Highlights

Micro-Expression Recognition using Dual-View Self-Supervised Contrastive Learning with Intensity Perception

Jingting Li, Haoliang Zhou, Yu Qian, Zizhao Dong, Su-Jing Wang

- A psychological experiment is designed to estimate the perceptual threshold for facial action intensity with consistent intensity filtering rules, serving as a reference for developing recognition models.
- Difference calculations is employed to preserve facial action information across multi-modalities.
- Dual-viewpoint contrastive learning based on different source modalities is implemented to extract high-level facial action features in a self-supervised manner, addressing the small sample size issue in micro-expression recognition from both algorithm design and input dimension perspectives.
- The effectiveness of our proposed method in enhancing recognition performance is demonstrated through comprehensive experiments.

Micro-Expression Recognition using Dual-View Self-Supervised Contrastive Learning with Intensity Perception

Jingting Li[a,b], Haoliang Zhou[c], Yu Qian[d,a], Zizhao Dong[a], Su-Jing Wang[a,b,*]

[a] *Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China*
[b] *Department of Psychology, University of the Chinese Academy of Sciences, Beijing, 100049, China*
[c] *Tianjin University of Technology, Tianjin, 300382, China*
[d] *Jiangsu University of Science and Technology, Zhenjiang, 212100, China*

Abstract

Micro-expressions, as indicators of true emotions, have significant applications in medical care and public safety. These expressions are characterized by their short duration, low intensity, and localized occurrence. These characteristics lead to the small sample problem in micro-expressions, making feature learning challenging and limiting the improvement of recognition performance. To address this issue, we propose a multimodal contrastive learning pre-training model based on Action Unit (AU) intensity perception. We conducted an experiment to determine the minimum threshold for recognizing facial expressions. Using this threshold, we filtered a large volume of unsupervised samples. The first stage involves unsupervised multimodal contrastive learning, where the model learns from differences in facial actions across various modalities. Subsequently, the model is trained on the micro-expression recognition task using a small amount of labeled data, overcoming the limitations of small sample sizes. Comparative experiments using the MEGC2019-CD and the multimodal dataset CAS(ME)^3 datasets demonstrate the superiority of our method. Our method is available at https://github.com/MELABIPCAS/DVSCL.git.

Keywords: Micro-Expression, Small Sample Size Problem, Contrastive

*Corresponding author: Su-Jing Wang (wangsujing@psych.ac.cn)

Learning, Self-supervised Learning.

1. Introduction

Facial expressions are crucial nonverbal cues for communication [20, 45]. However, when individuals attempt to conceal their true emotions, facial expressions become unreliable indicators. Micro-expressions, which are brief and low in intensity, are believed to be involuntary expressions of genuine emotions. The analysis of micro-expressions has potential applications in fields such as healthcare, national security, and financial risk assessment.

The identification of micro-expressions by the naked eye is a time-consuming and challenging task, even for those with professional training, with an accuracy rate of only about 50%. Therefore, the development of intelligent analysis of micro-expressions is necessary to better capture individuals' genuine emotions. Since the release of the CASME [60] and SMIC [30] databases in 2013, the algorithmic research on micro-expression recognition has gradually emerged. However, to date, only the 12 spontaneous micro-expression databases have been released (CASME [60], CASME II [59], CAS(ME)^2 [41], CAS(ME)^3 [27], SMIC [30], SMIC-E [48], 4DME [28], SAMM [7], SAMM-LV [61], MMEW [2], DFME [65] and MEVIEW [17]), with a total data volume of around 10,000. This is a typical small sample problem, which greatly limits the development of deep learning in the field of micro-expressions. In addition, the three characteristics of micro-expressions, i.e., short duration, subtle movement and local appearance, also make it challenging for deep learning algorithms to extract their features.

Despite the challenges posed by small sample sizes, micro-expression research has continued to grow. Initially, micro-expression recognition methods relied on manual feature extraction combined with traditional machine learning. This later evolved to deep learning models incorporating manually crafted features, and there have been attempts at end-to-end deep learning approaches. Transfer learning methods, such as knowledge distillation and teacher-student networks, have shown promising results by transferring knowledge from macro-expression recognition tasks to micro-expressions. However, these methods still depend heavily on annotated data and require training on specific macro-expression databases. Additionally, macro-expressions and micro-expressions do not always exhibit a one-to-one correspondence. Therefore, the use of self-supervised learning networks, which

can bypass the reliance on annotated data and instead learn features from large-scale, low-cost datasets, is an ideal solution to the small sample problem in micro-expression research.

Moreover, when micro-expression sample sizes are limited, a multi-modal, multi-view approach allows for deeper exploration of micro-expression features by expanding the feature space. Specifically, for the same micro-expression, different signal acquisition modalities—such as RGB, depth, and thermal imaging—can capture different aspects of its movement patterns. These complementary modalities help the network learn more effectively by providing information that may be missing from a single modality. Using self-supervised contrastive learning on multimodal datasets opens up new avenues for research under the constraints of limited sample sizes.

Based on the assumption that a powerful feature representation should model invariance across different viewpoints, we propose a dual-viewpoint self-supervised contrastive learning (DVSCL) framework. As illustrated in Fig. 1, this framework aims to maximize the mutual information of features for the same expression across different representation modalities. We first obtain the perceptual threshold of facial action units (AUs) through a psychological experiment. Next, we screen facial action video intervals through unsupervised intensity detection. By calculating differences, we highlight facial action information while removing ID information. In the self-supervised pretext task stage, we effectively learn facial action features through multi-modal contrastive learning based on numerous samples. Subsequently, we perform downstream recognition tasks on multi-modal micro-expression samples.

In summary, our contributions are: 1) designing a psychological experiment to estimate the perceptual threshold for facial action intensity with consistent intensity filtering rules, serving as a reference for developing recognition models; 2) employing difference calculations to preserve facial action information across multi-modalities; 3) Using multi-viewpoint contrastive learning to extract high-level facial action features from different source modalities in a self-supervised manner, addressing the small sample size issue in micro-expression recognition from both algorithm design and input dimension perspectives. Notably, this is the first time depth information has been used to construct a self-supervised model for micro-expression recognition; and 4) demonstrating the effectiveness of our module in enhancing recognition performance through comprehensive experiments.

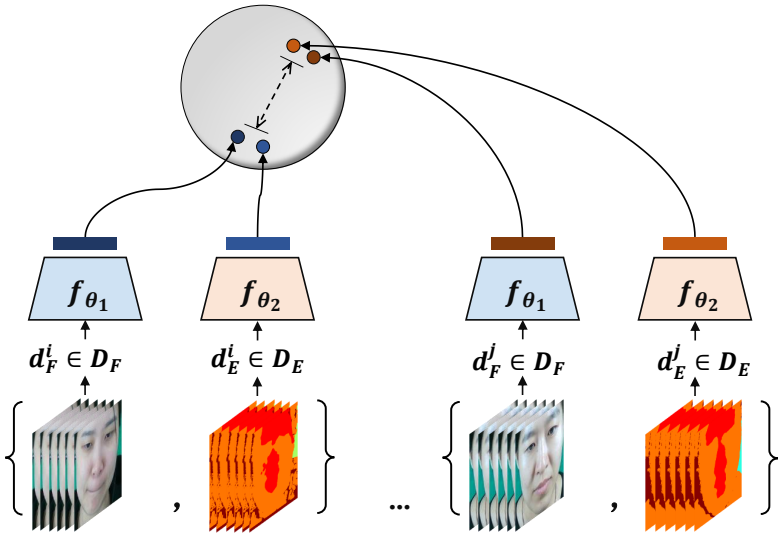


Figure 1: Training pipeline of the proposed DVSCL framework.

2. Related work

In this section, we will systematically review the advances in micro-expression recognition methods and introduce the related research on the contrastive learning.

2.1. Micro-expression recognition

Currently, the technology for micro-expression recognition is relatively mature in experimental environments. Methods can be broadly categorized into handcrafted feature methods and deep learning methods. The most common handcrafted feature methods are based on LBP-TOP, HOG, and optical flow. In terms of texture feature-based methods, Li et al. used a variant of Histogram of Oriented Gradients (HOG) for micro-expression recognition [29]. Le Ngo et al. used Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) and sparsity constraints to learn the important temporal features and spectral structures of micro-expressions [23]. Huang et al. constructed a discriminative LBP-TOP based on integral projection to enhance the inter-class differences of micro-expression classification [16]. Wang et al. proposed to extract LBP-TOP features separately in the color space to enhance micro-expression recognition performance [50]. For optical flow-based methods, Liu et al. proposed the MDMO method to describe

micro-expressions [36]. Liong used a Bi-Weighted Oriented Optical Flow (Bi-WOOF) feature extractor [35], while Happy et al. proposed a fuzzy HOOF method that ignored small motion amplitudes and only extracted features based on motion directions for micro-expression recognition [14]. These methods greatly improve the recognition performance of micro-expressions, but their robustness is not strong due to the short duration and low intensity of micro-expressions.

In recent years, combining deep learning with micro-expression recognition has become the main trend, leading to continuous improvements in recognition rates. Gupta et al. proposed effective feature encodings and a 2D convolutional neural network to recognize micro-expression [12]. Zhang et al. proposed a micro-expression recognition method based on Transformer spatiotemporal feature extraction. [64]. Mao et al. achieved objective category-based micro-expression recognition under partial occlusion using a region-based heuristic relation inference network [37]. Cai et al. proposed the MFDA network, which utilizes multi-level flow-driven attention to capture subtle motion variations in micro-expressions [3]. Gan et al. combines attention mechanisms and long short-term memory (LSTM) networks to capture subtle micro-expression features from facial sequences [10]. Wang et al. employs a hybrid transformer network to capture both global and local temporal dynamics of micro-expressions [52]. Furthermore, since micro-expressions are subtle and localized movements, some networks focus on action unit feature extraction to improve micro-expression recognition performance, such as [25, 31, 32, 68]. To further solve the small sample problem of micro-expressions, many methods have introduced transfer learning to enhance the performance of network feature extraction for micro-expressions [49, 44, 56].

