429	0.2000	0.5280	0.4513	0.6355
CapsuleNet [38]	Apex	0.7068	0.7018	0.5820	0.5877	0.6444	0.6448
Bi-WOOF [25]	TV-L1	0.7805	0.8026	0.5727	0.5829	0.6766	0.6928
GoogLeNet [37]	TV-L1	0.5989	0.6414	0.5123	0.5511	0.5556	0.5963
VGG16 [36]	TV-L1	0.8166	0.8202	0.5800	0.5964	0.6983	0.7083
OFF-ApexNet [12]	TV-L1	0.8764	0.8680	0.6817	0.6695	0.7791	0.7688
Dual-Inception [52]	TV-L1	0.8621	0.8560	0.6645	0.6726	0.7633	0.7643
STSTNet [24]	TV-L1	0.8382	0.8686	0.6801	0.7013	0.7592	0.7850
FeatRef [51]	TV-L1	0.8915	0.8873	**0.7011**	**0.7083**	0.7963	0.7978
SelfME	Learned	**0.9078**	**0.9290**	0.6972	0.7012	**0.8025**	**0.8151**

Table 3. Analysis of the impact of the learned motion. Learned motion of SelfME demonstrated superior performance than the motion derived from TV-L1 and TCAE.

Method	UF1	UAR
TV-L1+ViT	0.8060	0.8016
TV-L1+SCViT	0.8460	0.8305
TCAE+FC [23]	0.4836	0.5491
TCAE's motion+ViT	0.5681	0.5752
TCAE's motion+SCViT	0.6158	0.5926
SelfME's motion+ViT	0.8784	0.8960
SelfME's motion+SCViT	0.9078	0.9290



(a) Hyperparameter ω. (b) Hyperparameter τ.

Figure 4. Hyperparameter analysis on the SMIC-HS dataset. (a) The evaluation metrics for different ω with $\tau = 0.07$. (b) The evaluation metrics for different τ with $\omega = 0.1$.

ing TCAE for comparison. Initially, we used the exact pipeline of TCAE to encode faces into a low-dimensional facial action vector followed by FC for classification, but the performance was not satisfactory. We attribute this to the fact that the low-dimensional representation vector is insufficient for ME, which requires subtle motions. Then, we modified TCAE to obtain its motion field for classification, but its performance still did not improve significantly. While TCAE's advantages of disentangling facial actions and head movements may be useful for other facial tasks, they are not applicable to MER. Its less accurate motion field causes it to fail in MER. SelfME's superior performance is attributed to its ability to extract subtle and fine-grained motions, highlighting the importance of learning accurate motion representation for MER.

4.7. Hyperparameter Analysis

Trade-off Weight ω. ω balances the classification objective and the degree of symmetric contrast. Fig. 4a shows that the optimal ω is around 0.1, with 0.11 being the optimal. When ω approaches zero, the degree of SC reduces, resulting in a decline in performance. When ω grows excessively large, the performance decreases because the ultimate classification objective is affected.

Sharpen Temperature τ. τ provides the flexibility to slightly modify the entropy of $\mathbf{\Phi}_i$ in Eq. 14. Fig. 4b shows the effect with different τ. When $\tau > 1$, the distribution flattens, which reduces the model's sensitivity to predictions. When $\tau < 1$, it is equivalent to optimizing for hard positives and negatives, applying stronger constraints of the similarity between the left and right sides of the facial action features. The best temperature is $\tau = 0.07$.

Motion Amplification Factor γ. When the motion is extracted, we can amplify the motion with a factor γ before inputting it into the classification stage. The effects of MA are shown in Table 4. When we amplify the motion by a factor of 2, we observe a prominent increase, because the subtle ME motions are amplified to be observed easily. However, when the motion is amplified by 3 times, the performance drops significantly. The results suggest that when the motions are amplified too much, their distinct pattern might disappear and the noise will be amplified as well.

Table 4. Hyperparameter analysis on the SMIC-HS dataset for MA (γ). There is a notable increase when the motion is amplified by a factor of 2, since the subtle ME motions are amplified so that they can be detected easily.

