CMNet: Contrastive Magnification Network for Micro-Expression Recognition

Mengting Wei [1,2], Xingxun Jiang [1,2], Wenming Zheng [1*], Yuan Zong [1*], Cheng Lu[1,3], Jiateng Liu[1,2]

[1]Key Laboratory of Child Development and Learning Science of Ministry of Education
[2]School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
[3]School of Information Science and Engineering, Southeast University, Nanjing, China
{weimengting,jiangxingxun,wenming_zheng,xhzongyuan,cheng.lu,jiateng_liu}@seu.edu.cn

Abstract

Micro-Expression Recognition (MER) is challenging because the Micro-Expressions' (ME) motion is too weak to distinguish. This hurdle can be tackled by enhancing intensity for a more accurate acquisition of movements. However, existing magnification strategies tend to use the features of facial images that include not only intensity clues as intensity features, leading to the intensity representation deficient of credibility. In addition, the intensity variation over time, which is crucial for encoding movements, is also neglected. To this end, we provide a reliable scheme to extract intensity clues while considering their variation on the time scale. First, we devise an Intensity Distillation (ID) loss to acquire the intensity clues by contrasting the difference between frames, given that the difference in the same video lies only in the intensity. Then, the intensity clues are calibrated to follow the trend of the original video. Specifically, due to the lack of truth intensity annotation of the original video, we build the intensity tendency by setting each intensity vacancy an uncertain value, which guides the extracted intensity clues to converge towards this trend rather some fixed values. A Wilcoxon rank sum test (Wrst) method is enforced to implement the calibration. Experimental results on three public ME databases i.e. CASME II, SAMM, and SMIC-HS validate the superiority against state-of-the-art methods.

Introduction

Micro-expression (ME) is a spontaneous facial expression with many applicable contexts, e.g., health diagnosis, homeland security. Recently, an increasing number of methods for improving Micro-Expression Recognition (MER) performance have been proposed (Yan et al. 2013; Kim, Baddar, and Ro 2016).

Among these methods, convolutional neural networks (CNN) have been extensively applied to extract discriminative ME features. However, a ME is localized with slight movements and lasts only a short time, which makes it difficult to spot and recognize. To deal with this problem, some studies propose to magnify ME intensity to make movements more remarkable. Many MER methods adopting magnification stratgies, e.g., Eulerian motion magnification (EMM), Global Lagrangian motion magnification (GLMM), have produced promising experimental results, which demonstrates that increasing ME intensity contributes a lot to boosting MER performance (Bai, Goecke, and Herath 2021; Li, Huang, and Zhao 2018; Le Ngo et al. 2018; Wei et al. 2022a,b).

Generally, the magnification strategies implemented in these methods mainly include two ways, i.e, in the image space and in the feature space. In earlier works, when performing magnification, researchers borrow from a proven magnification technique and apply it to generate magnified images as the input for recognition (Yan et al. 2013; Kim, Baddar, and Ro 2016). This strategy is implemented based on the onset and apex frames between which there is some displacement of pixels. A magnification technique produces images by shifting the displacement with different extent, controlled by setting different amplification factors. However, this strategy can not adjust the intensity information specific to different ME instances. For example, in a ME video, some movements are unrelated for recognition, such as head shaking, eye blinking (Zhang et al. 2020), where the technique may fail to distinguish adaptively, leading to the produced images contaminated with large deformation. Moreover, due to the inconsistency of facial movements on different expressions, seeking appropriate amplification factors for different samples is also an intractable job.

To overcome this drawback, some recent studies (Liu, Zheng, and Zong 2020; Wang et al. 2018) propose to achieve magnification in the feature space, where the features can be dynamically changed during the training stage, thus the intensity clues are more applicable to different ME instances. In their approaches, a feature vector, considered as a representation of intensity, is extracted independently and constrainted by a loss, during the attenuation of which the network achieves intensity enhancement. Although this strategy attempts to magnify intensity tailored to MEs, it lacks explicit interpretation on the features. Specifically, the features they extracted contain not only intensity clues, but also facial texture clues, so it is difficult to interpret whether the performance is improved by magnifying intensity clues or other information. On the other hand, in the original video, the intensity changes with a certain trend over time (Liu, Zong, and Zheng 2022; Zong et al. 2018), while the features may be out of order under no restrictions on its tendency. In that case, the ME movements will be hard to encode.