While many micro-expression recognition algorithms are intricately designed to handle the unique characteristics of micro-expressions, these methods generally rely on labeled data, and their robustness is limited. Due to the subtle nature of micro-expressions, these algorithms can be heavily influenced by environmental conditions during data collection, and the labeling process may introduce inaccuracies. Furthermore, the small sample size inherent to micro-expression datasets exacerbates these issues. Exploring the use of self-supervised learning on large-scale, unlabeled facial motion data presents a promising research direction. Additionally, most existing methods focus solely on RGB/Gray signals, leaving multi-modal approaches underexplored. Leveraging different data modalities for multi-view learning represents a new and exciting avenue for advancing micro-expression recognition.

2.2. Contrastive learning

Contrastive learning shapes the embedding space by attracting positive pairs and repelling negative pairs, effectively realized through contrastive loss. This method has shown great promise in visual representation learning without annotations [4, 38, 39, 47, 54, 62]. More recently, contrastive-based methods have been proposed to learn invariant properties between various data augmentation or modality views of the same image. These methods achieve notable performance in downstream tasks, some even outperforming the supervised component [6, 5, 8, 11, 40, 58].

Methods that learn prominent representation by mining invariant information from the views of different modalities are also proposed. Given an image, CMC [47] extracts feature representations of the corresponding luminance, chromaminance, depth, and other views separately, and compels the model learn the view-invariant features.In this way, each sample not only has one positive pair but gets more positive pairs from multiple views and pulls them closer in the embedding space, i.e., maximizes mutually complementary information from multi-views. Similarly, CoCLR [13] intends to mine positive samples from the same semantic-class, while proposing a co-training scheme for self-supervised pre-training. CoCLR, in particular, utilizes two complementary pieces of information (RGB and Optical flow) in videos to obtain more positive sample pairs from the other view (modality) through one, performing exceptionally well in downstream tasks.

Contrastive learning has proven effective for both small sample sizes and low-intensity tasks across various domains. For small datasets, approaches like Weakly Contrastive Learning (BIDFC) [63] and ScatSimCLR [19] demonstrate significant improvements in tasks such as automatic target recognition and small-scale datasets, leveraging batch instance discrimination and feature clustering to maintain high accuracy with minimal data.For low-intensity tasks, Semantically Contrastive Learning [34] has enhanced low-light image restoration by using contrastive pairs of normal and over/underexposed images, improving brightness consistency and image quality. Additionally, SISR frameworks [53] use contrastive learning with high-pass filters and blur operations to optimize low-resolution images, further demonstrating the technique's versatility in handling subtle features and improving resolution. These methods exemplify contrastive learning's adaptability in resource-constrained settings.

In recent years, several approaches to micro-expression recognition using contrastive learning are published. Li et al. amplify the differences

between onset and apex frames and explore differences between different AUs to enhance the robustness of AU detection [33]. Jia et al. propose a bimodal contrastive learning network for micro-expression recognition, extracting common and distinctive features from RGB and optical flow sequences [18]. Lao et al. use a graph contrastive learning (GCL) framework with a transformer-based micro-expression feature encoder to differentiate between normal and abnormal micro-expression samples [22]. Wang et al. use adaptively temporal augmented momentum contrastive learning to address data scarcity and subtle changes in micro-expression recognition [51]. Zhi et al. propose a MER-Supcon framework that combines a supervised contrastive learning approach with a dual-terminal micro-expression acquisition strategy to improve the accuracy of micro-expression classification [66]. Xia et al. introduce a macro-to-micro transformation framework that aligns spatial and temporal features from both micro and macro-expression data using domain discriminators, relation classifiers, and contrastive loss [55]. Song et al. fuse RGB features, flow features, and text information from the Facial Action Coding System to improve micro-expression recognition in scenarios where micro-expression samples are limited [43]. Zhu et al. uses contrastive learning to enhance the discriminative power of holistic representations of micro-expressions [69]. However, the aforementioned methods overlook the extraction of facial action features from large amounts of unlabeled data in an unsupervised manner. They also do not utilize features other than RGB, such as scene depth information, or those derived from RGB for feature learning and extraction.

To address these limitations, we propose a solution leveraging dual-view contrastive learning across modalities. By capturing and comparing features from different modalities like depth information, RGB, and grayscale images, our model can extract meaningful facial movement and emotion features from a large amount of unlabeled data, achieving robust micro-expression recognition.

3. Perception of Facial Action Unit Intensity based on Human Visual Observation

In our research, we use a large number of unlabeled facial video samples and select facial actions with prominent motion intensity as the pretext task input for the self-supervised learning model. We conducted a psychological experiment based on human visual perception to determine physiologi-

cally meaningful intensity screening thresholds. This research also provides valuable prior information for intelligent analysis of expressions and micro-expressions.

3.1. Method

3.1.1. Participants

The experiment involved 30 volunteers, including undergraduate and graduate students, with a mean age of 24 years (SD = 0.77). They all had a normal or corrected-to-normal vision and no known psychiatric disorders. Each participant signed an informed consent form before the experiment began and received compensation upon completion of the experiment. The study adhered to the Declaration of Helsinki and was approved by the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences.

3.1.2. Materials and Procedure

The stimuli were selected from the facial expression pictures in the CAS(ME)^3 database. Since AUs are key components of facial movements [46], we use AUs as the focus of this experiment. We evaluated AU intensity using OpenFace and selected 11 pictures per intensity (0-5), totaling 66 stimulus pictures. Since facial expressions exhibit universal characteristics [9], and we aim to establish an intensity perception standard that is applicable to a wide range of individuals, there is no restriction that the materials all come from the same subject. Furthermore, to eliminate the influence of facial familiarity on intensity evaluation, we ensured that the facial samples for each intensity were non-repeating. Additionally, we minimized the repetition of facial samples throughout the entire experiment.

During the experiment, participants viewed facial expression pictures of varying intensities and answered related questions. To help participants intuitively perceive the intensity changes in facial expressions, examples of intensity levels represented by the stimuli were shown before the experiment began, as illustrated in Fig. 2.

Before the formal experiment, participants completed six practice stimuli to familiarize themselves with the experimental process. The stimuli used in the practice session did not overlap with those in the formal experiment, and the data from the practice session were excluded from the final analysis. In the experiment, participants were presented with a facial expression picture upon pressing the space bar. They were then required to quickly and

accurately answer two questions to assess their perception of the intensity of AUs. A detailed experimental procedure is presented in Fig. 3.

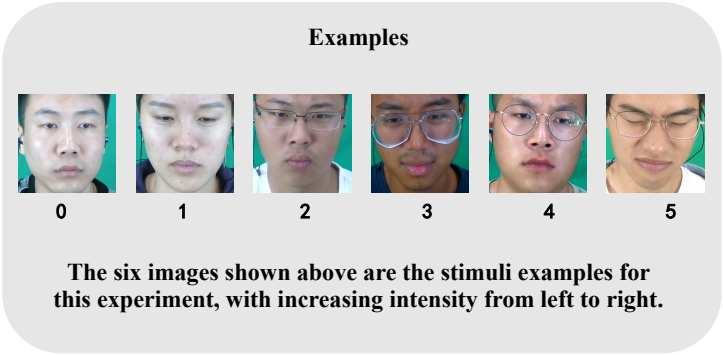


Figure 2: Instruction interface presented to participants during our behavioral experiment. In addition to the examples of AU movement intensity shown at the top, the text at the bottom was also displayed to the participants to help them better understand the experimental task.

3.2. Results and Discussion

In the data analysis process, we first excluded trials with contradictory responses to the two questions. Specifically, we removed trials where no facial movement was perceived (rating of 0) but the intensity rating was not 0, and trials where a facial movement was perceived (rating of 1) but the intensity rating was 0. This step ensured that the data reflected deliberate and conscious ratings from participants rather than random responses.

Next, we used an intensity rating of 2 as a cutoff for further analysis. We conducted a Chi-Square Test on the perception ratings and an Independent-Samples T-Test on the intensity scores separately.

We found a significant difference in whether participants perceived facial movements for AU intensities below 2 versus those at or above 2 ($x^2 = 6.748$, $p = 0.009^{**}<0.01$). Participants were more likely to perceive facial movements with AU intensity ratings of 2 or higher. Additionally, the intensity score of facial movements perceived by participants for AU intensities below 2 (M = 1.30, SD = 1.24) was significantly lower than those at or above 2 (M = 1.75, SD = 1.50), t = -6.853, $p = 0.000^{**}<0.01$. This suggests that participants not only perceive facial movements with an AU intensity rating of 2 or higher

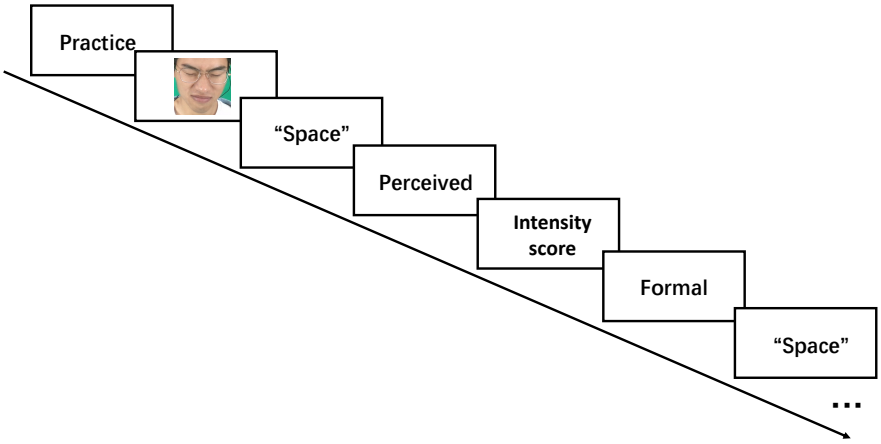


Figure 3: Experiment procedure. At the beginning of the experiment, participants were shown a facial expression picture. Once they had observed it carefully, they pressed the “space key” to start answering two questions. Participants were required to respond based on their perception, including perceived and intensity score: (1) Do you perceive any facial movements? (2) What is the intensity of the facial movements? 0-5 indicating from "no intensity" to "very strong intensity." The practice experiment consisted of 6 trials, and the formal experiment consisted of 60 trials.

more easily but also rate their intensity higher compared to movements with an AU intensity rating below 2.

The expert knowledge about AU intensities that can be perceived by the human eye is crucial for constructing the self-supervised AU intensity learning network, and it offers a theoretical foundation for establishing the AU intensity thresholds used in our following proposed computer vision-based method.