MA (γ)	UF1	UAR
$\times$ 1	0.6768	0.6798
$\times$ 2	**0.6972**	**0.7012**
$\times$ 3	0.6523	0.6622

4.8. Visualization

We conducted visualizations of the motion learned by self-supervision, and a gradient-weighted class activation mapping (GradCAM) visualization [34] in Fig. 5.

Learned Motion by Self-supervision. We present the learned motion representations obtained through self-supervision in column 3 of Fig. 5. We observe that the pulling up of lip corners, which is challenging to detect in the positive ME sequence with onset and apex frames, is clearly visible in the learned motion despite some noise. Similarly, the depression of the lip corners, which is a key feature of negative ME, is not readily discernible in the original images but is evident in the learned motion. The learned motions capture critical facial features around the eyes and mouth regions, where MEs typically manifest. Although the learned motions appear to be sufficient, there is a distinct discrepancy in motion intensity between the left and right sides, which may affect MER. In order to address this issue, we propose SC to constrain the learning of similar features for the left and right parts of facial action features.

Symmetric Contrastiveness. To provide further evidence of the effectiveness of SC, we utilized GradCAM to visualize the attention maps in columns 4 and 5 of Fig. 5. For the positive ME, the model with SC exhibited better symmetrical attention around the corners of the lips, while the model without SC did not show this behavior and was confused by the noisy motions. For the negative ME, the model with SC constraints demonstrated improved symmetrical attention for the positive lip corner pulling-up action, whereas the method without SC constraints was misled by the eye-opening action, which did not contribute to the negative ME. These visualizations demonstrate that SC effectively encourages the model to focus on symmetrical features, which are more indicative of MEs.

5. Limitation and Ethical Concern

Limitation. SelfME cannot currently handle grayscale input like SAMM dataset [5]. The disparity between color and grayscale spaces may cause instability in training and prevent accurate extraction of motions. SelfME requires detection of the onset and apex in advance. If this is inaccurate, performance may be poor. Recognition of non-frontal faces, where symmetrical compensation may not work, was not evaluated and this could significantly impact performance. Future work will focus on developing more robust and efficient methods to address these limitations.

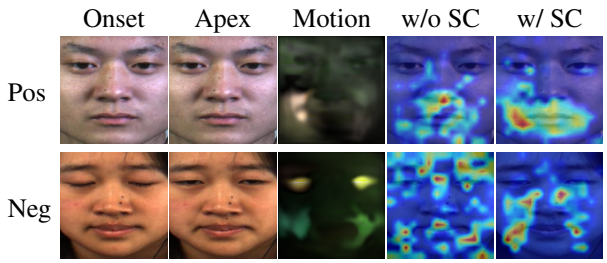


Figure 5. Motion representation learned by self-supervision, and GradCAM visualization with (w/) and without (w/o) SC of the positive (Pos) and negative (Neg) MEs. Although the motions seem satisfactory, there is a distinct difference between the left and right motion intensities. SC demonstrated better symmetrical attention to the facial actions on the left and right sides of the faces, and was robust to noisy actions that did not contribute to the corresponding MEs.

Ethical Concern. SelfME may be biased due to biases in collecting and labeling the training data. Current ME data are lab-controlled scenarios with subjective labeling by annotators. Deep learning models could learn those biases and may be inaccurate for some demographic groups. Personal and sensitive information could be revealed by ME, so informed consent is needed for their ethical development and deployment. Weights in these models are patterns learned from the data and may reveal sensitive information about individuals. The privacy of the raw data and the learned patterns in deep learning models must be guarded. Secure model storage and privacy techniques are needed to prevent leakage of sensitive information. Addressing these concerns is crucial to ensure the ethical development and deployment of MER systems.

6. Conclusion

In this study, we presented the SelfME framework for MER. SelfME advances the pipeline for MER by using self-supervised motion representation instead of sophisticated traditional optical flow inputs. In addition, by pulling together the representations of left and right facial motions, our SCViT demonstrated superior performance. Effective results were demonstrated on CASME II and SMIC-HS datasets. In the future, we will address the limitations, and expand the SelfME framework to create a unified pipeline for both ME spotting and MER tasks.