4. Proposed Method: DVSCL

Inspired by the human perception of facial action intensity, we designed a micro-expression recognition framework based on multimodal self-supervised contrastive learning, named DVSCL. This framework uses a large amount of unlabeled facial expression data to learn patterns of facial AU intensity changes. Then, it employs a small amount of labeled data to train the model

for the micro-expression recognition task, thus overcoming the limitation of small sample sizes. By progressively learning the patterns of facial action changes and micro-expression features, the framework enhances recognition performance.

4.1. Data Preprocessing

A large dataset is essential for self-supervised models to effectively learn features during the pretext task, providing robustness and versatility for downstream tasks. The recently released multimodal CAS(ME)^3 database, with over 8 million frames of facial images, was used as our data source. This dataset includes not only RGB but also depth information, capturing a wide range of facial expressions while subjects view emotionally evocative materials. However, not all frames contain facial expressions, and the distribution is sparse. To focus on facial movements, we preprocessed the data by selecting onset-apex pairs from expression segments. For Part A (annotated samples), onset-apex pairs were directly matched, while for Part B (unannotated samples), AU intensity was used to create pairs. In total, we constructed 118,542 sample pairs. Additionally, since both Part A and Part B were collected under identical conditions, training on this combined dataset facilitates better transfer learning for micro-expression recognition, minimizing the impact of environmental factors.

4.1.1. Construct Candidate Set from Unlabeled Dataset

Building a pretraining model for self-supervised learning solely based on annotated micro-expression samples is inadequate. Within Part B of the CAS(ME)^3 database, we have 2808 unlabeled videos, each documenting various facial expressions from 216 distinct subjects. In light of this, we preprocessed the data in Part B to obtain samples better suited for the self-supervised network to learn facial action features.

Specifically, to mitigate the intrusion of ID-related information and to steer the network's attention towards the dynamics of facial expressions, we formulated the onset and apex of each action sequence as the sample pair. This construction process is divided into two stages: selecting apex frames with maximum action intensity and inferring corresponding onset frames.

We used OpenFace [1] for coarse-grained screening of the AU activation status in each video sample in Part B. Given a facial expression image x_i in set $\mathbf{D} = \{x_i\}_{i=1}^{N}$, we compute the probability $\hat{p}_i = \{\hat{p}_{ij}\}_{j=1}^{M}$ of each AU

occurring, where N denotes the number of frames in one video, and M indicates the number of AUs. We then convert the probability of each AU being anticipated to the activated intensity of AUs, i.e., the regression scores of activated AUs. The activated intensity is defined as follows:

$$\begin{aligned} \hat{\mathbf{p}}_{\mathbf{i}}^{(\mathbf{int})} &= \left\{ \hat{p}_{i1}^{(int)}, \hat{p}_{i2}^{(int)}, ..., \hat{p}_{ij}^{(int)} \right\}_{j=1}^{M} \\ \text{where} \quad \hat{p}_{ij}^{(int)} &= \hat{p}_{ij} \cdot T \end{aligned} \tag{1}$$

where $T \in [0, 5]$ denotes the activated intensity of AUs. A frame x_i is selected as a candidate frame if any element in the $\hat{\mathbf{p}}^{(\mathbf{int})}$ exceeds the threshold value ϕ. (default setting: $\phi = 2.1$).

$$\mathbf{C}_{\mathbf{apex}} = \left\{ x_i \mid \exists \hat{p}_{ij}^{(int)} \in \hat{\mathbf{p}}_{\mathbf{i}}^{(\mathbf{int})}, \hat{p}_{ij}^{(int)} > \phi \right\} \tag{2}$$

Based on the human visual perception experiment in Section 3, when the AU activation scores threshold exceeds 2, the motion features of facial images are noticeable to individual vision, we hypothesize these motion features should also be evident to a self-supervised learning model. Furthermore, as shown in Fig. 4, we performed an analysis on the number of samples retained under various AU intensity threshold conditions. At a threshold (ϕ) of 2.1, the number of samples is relatively abundant, contributing to the development of a robust model. This threshold represents a trade-off between the number of samples and the level of conspicuousness of the features. A larger ϕ would result in a significant drop in the number of samples. Conversely, if ϕ were to be lowered, the motion features of the face would not be as pronounced, thereby posing challenges for the model to learn effective action features. Specifically, in order to obtain as many unsupervised motion sample pairs as possible, when there are consecutive frames with apex intensity greater than ϕ, we construct an onset-apex pair for each frame.

According to the conventionally accepted definition of micro-expressions, their duration is typically less than 500 milliseconds [27]. Additionally, frames with the highest magnitude of action, i.e., apex frames are mostly distributed within the middle of the micro-expression segment. Therefore, upon identifying frames with higher AU activation scores (apex), based on the frame rate of CAS(ME)^3 (30fps), the eighth frame before the apex frame (~250ms) is considered as the corresponding onset frame.

$$\mathbf{C}_{\mathbf{onset}} = \{ o_i = x_i - 8 \mid x_i \in \mathbf{C}_{\mathbf{apex}} \} \tag{3}$$

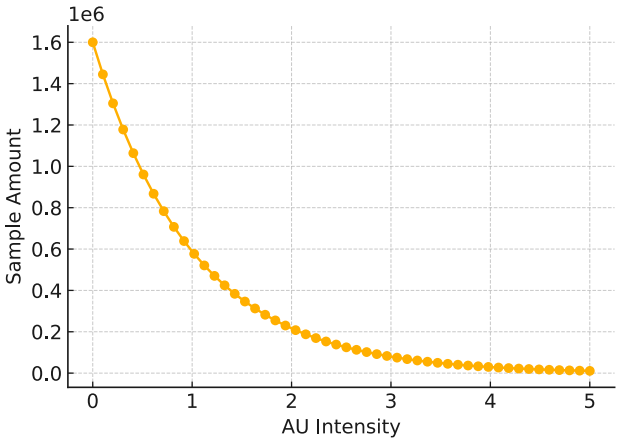


Figure 4: Comparison of the number of candidate images with different AU thresholds in CAS(ME)^3-Part B.

Similarly, for modalities derived from sources other than RGB, such as the depth modality, we retain the corresponding candidate sets in the depth modality sequence based on the same temporal location.

$$\mathcal{U}_m = \{[o_i, x_i] \mid o_i \in C_{onset}, x_i \in C_{apex}\} \tag{4}$$

where $\mathcal{U}$ denotes the candidate sets, including the AU-selected apex frames and the corresponding onset frames; m represents different modality, such as RGB, grayscale and depth information.

4.1.2. Facial Region Identification and Cropping

Starting from this section, we denote the set of apex and onset frames from labeled micro-expression segments in CAS(ME)^3-Part A as $\mathcal{M}$. To maximize sample size, we combined $\mathcal{M}$ with $\mathcal{U}$, the candidate set selected from unlabeled videos (as described in Section 4.1.1). Face cropping is performed on image pairs to eliminate background interference in micro-expression recognition.

The decision to crop the face region after selecting the candidate pairs, rather than before, is based on the need for an effective self-supervised model. Cropping early could result in distorted features due to head movement if the initial frame is used for reference. A sliding window approach would be computationally expensive given the large video volume.

Specifically, using facial landmarks detected in the onset frame as a reference, the apex frame is cropped at the same pixel positions [67]. The brow

region is adjusted based on the average distance between the left and right below the brows, limiting expansion to avoid distractions such as hair. The cropped images are resized to 310 × 310 pixels, and the same cropping is applied across all modalities.

4.1.3. Frame Difference Estimation

In this subsection, we introduce the frame difference calculation, which highlights the regions where significant facial movements occur, and provides a representation of the temporal changes in MEs. In particular, we choose to directly use frame differences instead of the prevailing optical flow-based methods [36, 35], based on the following reasons.

- Intuitive Depth Variation: The depth information is directly assigned to the value of each pixel, allowing variations in depth to be intuitively reflected by the difference between two depth images. For facial expressions, facial movements can be straightforwardly correlated with their resulting changes in depth values. Additionally, since we use multimodal data, calculating frame differences is suitable for the depth modality as well as other modalities to reflect changes caused by facial movements. This makes it the most generalized approach in preprocessing to extract variations in the facial region across different modalities.

- Reduced Computational Burden: The computation required for effective optical flow estimation is often computationally intensive, limiting its real-time applicability. By directly calculating frame differences, we aim to reduce the computational load while still capturing the dynamic facial motion characteristics in micro-expressions.

Specifically, the frame pair set in RGB view consist of onset frame $F_i^O \in \{F_1^O, F_2^O, ..., F_K^O\}$ and the apex frame $F_i^A \in \{F_1^A, F_2^A, ..., F_K^A\}$, where i and K denote the index and the total amount of frame pairs, respectively. As shown in 5, through channel-wise subtraction between the onset frame and the apex frame per pixel, we obtain the frame difference in RGB view.

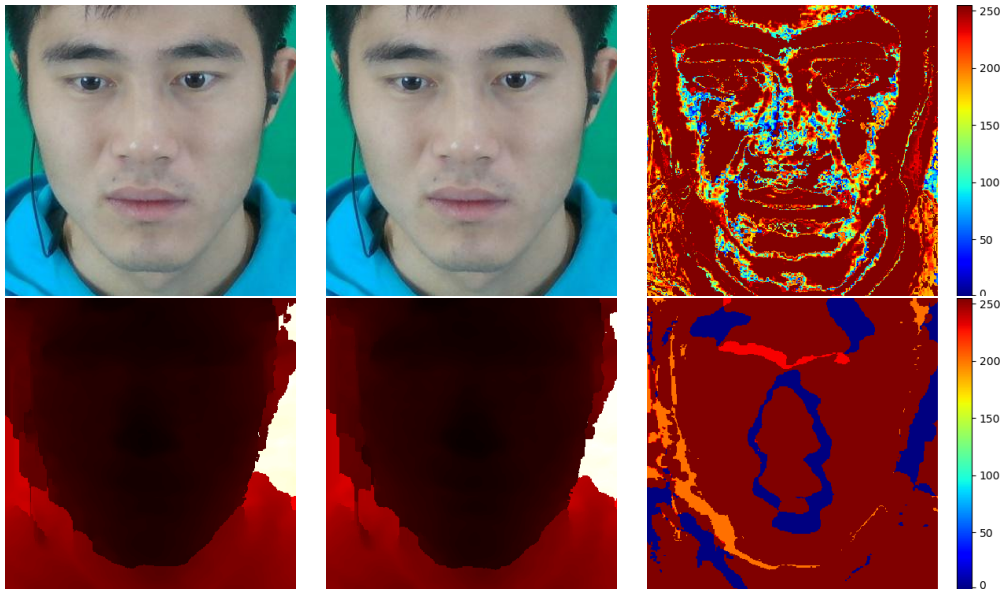
$$d_i^F = F_i^A - F_i^O \tag{5}$$

Subsequently, this approach was extended to other modalities such as depth and grayscale. For the onset frame and apex frame in each modality,

denoted by $E_i^O \in \{E_1^O, E_2^O, ..., E_k^O\}$ and $E_i^A \in \{E_1^A, E_2^A, ..., E_k^A\}$, respectively, we calculate the frame difference as:

$$d_i^E = E_i^A - E_i^O \tag{6}$$

This step enables us to capture the dynamic facial information present in different modalities, contributing to a comprehensive understanding of MEs across various visual cues. For instance, Fig. 5 present the results of our frame difference calculation for RGB view and view of the depth information, the two types of modality differences can corroborate each other and reflect changes in the facial region. Finally, d_i^E and d_i^F comprise the final training set.



(**a**) Onset (**b**) Apex (**c**) Frame diff.

Figure 5: Frame difference sample of spNO.140 in CAS(ME)^3. The facial variations in RGB differences closely correspond to disparities in depth information, with depth data offering more refined motion cues.

4.2. Dual-view Contrastive Learning

Our proposed DVSCL framework leverages invariant properties from multiple modalities for self-supervised contrastive learning. It seeks positive samples from various modalities and maximizes their mutual information via

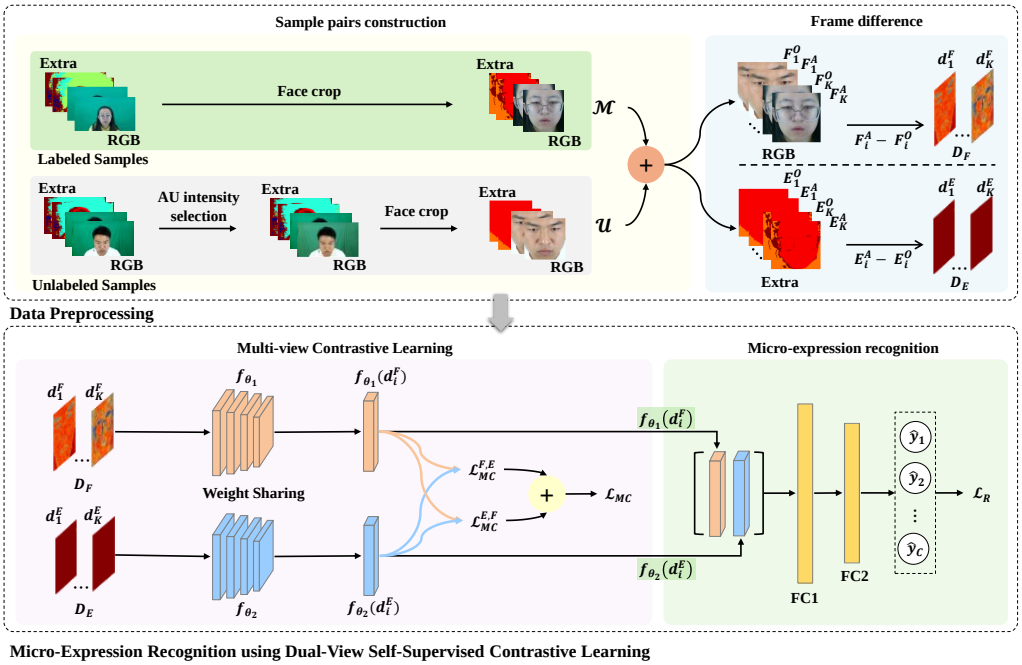


Figure 6: Training pipeline of the proposed DVSC framework. The data comes from facial samples, including RGB and depth modalities. The upper block of the framework is dedicated to data preprocessing, which includes sample selection based on AU intensity and frame differencing to highlight movements. The lower block focuses on dual-view unsupervised contrastive learning and downstream tasks, specifically micro-expression recognition.

contrastive loss. DVSCL consists of two parts: self-supervised contrastive pre-training and downstream micro-expression recognition.

During pre-training, substantial unlabeled data mines mutual information between modalities. We implemented this using a dual-view contrastive loss, learning a robust representation for the downstream task. The model was fine-tuned with annotated micro-expression data, replacing the contrastive module with a linear classification layer optimized with cross-entropy loss. Following we present the specifics of our proposed method. The overview of the framework is illustrated as Fig. 6.

4.2.1. Input

The underlying idea behind our proposed DVSCL is to learn an embedding that discriminates between samples from two different views. Due to the subtle, brief, and localized nature of micro-expressions, using traditional

data augmentation methods to construct positive pairs might actually interfere with the model's learning process. However, through contrastive learning between multiple modalities, the model can learn features from multiple perspectives, compensating for the low performance that may result from focusing on a single modality. Therefore, our proposed DVSCL aims to unsupervisedly learn the facial action features embedded across different modalities.

Regarding the feature dimensionality, DVSCL is flexible and supports switching between different feature dimensions. For the RGB modality, we encode the input using a three-channel network, while for single-channel modalities, such as depth or grayscale, we use a single-channel network for encoding. The contrastive learning is then performed at the feature level.

Specifically, as mentioned earlier, in the CAS(ME)^3 dataset, the information in the micro-expression fragments is represented by both the RGB color map and the depth map, and there is interrelated mutual information between the two different modalities. Hence, we primarily utilize the RGB and depth information from CAS(ME)^3 to construct the network and perform performance verification, i.e., the construction of positive pairs in the following context is mainly based on these two modalities. Not only that, our network is applicable to any two two-dimensional features, not just limited to RGB and depth. For example, we also analyzed the network performance based on other modalities like grayscale in the experiments.

4.2.2. Encoder Network

Two ResNet18 backbones with weight sharing are employed as the main network architecture to process RGB and depth image samples in parallel. Specifically, we input RGB and depth images into two parallel encoders, denoted as $f_{\theta_1}(\cdot)$ and $f_{\theta_2}(\cdot)$, respectively, which share the same convolutional layers and weight parameters. This weight-sharing approach significantly reduces the model's parameter count, making it easier to train and optimize. Additionally, because both branches share the same feature extractor, this method also enhances the model's generalization ability, enabling it to better handle new multimodal input data.

4.2.3. Dual-view Contrastive Loss

Fig. 1 illustrates this self-supervised contrastive learning process using RGB and depth modalities. Specifically, D_F and D_E represent the data sets for the RGB and depth modalities. Each set contains N samples, represented

by d_i^F and d_i^E respectively, where $i \in \{1, 2, \ldots, N\}$. Specifically:

$$D_F = \{d_i^F\}_{i=1}^N,\ D_E = \{d_i^E\}_{i=1}^N$$

Representations of a single sample across modalities, i.e., $\{d_i^F, d_i^E\}$, are termed *positive pairs*, while representations of different samples, i.e., $\{d_i^F, d_j^E\}$, are *negative pairs*. Formally:

1. **Positive pairs:**
$$X_{\text{positive}} = \{(d_i^F, d_i^E)\}_{i=1}^N$$
2. **Negative pairs:**
$$X_{\text{negative}} = \{(d_i^F, d_j^E)\}_{i,j=1,i\neq j}^N$$

Specifically, we first fix any sample, such as d_1^F in one view, enumerate all samples in another view, i.e., $d_1^E, d_2^E, ...d_k^E$, and calculate their similarity. The objective function $\mathcal{L}_i^{D_F,D_E}$ is defined to be:

$$\mathcal{L}_i^{D_F,D_E} = \\ -\log\frac{\exp(\text{sim}(d_i^F, d_i^E))/\tau}{\exp(\text{sim}(d_i^F, d_i^E)) + \sum_{j=1}^K \exp(\text{sim}(d_i^F, d_j^E))/\tau} \tag{7}$$

where τ denotes the temperature coefficient, $sim(x, y)$ is the similarity calculation function, which is implemented by the cosine similarity after $L2$ normalization.

In particular, as mentioned in previous section, to extract the potential feature representations of d^F and d^E, we feed them into two weight-sharing encoders, $f_{\theta_1}(\cdot)$ and $f_{\theta_2}(\cdot)$, respectively. Then we obtain the feature representations $f_{\theta_1}(d^F)$ and $f_{\theta_2}(d^E)$ and perform similarity calculation. The similarity function $sim(d^F, d^E)$ is shown as follow:

$$sim(d^F, d^E) = \frac{f_{\theta_1}(d^F) \cdot f_{\theta_2}(d^E)}{\|f_{\theta_1}(d^F)\| \cdot \|f_{\theta_2}(d^E)\|} \tag{8}$$

Following the above, we maximize the similarity between the anchor and the positive pair while minimizing the similarity to the negative pairs.

Subsequently, we calculate the mean loss $\mathcal{L}_{MC}^{F,E}$ across all samples as follows:

$$\mathcal{L}_{MC}^{F,E} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_i^{D_F,D_E} \tag{9}$$

Since $\mathcal{L}_{MC}^{F,E}$ is the comparison loss calculated by fixing D_F traversing D_E. Symmetrically, calculating $\mathcal{L}_{MC}^{E,F}$, and the final loss is:

$$\mathcal{L}^{MC} = \mathcal{L}_{MC}^{F,E} + \mathcal{L}_{MC}^{E,F} \tag{10}$$

By employing Dual-view Contrastive Loss for self-supervised contrastive learning, we can effectively acquire feature representations of micro-expressions from two different perspectives: RGB color images and depth maps. This multimodal self-supervised learning approach leverages the mutual information between the two modalities, enhancing the model's generalization capability and feature expressiveness, thereby achieving better performance in micro-expression recognition tasks.

4.3. Downstream Task: Micro-Expression Recognition

After the self-supervised contrastive pre-training, the representations are further transferred to a downstream micro-expression recognition task to validate the efficacy of the learned representations.

Specifically, in the downstream task, contrastive learning, utilized to discriminate between samples during pre-training, is no longer required. Instead, the network, which has learned facial action features ($f_{\theta_1}(d^F)$, $f_{\theta_2}(d^E)$), is employed to perform a micro-expression classification task. Consequently, the contrastive module, pivotal during the pretext task, is substituted with a linear classification layer for this phase. To optimize the classification, a cross-entropy (CE) loss function, defined as:

$$\mathcal{L}^R = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{11}$$

where C represents the number of classes, is utilized. Throughout this stage, to exclusively train the newly added linear layer, the remainder of the model is kept frozen.