Acknowledgement

This work is supported by Hong Kong Research Grants Council (Project 11204821) and Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

References

[1] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006. 4

[2] Xianye Ben, Xitong Jia, Rui Yan, Xin Zhang, and Weixiao Meng. Learning effective binary descriptors for micro-expression recognition transferred by macro-information. *Pattern Recognition Letters*, 107:50–58, 2018. 1, 2

[3] 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

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[5] 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, 2016. 8

[6] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015. 2

[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 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4

[8] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969. 1

[9] Paul Ekman, Joseph C Hager, and Wallace V Friesen. The symmetry of emotional and deliberate facial actions. *Psychophysiology*, 18(2):101–106, 1981. 1, 4

[10] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, pages 363–370. Springer, 2003. 1

[11] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978. 5

[12] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. Off-apexnet on micro-expression recognition system. *Signal Processing: Image Communication*, 74:129–139, 2019. 1, 2, 7

[13] Joseph C Hager and Paul Ekman. The asymmetry of facial actions is inconsistent with models of hemispheric specialization. *Psychophysiology*, 22(3):307–318, 1985. 1, 4

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[15] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 4

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2, 5

[17] A Sophia Koepke, Olivia Wiles, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *British Machine Vision Conference*, page 302, 2018. 2

[18] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016. 2

[19] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017. 2

[20] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

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

[22] Yong Li, Jiabei Zeng, and Shiguang Shan. Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):302–317, 2020. 2

[23] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019. 2, 7

[24] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional cnn for micro-expression recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–5. IEEE, 2019. 1, 2, 7

[25] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018. 1, 2, 6, 7

[26] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019. 2

[27] Pengpeng Liu, Michael R Lyu, Irwin King, and Jia Xu. Learning by distillation: a self-supervised learning framework for optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5026–5041, 2021. 2

[28] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. *Annual Conference of the International Speech Communication Association*, pages 2616–2620, 2017. 6

[29] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 2

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 6

[31] Devangini Patel, Xiaopeng Hong, and Guoying Zhao. Selective deep features for micro-expression recognition. In *International Conference on Pattern Recognition*, pages 2258–2263. IEEE, 2016. 2

[32] Devangini Patel, Guoying Zhao, and Matti Pietikäinen. Spatiotemporal integration of optical flow vectors for micro-expression detection. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 369–380. Springer, 2015. 1, 2

[33] Huaqi Qiu, Chen Qin, Andreas Schuh, Kerstin Hammernik, and Daniel Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. In *Medical Imaging with Deep Learning*, 2021. 4

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[35] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32:7137–7147, 2019. 3

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 7

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 7

[38] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. Capsulenet for micro-expression recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–7. IEEE, 2019. 7

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 5

[40] Su-Jing Wang, Hui-Ling Chen, Wen-Jing Yan, Yu-Hsin Chen, and Xiaolan Fu. Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural Processing Letters*, 39(1):25–43, 2014. 1

[41] 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. 1

[42] Su-Jing Wang, Shuhang Wu, Xingsheng Qian, Jingxiu Li, and Xiaolan Fu. A main directional maximal difference analysis for spotting facial movements from long-term videos. *Neurocomputing*, 230:382–389, 2017. 1, 2

[43] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, Chun-Guang Zhou, Xiaolan Fu, Minghao Yang, and Jianhua Tao. Micro-expression recognition using color spaces. *IEEE Transactions on Image Processing*, 24(12):6034–6047, 2015. 1

[44] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition. In *Asian Conference on Computer Vision*, pages 525–537. Springer, 2014. 1, 2

[45] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1910–1919, 2019. 2

[46] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[47] Feng Xu, Junping Zhang, and James Z Wang. Microexpression identification and categorization using a facial dynamics map. *IEEE Transactions on Affective Computing*, 8(2):254–267, 2017. 1

[48] 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):e86041, 2014. 5

[49] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime TV-L1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007. 1, 6

[50] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. 1, 2, 6, 7

[51] 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. 1, 2, 6, 7

[52] Ling Zhou, Qirong Mao, and Luoyang Xue. Dual-inception network for cross-database micro-expression recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–5. IEEE, 2019. 1, 2, 7