5. Experiments

In this section, we will first describe our experimental details, which include datasets and experimental settings for both pre-training and downstream evaluation. Next, comprehensive experimental results about DVSCL are reported. We also conduct multiple ablation studies to demonstrate the efficacy of each module in our framework. Finally, we provide the analysis regarding micro-expression recognition.

5.1. Dataset

Extensive experiments are conducted on CAS(ME)^3 [27] dataset and MEGC2019-CD [42], which contains three spontaneous datasets namely SMIC [30], CASME II [59] and SAMM [7].

5.1.1. CAS(ME)^3

CAS(ME)^3 [27] is an multi-model dataset. Among all the currently available datasets, it is the only micro-expression database that captures RGB and depth information simultaneously. Table 1 shows an overview of this database. The pre-training dataset consists of both micro- and macro-expression in Part A and unlabeled images after AU selection in Part B. Downstream micro-expression recognition task is implemented on the "labeled micro-expression Data" portion of CAS(ME)^3 dataset, which contains 860 micro-expression samples, including RGB and Depth pairs of difference between apex frame and onset frame.

Table 1: The overview of CAS(ME)^3, based on the second version of annotation. ME and MaE represent micro- and macro-expression, respectively. The numbers in the table represent the quantities under the respective headers.

Part	ME	MaE	Subject	Emotion class	Modality
A	860	3226	100	7	RGB+Depth
B	N/A	N/A	116	N/A	

5.1.2. MEGC2019-CD

The 3DB-combined database is proposed for Composite Database Evaluation (CDE) by MEGC2019, called **MEGC2019-CD** [42]. It consists of a combination of three spontaneous datasets, namely SMIC, CASME II, and SAMM, with three emotion categories, namely Negative (including Repression, Anger, Contempt, Disgust, Fear, and Sadness), Positive (i.e., 'Happiness') and Surprise. Table 2 summarizes the sample distribution for the three datasets and their combination MEGC2019-CD. The detailed information of these three datasets is described as follows:

SMIC [30]. The high-speed camera (HS) verson of the Spontaneous Micro-Expression Corpus (SMIC) is utilized, thus unifying the evaluation with the two CASME II and SAMM described next. SMIC-HS contains 164 video clips from 16 subjects and was recorded using a 100 fps high-speed camera with a resolution of 640×480.

Table 2: Sample amount of the Micro-expression Composite Database (MEGC2019-CD), which is comprised of SMIC, CASME II, and SAMM.

Datasets	**Emotion class**			**Total**	**Modality**
	Negative	**Positive**	**Surprise**		
SMIC (SMIC-HS) [30]	70	51	43	164	RGB
CASME II [59]	88	32	25	145	
SAMM [7]	92	26	15	133	
MEGC2019-CD [42]	250	109	83	442	

CASME II [59]. The Chinese Academy of Sciences Micro-Expression II (CASME II) dataset was obtained with a 200 fps high-speed camera and contained 255 micro-expression samples from 26 participants. For each frame, the raw resolution is 640×480, and the facial region is 280×340 pixels.

SAMM [7]. The Spontaneous Actions and Micro-Movement (SAMM) dataset contains 159 micro-expression clips from 29 participants. It was captured at 200 fps using a high-speed camera and coded with onset, apex, and offset frames. The original resolution for each micro-expression frame in SAMM is 2040×1088, and the facial area is approximately 400 × 400 pixels.

5.2. Experimental Setting

5.2.1. Setup for Self-supervised Pre-training

Parameters Setting. For self-supervised pre-training stage, we train the network with our DVSCL framework for 240 epochs, and adopt SGD as our optimizer with a momentum of 0.9 and a weight decay of 1e-4. The initial learning rate is 0.003 with a cosine learning rate schedule used. We set a batch size as 128 and contrast each positive pair with 127 negative pairs. Following [CMC], we set the temperature τ as 0.07, and the dimension of the embedding which extracted by the network is 128-d. All the experiments are conducted on 8 NVIDIA-RTX-1060Ti GPUs.

Data Augmentation. In order to make the learned features more robust to the invariance of multi views (i.e. different modalities of a micro-expression sample) and different transformation, we applied 4 kinds of data augmentation for each sample, including random resized crop, random horizontal flip, color jitter and grayscale which provided by PyTorch [CMC's 59] [47].

5.2.2. Setup for Micro-Expression Recognition Experiments

Evaluation Metrics. Leave-one-subject-out (LOSO) cross-validation is utilized in all of the evaluation experiments, where samples from one subject

are held out as the testing set while all other samples are used for training. The following metrics are used to result evaluation.

$$Acc = \frac{num(true)}{num(total)} \tag{12}$$

$$UAR = \sum_{i=1}^{K} \frac{Acc_i}{K}, \text{ where } Acc_i = \frac{TP_i}{N_i} \tag{13}$$

$$\begin{aligned} UF1 &= \sum_{i=1}^{K} \frac{UF1_i}{K}, \\ \text{where } UF1_i &= \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i} \end{aligned} \tag{14}$$

Parameters Setting. After self-supervised learning, we verify the representation which trained with DVSCL by transferring the weights of encoder network to micro-expression recognition. We perform a linear classification task with frozen weights and only a single linear layer which on the top frozen layer is trained with cross-entropy loss on CAS(ME)^3 dataset. The learning rate is initialized as 0.0001 with a cosine learning rate schedule applied.

5.3. Experimental Results with SOTA Comparisons

For the SOTA comparisons, we focused on methods with available open-source code that we could accurately reproduce. We compared our method with several influential approaches and the most recent methods published in the last five years. Micro-expressions are extremely subtle, and different preprocessing or data validation techniques can significantly impact recognition performance. Therefore, we typically reproduce SOTA methods to ensure that all algorithms are evaluated on the same data basis for a fair comparison. Specifically, to demonstrate that self-supervised contrastive learning can effectively capture motion patterns through RGB-related inputs, we standardized the input features across all methods being compared, including ours and the SOTA approaches.

As shown in Table 3, our proposed method achieves the best performance on the CAS(ME)^3 dataset. This can be attributed to the fact that CAS(ME)^3 contains the largest number of samples among the datasets, and there are

Table 3: Comparison of Our Method with SOTA Approaches on the Multimodal CAS(ME)^3 Dataset (RGB+Depth) and the MEGC2019-CD Dataset (CASME II, SAMM, SMIC) Using RGB Modalities. Note: All input features are unified to RGB-related features, with RGB difference being used unless otherwise specified. FGRL requires AU labels, and as SMIC in MEGC2019-CD lacks AU labels, no corresponding results are available for MEGC2019-CD.

Methods		CAS(ME)3		CASME II		SAMM		SMIC		MGEC2019-CD	
		UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR	UF1
RCN (2020) [57]	RCN-A	0.2602	0.2109	0.2685	0.3277	0.2636	0.3152	0.3172	0.3547	0.3465	0.3096
	RCN-S	0.2534	0.1856	0.2482	0.3258	0.2643	0.3284	0.2808	0.3226	0.3288	0.2597
	RCN-W	0.2723	0.2308	0.2962	0.3486	0.2618	0.3116	0.2718	0.2881	0.3303	0.2978
	RCN-F	0.2606	0.2031	0.3463	0.3829	0.3257	0.3559	0.3257	0.3599	0.3327	0.2931
	RCN-C	0.2811	0.2527	0.3527	0.3792	0.3237	0.3351	0.3353	0.3285	0.3396	0.2991
	RCN-P	0.2500	0.1752	0.2500	0.3295	0.2655	0.3188	0.2386	0.2762	0.3271	0.2540
FGRL (onset-apex) (2021) [24]		0.1735	0.2500	-	-	-	-	-	-	-	-
MMNet (onset-apex) (2022) [26]		0.3101	0.3054	0.4073	0.4086	0.2467	0.2651	0.2585	0.2272	0.3897	0.3837
FR (2022) [68]	FR (R,G)	0.3020	0.3034	**0.5371**	**0.5074**	0.2496	0.2862	0.3495	0.3506	**0.4869**	**0.4784**
	FR (R,B)	0.2965	0.2921	0.3387	0.3507	0.2740	0.3099	0.4076	0.4084	0.3981	0.3970
	FR (G,B)	0.3165	0.3107	0.3804	0.3744	0.2747	0.3079	0.3636	0.3919	0.4124	0.4122
HTNet (2024) [52]		0.3020	0.3020	0.4191	0.4332	0.3063	0.2743	0.3803	0.3605	0.3521	0.3638
DVSCL	RGB+Gray	0.3264	0.3139	0.4543	0.4600	**0.4034**	**0.3861**	0.3114	0.2919	0.4004	0.4039
	RGB+Depth	**0.3373**	**0.3252**	-	-	-	-	-	-	-	-

no cross-database differences. Specifically, networks trained with our self-supervised approach, DVSCL, on large-scale facial motion data demonstrate advantages over those trained on micro-expression datasets in a fully supervised manner. Additionally, the results of DVSCL with RGB+Depth outperform those with RGB+Gray, indicating that the additional modality (depth information) effectively helps the self-supervised contrastive learning network to learn more features, thereby improving micro-expression recognition performance. Furthermore, on the MEGC2019-CD dataset (CASME II, SAMM, SMIC), our method performs comparably with SOTA methods. This is particularly notable as our primary design focus was on the self-supervised contrastive learning aspect, while the micro-expression recognition component of our method remains relatively simple. This demonstrates the potential of self-supervised learning in advancing micro-expression recognition.

The relatively poor performance on the CAS(ME)^3 database is due to the complexity of the samples, which include facial occlusions and significant head movements. The suboptimal performance on the SMIC database is primarily attributed to its lower facial resolution, averaging 190×230 pixels, making it the most challenging database for the network to extract facial motion features.

5.4. Comprehensive Experimental Analysis on Model Design

To demonstrate the validity of the proposed method, we conducted an extensive experimental analysis on the CAS(ME)^3 dataset, including pre-training dataset scales, multi-modal inputs, feature extraction networks, sampling distribution, and different zero-filling methods for depth image.

5.4.1. Number of Samples for Pre-training

Table 4: Comparative study on different sample sizes and modalities for self-supervised learning (Part-A vs Part-A&B)

View	**Part A**		**Part-A&B**	
	UAR	**UF1**	**UAR**	**UF1**
RGB+Depth	0.2547	0.2243	**0.3373**	**0.3252**
Depth+Gray	0.2449	0.2435	0.2573	0.2533
R+G	0.2405	0.2392	0.2477	0.246
G+B	0.2367	0.2341	0.2465	0.2442
R+B	0.2274	0.2195	0.2306	0.2252

To investigate the impact of data volume during the self-supervised pre-training phase on final micro-expression recognition performance, we conducted ablation experiments on different-sized datasets of CAS(ME)^3. Specifically, we considered two dataset sizes: Part-A only and Part-A&B. Part-A consists of a labeled subset of the CAS(ME)^3 dataset, while Part-A&B also includes Part-B, a large amount of unlabeled emotional samples. In particular, the sample size of part A is limited to 860 samples. In contrast, Part-A&B combined has 119,420 samples, which is 138 times larger. As listed in Table Table 4, the results clearly indicate that, compared to modality switching, the sample size has a much more significant impact on recognition performance. Hence, by using more training data, DVSCL robustly learned discriminative emotional features and improved the model's performance during fine-tuning.

5.4.2. Effectiveness of Different Inputs

1) Compare Depth and Gray modalities: RGB+Depth vs RGB+Gray. To investigate the effect of data modality on model performance, we used two different modality combinations for self-supervised contrastive learning: RGB+Depth and RGB+Gray. We used ResNet18 as the feature extraction network and trained the model in both supervised and self-supervised

manner. As shown in Table 3, the RGB+depth combination achieved higher micro-expression recognition performance in both supervised and self-supervised settings. This is because depth information and RGB information are heterogeneous, allowing the network to learn different facial action patterns from these two distinct modalities. Specifically, the Depth+RGB combination can leverage the lighting insensitivity of depth information and the rich detail-capturing ability of RGB information, providing comprehensive and rich facial feature information, enhancing the model's performance under various conditions. In contrast, grayscale and RGB essentially represent the same perspective. Although increasing the dimensions improves performance, they lack the complementary or new information provided by different modalities.

2) Compare single-channel inputs: To investigate the effect of single-channel combinations on model performance, we compared the combinations of grayscale and depth channels with single channels in RGB. According to Table 4, the Gray+Depth combination achieved the best performance. This result also supports the previous conclusion. Depth information and grayscale information come from different modalities. Depth information captures the 3D structure and distance of the face, while grayscale information primarily reflects texture and brightness variations. Combining these two types of information provides a more comprehensive description of facial features, improving model performance. However, combinations of RGB channels (such as R+G/R+B/G+B) essentially derive from the same perspective (RGB images), differing only in color information. Therefore, these combinations contain a lot of redundant information and do not provide new features.

5.4.3. Effectiveness of Different Backbones

To investigate the effect of network depth on model performance, we conducted ablation experiments on different feature extraction networks. Specifically, we selected ResNet18, ResNet50, and AlexNet as three backbone networks and trained them with RGBD combinations. We evaluated their performance on the micro-expression dataset, as shown in Table 5. The results showed that ResNet18 achieved the best performance. The performance of the other backbone networks was slightly worse, which may be due to factors such as their depth, dataset size, and overfitting. To further analyze these results, we noted that ResNet18's network depth is between ResNet50 and AlexNet. On the one hand, it shows better feature extraction capabilities than the shallower AlexNet. On the other hand, it avoids overfitting caused

by deep networks, as occurred in ResNet50. In addition, our experimental dataset is relatively small, which may limit the performance of deeper networks. These factors may help explain why ResNet18 achieved the best performance.

Table 5: Comparative study on backbones for feature extraction on CAS(ME)^3 under RGB+Depth view

	Supervised		**Self-Supervised**	
	UAR	**UF1**	**UAR**	**UF1**
AlexNet [21]	0.2290	0.2155	0.2437	0.2437
ResNet18 [15]	0.2598	0.2229	**0.3373**	**0.3252**
ResNet50 [15]	0.2645	0.2261	0.2880	0.2858

5.4.4. Effectiveness of Sample Distribution

1) Different sampling strategies: Sample distribution has a significant impact on the performance of the model. We used three different sampling strategies, including the original dataset, oversampling, and semi-oversampling. Specifically, in oversampling, we oversampled Positive, Surprise, and Others samples by 8x, 3x, and 3x, respectively, while keeping the Negative samples unchanged. In semi-oversampling, we oversampled these three sample types by 4x, 1.5x, and 1.5x, respectively, while keeping the Negative samples unchanged. The micro-expression recognition results of the three different sampling strategies are shown in Table 6, with semi-oversampling achieving the best performance. This may be because semi-oversampling can increase the number of minority samples while avoiding introducing too much noise and overfitting issues.

2) Different loss fuction: We also compared three different loss functions, including the standard CE loss, Focal Loss, and weighted CE loss. In the weighted CE loss, we calculated the weight of each class based on the sample size of each class in the dataset. As listed in the Table 6, standard CE loss achieved the best performance. We analyzed the possible reasons as follows: First, Focal Loss and weighted CE loss introduce additional parameters or weights, which might cause the model to overfit the training data. Second, many micro-expressions belonging to the "negative" and "others" categories are difficult to classify, and adjusting their weights might negatively impact model training. In contrast, the standard CE loss is relatively simple, reducing the risk of overfitting and thus improving generalization ability.

Table 6: Comparative Study on different sample distributions and loss functions on CAS(ME)^3 under RGB+Depth view

Sampling	Loss	UAR	UF1
Original	CE	0.2880	0.2858
	weight-CE	0.2861	0.2536
	Focal	0.2500	0.2066
Oversampling	CE	0.2539	0.2538
Half-Oversampling		**0.3373**	**0.3252**

5.4.5. Effectiveness of Different Zero-fill Methods for Depth Modality

Due to the limitations of depth cameras, the collected depth maps inevitably contain zero values. To avoid the interference of these zero values in action estimation, we employed different zero-filling methods to complete the pixels where the camera failed to capture depth information. This operation aims to enhance the performance of micro-expression recognition based on depth modality.

Specifically, as listed in following formulas, we employed three methods for zero-filling: minimum value, mean value, and neighborhood value, respectively.

$$D'(x,y) = \begin{cases} D(x,y) & \text{if } D(x,y) \neq 0 \\ \min(D) & \text{if } D(x,y) = 0 \end{cases} \tag{15}$$

$$D'(x,y) = \begin{cases} D(x,y) & \text{if } D(x,y) \neq 0 \\ \frac{1}{|\mathcal{N}(x,y)|} \sum_{(i,j)\in\mathcal{N}(x,y)} D(i,j) & \text{if } D(x,y) = 0 \end{cases} \tag{16}$$

$$D'(x,y) = \begin{cases} D(x,y) & \text{if } D(x,y) \neq 0 \\ \frac{1}{N} \sum_{(i,j)\in D \text{ where } D(i,j)\neq 0} D(i,j) & \text{if } D(x,y) = 0 \end{cases} \tag{17}$$

where $D(x,y)$ represents the original depth value at position (x,y), $D'(x,y)$ represents the depth value after zero-filling, $\mathcal{N}(x,y)$ denotes the neighborhood of the pixel (x,y), $|\mathcal{N}(x,y)|$ is the number of non-zero pixels in the neighborhood, $\min(D)$ is the minimum value in the depth map, and N is the number of non-zero values in the depth map.

As listed in Table 7, the supervised learning (Sup.) method achieved the highest ACC (0.5460) using the RGB+Depth (min) zero-filling approach. However, the corresponding UAR and UF1 were relatively low, indicating

Table 7: Comparative study on different zero-fill methods for depth modality on on CAS(ME)^3 under RGB+Depth view

Zero-fill method	**Sup.**		**DVSCL**	
	UAR	**UF1**	**UAR**	**UF1**
Depth (min)	0.2605	0.2027	0.2941	0.2947
Depth (mean)	0.2625	0.2172	**0.3187**	**0.3051**
Depth (nearest)	0.2608	0.2229	0.3373	0.3252

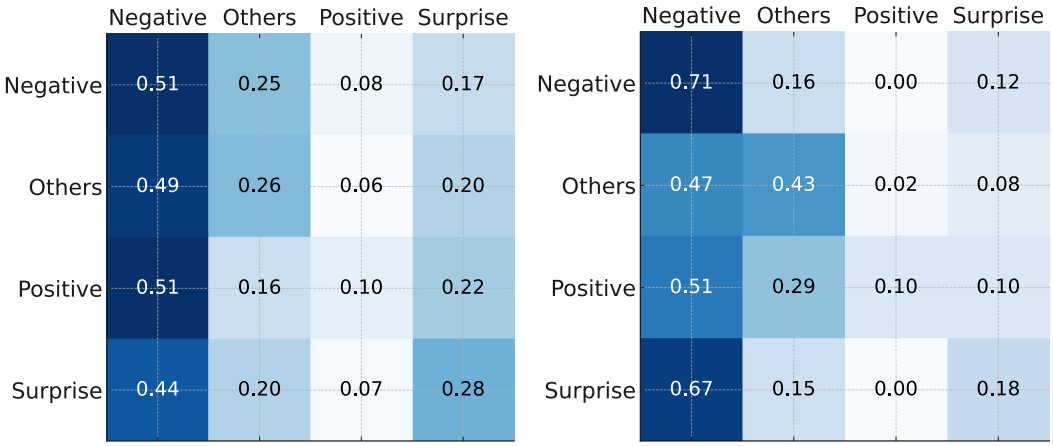
* In the supervised learning method (Sup.), after feature extraction with ResNet18, no contrastive learning was performed, and the micro-expression recognition task was carried out directly.

that while the overall accuracy was high, the balanced performance across different classes was poor. Our proposed DVSCL metho generally has higher UAR and UF1 compared to the supervised learning methods. This suggests that unsupervised learning is better at capturing balanced performance across categories, particularly when dealing with imbalanced datasets or larger sample sizes. Specifically, using the RGB+Depth (nearest) zero-filling approach, both UAR and UF1 reached their highest values.

5.5. Qualitative and Quantitative Analysis

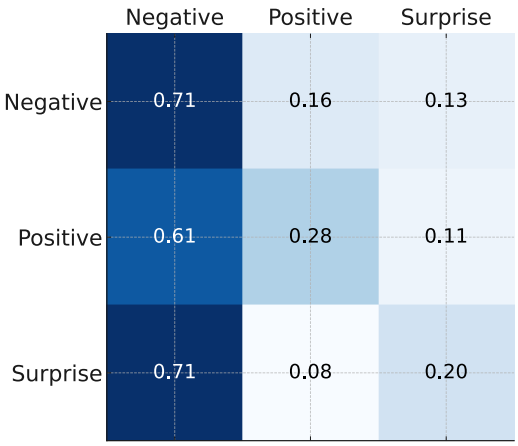
As illustrated in Fig. 7, the confusion matrix reveals that categories with larger sample sizes generally have higher recognition rates. This is evident across various emotion categories in the table, such as the high recognition rates of negative emotions in CASME II and CAS(ME)^3. Additionally, in self-supervised learning methods, the similarity between the samples used in the upstream model training and the downstream samples significantly impacts recognition performance. The CASME II database has the highest similarity in terms of acquisition environment, equipment, and participant groups compared to CAS(ME)^3, which results in optimal performance in recognizing negative and surprise emotions in CASME II.

Additionally, when comparing the performance of RGBD and RGB-gray on CAS(ME)^3, we observe that positive and negative emotions tend to exhibit more significant texture changes rather than geometric deformations along the camera axis. As a result, the recognition performance for these emotions is higher using RGB-Gray compared to depth-based methods. In contrast, the emotion of surprise typically involves eye socket expansion, which causes more noticeable geometric deformation than texture changes. This leads

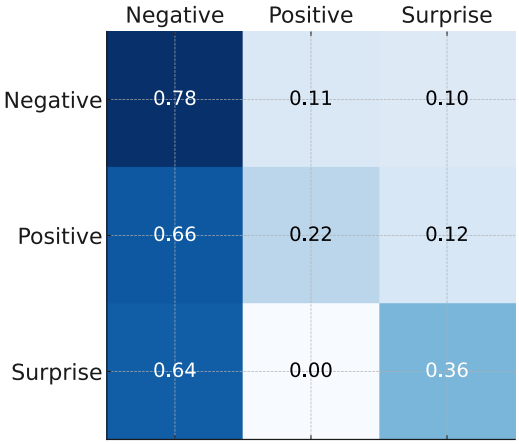


(a) RGB-D on CAS(ME)^3

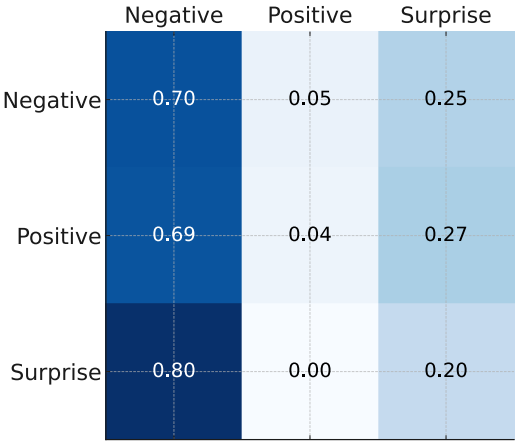
(b) RGB-Gray on CAS(ME)^3



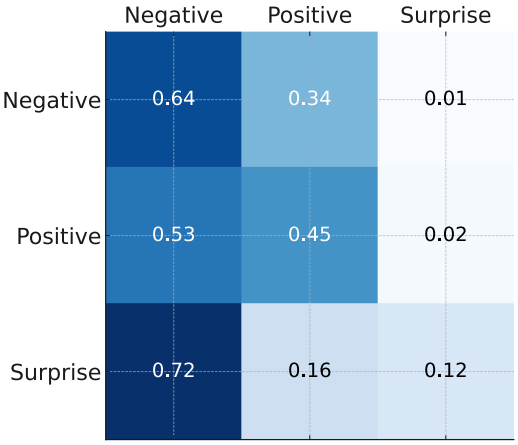
(c) RGB-Gray on MEGC2019-CD



(d) RGB-Gray on CASME II



(e) RGB-Gray on SAMM



(f) RGB-Gray on SMIC

Figure 7: The confusion matrixs on CAS(ME)^3 and MEGC2019-CD. In particular, MEGC2019-CD is composed of the CASME II, SAMM, and SMIC databases. The MEGC2019 evaluation method involves validating the model using all samples from the three databases, while also presenting the results for individual datasets.

to better performance for the RGBD method compared to the RGB-Gray method.

In summary, the characteristics and acquisition methods of different databases, as well as the impact of sample size and geometric deformation, all significantly influence micro-expression recognition performance. Designing more universally applicable unsupervised models remains a worthwhile area of exploration.

6. Conclusion and Perspectives

In the current era of rapidly advancing human-computer interaction, intelligent recognition of micro-expressions can help machines better understand individuals' true emotions. However, micro-expression recognition faces several challenges, such as the inherently subtle and brief nature of micro-expressions, which makes them difficult for models to learn, and the small sample problem due to the difficulties in eliciting and annotating these expressions. To address the small sample problem in micro-expression recognition, we propose leveraging facial action patterns from a large number of unsupervised samples to build models with generalized learning capabilities for facial actions. Additionally, the selection of these action features has been validated through psychological experiments involving human perception. Furthermore, when sample sizes are limited, learning the geometric transformations and texture changes brought by micro-expression actions through additional modalities can enhance recognition performance. Experiments have demonstrated that self-supervised learning based on a large number of unlabeled samples can better capture balanced performance across different categories, especially when dealing with imbalanced datasets or larger sample sizes.

In future work, designing upstream tasks in unsupervised networks to balance the learning of various spatiotemporal facial features caused by facial actions, such as geometric deformations and texture changes, will be crucial for further enhancing the learning performance of micro-expression features. Additionally, leveraging the emerging capabilities of large models in visual feature learning and inference represents another promising direction for continued exploration.

Acknowledgment

This work is supported, in part, by grants from the National Natural Science Foundation of China (62276252, 62106256, 62476269), and in part, by a grant from the Youth Innovation Promotion Association CAS.

References

[1] Baltrušaitis, T., Robinson, P., Morency, L.P., 2016. OpenFace: An open source facial behavior analysis toolkit, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. doi:10.1109/WACV.2016.7477553.

[2] Ben, X., Ren, Y., Zhang, J., Wang, S.J., Kpalma, K., Meng, W., Liu, Y.J., 2022. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 5826–5846. doi:10.1109/TPAMI.2021.3067464.

[3] Cai, W., Zhao, J., Yi, R., Yu, M., Duan, F., Pan, Z., Liu, Y.J., 2024. Mfdan: Multi-level flow-driven attention network for micro-expression recognition. IEEE Transactions on Circuits and Systems for Video Technology , 1–1doi:10.1109/TCSVT.2024.3437481.

[4] Caron, M., Bojanowski, P., Mairal, J., Joulin, A., 2019. Unsupervised pre-training of image features on non-curated data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2959–2968. doi:10.1109/ICCV.2019.00305.

[5] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660. doi:10.1109/ICCV48922.2021.00951.

[6] Chen, X., Xie, S., He, K., 2021. An empirical study of training self-supervised vision transformers, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9620–9629. doi:10.1109/ICCV48922.2021.00950.

[7] Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H., 2018. SAMM: A spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing 9, 116–129. doi:10.1109/TAFFC.2016.2573832.

[8] Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A., 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9568–9577. doi:10.1109/ICCV48922.2021.00945.

[9] Ekman, P., 1992. An argument for basic emotions. Cognition and Emotion 6, 169–200. doi:10.1080/02699939208411068.

[10] Gan, Y.S., Lien, S.E., Chiang, Y.C., Liong, S.T., 2024. Laenet for micro-expression recognition. The Visual Computer 40, 585–599. doi:10.1007/s00371-023-02803-3.

[11] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems 33, 21271–21284. doi:10.5555/3495724.3497510.

[12] Gupta, P., 2023. MERASTC: Micro-expression recognition using effective feature encodings and 2d convolutional neural network. IEEE Transactions on Affective Computing 14, 1431–1441. doi:10.1109/TAFFC.2021.3061967.

[13] Han, T., Xie, W., Zisserman, A., 2020. Self-supervised co-training for video representation learning. Advances in Neural Information Processing Systems 33, 5679–5690. doi:10.5555/3495724.3496201.

[14] Happy, S.L., Routray, A., 2019. Fuzzy histogram of optical flow orientations for micro-expression recognition. IEEE Transactions on Affective Computing 10, 394–406. doi:10.1109/TAFFC.2017.2723386.

[15] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. doi:10.1109/CVPR.2016.90.

[16] Huang, X., Wang, S.J., Liu, X., Zhao, G., Feng, X., Pietikäinen, M., 2019. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Transactions on Affective Computing 10, 32–47. doi:10.1109/TAFFC.2017.2713359.

[17] Husák, P., Cech, J., Matas, J., 2017. Spotting facial micro-expressions "in the wild", in: 22nd Computer Vision Winter Workshop (Retz), pp. 1–9.

[18] Jia, W., Song, Y., Wang, P., Chen, L., Ben, X., 2022. Micro-expression recognition based on bimodal contrastive learning, in: 2022 the 5th International Conference on Image and Graphics Processing (ICIGP), pp. 139–144. doi:10.1504/IJCAT.2023.132402.

[19] Kinakh, V., Taran, O., Voloshynovskiy, S., 2021. ScatSimCLR: self-supervised contrastive learning with pretext task regularization for small-scale datasets, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1098–1106. doi:10.1109/ICCVW54120.2021.00129.

[20] Kollias, D., 2022. ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2327–2335. doi:10.1109/CVPRW56347.2022.00259.

[21] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25. doi:10.1145/3065386.

[22] Lao, L., Li, Y., Liu, M.L., Xu, C., Cui, Z., 2022. Temporal discriminative micro-expression recognition via graph contrastive learning, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE. pp. 1033–1040. doi:10.1109/ICPR56361.2022.9956075.

[23] Le Ngo, A.C., See, J., Phan, R.C.W., 2016. Sparsity in dynamics of spontaneous subtle emotions: analysis and application. IEEE Transactions on Affective Computing 8, 396–411. doi:10.1109/TAFFC.2016.2633283.

[24] Lei, L., Chen, T., Li, S., Li, J., 2021. 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, pp. 1571–1580. doi:10.1109/CVPRW53098.2021.00173.

[25] Lei, L., Li, J., Chen, T., Li, S., 2020. A novel graph-tcn with a graph structured representation for micro-expression recognition, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2237–2245. doi:10.1145/3394171.3413714.

[26] Li, H., Sui, M., Zhu, Z., Zhao, F., 2022. MMNet: Muscle motion-guided network for micro-expression recognition, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI), pp. 1074–1080. doi:10.24963/ijcai.2022/150.

[27] Li, J., Dong, Z., Lu, S., Wang, S.J., Yan, W.J., Ma, Y., Liu, Y., Huang, C., Fu, X., 2023. 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 45, 2782–2800. doi:10.1109/TPAMI.2022.3174895.

[28] Li, X., Cheng, S., Li, Y., Behzad, M., Shen, J., Zafeiriou, S., Pantic, M., Zhao, G., . 4DME: A spontaneous 4d micro-expression dataset with multimodalities. IEEE Transactions on Affective Computing 14, 3031–3047. doi:10.1109/TAFFC.2022.3182342.

[29] Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikäinen, M., 2017. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Transactions on Affective Computing 9, 563–577. doi:10.1109/TAFFC.2017.2667642.

[30] Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M., 2013. A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), IEEE. pp. 1–6. doi:10.1109/FG.2013.6553717.

[31] Li, Y., Huang, X., Zhao, G., 2021a. Joint local and global information learning with single apex frame detection for micro-expression recognition. IEEE Transactions on Image Processing 30, 249–263. doi:10.1109/TIP.2020.3035042.

[32] Li, Y., Huang, X., Zhao, G., 2021b. Micro-expression action unit detection with spatial and channel attention. Neurocomputing 436, 221–231. URL: https://www.sciencedirect.com/science/article/pii/S0925231221000539, doi:https://doi.org/10.1016/j.neucom.2021.01.032.

[33] Li, Y., Zhao, G., 2021. Intra-and inter-contrastive learning for micro-expression action unit detection, in: Proceedings of the 2021 International Conference on Multimodal Interaction, pp. 702–706. doi:10.1145/3462244.3479956.

[34] Liang, D., Li, L., Wei, M., Yang, S., Zhang, L., Yang, W., Du, Y., Zhou, H., 2022. Semantically contrastive learning for low-light image enhancement, in: Proceedings of the AAAI conference on artificial intelligence, pp. 1555–1563. doi:10.1609/aaai.v36i2.20046.

[35] Liong, S.T., See, J., Phan, R.C.W., Wong, K., Tan, S.W., 2018. Hybrid facial regions extraction for micro-expression recognition system. Journal of Signal Processing Systems 90, 601–617. doi:10.1007/s11265-017-1276-0.

[36] Liu, Y.J., Zhang, J.K., Yan, W.J., Wang, S.J., Zhao, G., Fu, X., 2015. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing 7, 299–310. doi:10.1109/TAFFC.2015.2485205.

[37] Mao, Q., Zhou, L., Zheng, W., Shao, X., Huang, X., 2022. Objective class-based micro-expression recognition under partial occlusion via region-inspired relation reasoning network. IEEE Transactions on Affective Computing 13, 1998–2016. doi:10.1109/TAFFC.2022.3197785.

[38] Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707–6717. doi:10.1109/CVPR42600.2020.00674.

[39] Van den Oord, A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv e-prints , arXiv–1807doi:10.48550/arXiv.1807.03748.

[40] Peng, X., Wang, K., Zhu, Z., Wang, M., You, Y., 2022. Crafting better contrastive views for siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16031–16040. doi:10.1109/CVPR52688.2022.01556.

[41] Qu, F., Wang, S.J., Yan, W.J., Li, H., Wu, S., Fu, X., 2017. CAS(ME)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Transactions on Affective Computing 9, 424–436. doi:10.1109/TAFFC.2017.2654440.

[42] See, J., Yap, M.H., Li, J., Hong, X., Wang, S.J., 2019. MEGC 2019– the second facial micro-expressions grand challenge, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE. pp. 1–5. doi:10.1109/FG.2019.8756611.

[43] Song, Y., Wang, J., Wu, T., Huang, Z., Xiao, J., 2022. Micro-expression recognition based on attribute information embedding and cross-modal contrastive learning, in: 2022 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–7. doi:10.1109/IJCNN55064.2022.9892347.

[44] Sun, B., Cao, S., Li, D., He, J., Yu, L., 2020. Dynamic micro-expression recognition using knowledge distillation. IEEE Transactions on Affective Computing 13, 1037–1043. doi:10.1109/TAFFC.2020.2986962.

[45] Tang, C., Li, S., Zheng, W., Zong, Y., Zhang, S., Lu, C., Zhao, Y., 2024. CFEW: A large-scale database for understanding child facial expression in real world. IEEE Transactions on Affective Computing 15, 990–1003. doi:10.1109/TAFFC.2023.3313782.

[46] Tang, C., Lu, C., Zheng, W., Zong, Y., Li, S., 2021. Multi-view facial action unit detection via deep feature enhancement. Electronics Letters 57, 970–972. doi:10.1049/ell2.12322.

[47] Tian, Y., Krishnan, D., Isola, P., 2020. Contrastive multiview coding, in: European Conference on Computer Vision, Springer. pp. 776–794. doi:10.48550/arXiv.1906.05849.

[48] Tran, T.K., Vo, Q.N., Hong, X., Li, X., Zhao, G., 2021. Micro-expression spotting: A new benchmark. Neurocomputing 443, 356–368. doi:10.1016/j.neucom.2021.02.022.

[49] Wang, S.J., Li, B.J., Liu, Y.J., Yan, W.J., Ou, X., Huang, X., Xu, F., Fu, X., 2018. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. Neurocomputing 312, 251–262. doi:10.1016/j.neucom.2018.05.107.

[50] Wang, S.J., Yan, W.J., Li, X., Zhao, G., Zhou, C.G., Fu, X., Yang, M., Tao, J., 2015. Micro-expression recognition using color spaces. IEEE Transactions on Image Processing 24, 6034–6047. doi:10.1109/TIP.2015.2496314.

[51] Wang, T., Shang, L., 2023. Temporal augmented contrastive learning for micro-expression recognition. Pattern Recognition Letters 167, 122–131. doi:10.1016/j.patrec.2023.02.012.

[52] Wang, Z., Zhang, K., Luo, W., Sankaranarayana, R., 2024. Htnet for micro-expression recognition. Neurocomputing 602, 128196. URL: https://www.sciencedirect.com/science/article/pii/S0925231224009676, doi:10.1016/j.neucom.2024.128196.

[53] Wu, G., Jiang, J., Liu, X., 2023. A practical contrastive learning framework for single-image super-resolution. IEEE Transactions on Neural Networks and Learning Systems , 1–12doi:10.1109/TNNLS.2023.3290038.

[54] Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 3733–3742. doi:10.1109/CVPR.2018.00393.

[55] Xia, B., Wang, S., 2021. Micro-expression recognition enhanced by macro-expression from spatial-temporal domain, in: IJCAI, pp. 1186–1193. doi:10.24963/ijcai.2021/164.

[56] Xia, B., Wang, W., Wang, S., Chen, E., 2020a. Learning from macro-expression: a micro-expression recognition framework, in: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2936–2944. doi:10.1145/3394171.3413774.

[57] Xia, Z., Peng, W., Khor, H.Q., Feng, X., Zhao, G., 2020b. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. IEEE Transactions on Image Processing 29, 8590–8605. doi:10.1109/tip.2020.3018222.

[58] Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., Hu, H., 2021. Self-supervised learning with swin transformers. arXiv preprint arXiv:2105.04553 doi:10.48550/arXiv.2105.04553.

[59] Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X., 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PloS One 9, e86041. doi:10.1371/journal.pone.0086041.

[60] Yan, W.J., Wu, Q., Liu, Y.J., Wang, S.J., Fu, X., 2013. 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), IEEE. pp. 1–7. doi:10.1109/FG.2013.6553799.

[61] Yap, C.H., Kendrick, C., Yap, M.H., 2020. SAMM long videos: A spontaneous facial micro-and macro-expressions dataset, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE. pp. 771–776. doi:10.1109/FG47880.2020.00029.

[62] Ye, M., Zhang, X., Yuen, P.C., Chang, S.F., 2019. Unsupervised embedding learning via invariant and spreading instance feature, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6210–6219. doi:10.1109/CVPR.2019.00637.

[63] Zhai, Y., Zhou, W., Sun, B., Li, J., Ke, Q., Ying, Z., Gan, J., Mai, C., Labati, R.D., Piuri, V., Scotti, F., 2022. Weakly contrastive learning via batch instance discrimination and feature clustering for small sample sar atr. IEEE Transactions on Geoscience and Remote Sensing 60, 1–17. doi:10.1109/TGRS.2021.3066195.

[64] Zhang, L., Hong, X., Arandjelović, O., Zhao, G., 2022. Short and long range relation based spatio-temporal transformer for micro-expression recognition. IEEE Transactions on Affective Computing 13, 1973–1985. doi:10.1109/TAFFC.2022.3213509.

[65] Zhao, S., Tang, H., Mao, X., Liu, S., Zhang, Y., Wang, H., Xu, T., Chen, E., 2023. DFME: A new benchmark for dynamic facial micro-expression recognition. IEEE Transactions on Affective Computing , 1–16doi:10.1109/TAFFC.2023.3341918.

[66] Zhi, R., Hu, J., Wan, F., 2022. Micro-expression recognition with supervised contrastive learning. Pattern Recognition Letters 163, 25–31. doi:10.1016/j.patrec.2022.09.006.

[67] Zhou, H., Huang, S., Li, J., Wang, S.J., 2023. Dual-ATME: Dual-branch attention network for micro-expression recognition. Entropy 25, 460. doi:10.3390/e25030460.

[68] Zhou, L., Mao, Q., Huang, X., Zhang, F., Zhang, Z., 2022. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. Pattern Recognition 122, 108275. doi:10.1016/j.patcog.2021.108275.

[69] Zhu, J., He, W., Wang, F., Chang, H., Lu, C., Zong, Y., 2024. Exploring holistic discriminative representation for micro-expression recognition via contrastive learning. Image and Vision Computing 149, 105186. doi:10.1016/j.imavis.2024.105186.