*Corresponding authors

Based on the limitations above, in this paper, we propose a novel contrastive magnification network for micro-expression recognition, which achieves coordination between increasing intensity as well as confining tendency. We consider in two points of view: intensity enhancement and tendency consistency. For intensity enhancement, we devise an Intensity Distillation Loss, which leverages the contrastive method to distill intensity clues by contrasting the difference between frames, given that the difference between frames lies only in intensity. In its optimization, due to the ME movement is so subtle that may hinder the distinction between frames, we compare different ways, i.e., deterministic, random and probabilistic, to sample negative candidates, the results of which prove that probabilistic sampling produces more discriminative intensity features. For tendency consistency, we aim to calibrate the variation of intensity, implemented by keeping the intensity clues corresponding to each frame varying with the same tendency as the original video. Specifically, we build the variation curve as a prototype by injecting uncertainty to the original intensity vacancies.The uncertainty facilitates the network to perceive the overall tendency, rather than guiding it to converage towards specific values. The augmented intensity clues are then optimized to follow the tendency of the modeled prototype, driven by a Rank Sum Test method. By directing the intensity clues numerically increase and vary consistently with the original video, the network could acquire the intensity clues dynamically during iterations, and ensure its variation conforming to proper tendency in the training stage. Our contributions are summarized as follows:

- We propose an Intensity Distillation loss to encode explicit intensity features, underpinned by the difference between ME frames in a video clip.
- We achieve intensity variation consistency by enforcing a Wilcoxon rank sum test loss, which calibrates the extracted intensity clues to optimize following the built tendency.
- Extensive experiments on three public ME databases validate the efficacy of the proposed method.

Related works

Magnification for Micro-Expressions

Magnification strategies are very helpful to alleviate the hard-to-perceive problem induced by low intensity. In earlier studies, researchers focus on using magnified images, which present more intensive variation, as the network's input. Peng et al. (Peng et al. 2019) employ Eulerian magnification method (EMM) and use magnified images for recognition. Li et al. (Lei et al. 2020, 2021) apply a deep learned magnification filter to produce less noise on magnified images. In general, this strategy only produce static images, where the degree of magnification is impossible to adjust in the training stage, resulting in its inadaptability to different ME samples. Therefore, a more flexible strategy, performed in the feature space, is proposed. Liu et al. (Liu, Zheng, and Zong 2020) propose a deviation enhancement loss which aims to enlarge the distance of features, where the features are generated from different frames. Xia et al. (Xia et al. 2020) impose a loss inequality regularization to calibrate the MicroNet, so the pattern learned in the MacroNet can be involved in ME features. These works provide novel insights in extracting adaptive intensity clues, but is less persuasive on the features constrainted by loss. In addition, they neglect the intensity variation inherent in ME videos, which may lead to the disorder of intensity features along the time axis. For these deficiencies, we manage to acquire explicit intensity clues achieved by a contrastive method, and merge intensity enhancement into confining intensity tendency.

Contrastive Learning

Throughout recent years, contrastive learning has achieved great progress in self-supervised learning. Basically, it aims at learning transferable representations invariant to different data augmentations (Tian et al. 2020), implemented by choosing a set of positive and negative samples anchored by a sample, where the positives share majority similarity with the anchor, and the negatives share barely. The contrastive loss is to optimize the pairwise (dis)similarity at the feature level. Hence, the features reserved, are actually a reflection of the discrepancy between the positives and negatives. Inspired by this efficacy, many methods exploit it as a feature extractor. Tian et al (Tian et al. 2020) studied the influence of different augmentation strategies in generating the positives homologous to the anchor. Their results demonstrated imposing various data augmentations facilitates to keep the task-relevant information intact. Khosla et al (Khosla et al. 2020) extend it into supervised learning to narrow the intra-class distance and to enlarge the inter-class distance. Based on existing works, we introduce the idea of contrastive learning into MER. Our work aims to build the contrast between different frames to extract intensity clues. Furthermore, considerating that the subtle movements may deteriorate the contrast result, we compare three methods in sampling the negative candidates, which shows that probabilistic manner encourages more discriminative learning of intensity clues.

Wilcoxon Rank Sum Test

Statistically, if tests and models conform to parametric standards, i.e., they subject to a known distribution, a parametric test can be used to measure their conformity. Otherwise, nonparametric standards are adopted. The Wilcoxon rank sum test is a nonparametric method introduced by Wilcoxon (Wilcoxon 1992) and further expanded by Mann and Whitney (Mann and Whitney 1947), widely used to measure whether two independent statistics come from the populations with the same distribution (Larson, Farber, and Farber 2009). It is developed under the null hypothesis, i.e, two statistics have the same distribution (Bluman 2014). When testing the hypothesis, a significance level, also called risk level, is predifined as the probability of rejection. For two observations, the test computes the p-value to verify if the hypothesis can be accepted, based on the sum of their ranks. The hypothesis is accepted if and only if the obtained p-value is larger than the chosen significance level. In our approach, we adopted this test to calibrate the extracted intensity clues to conform to the variation of the built prototype.

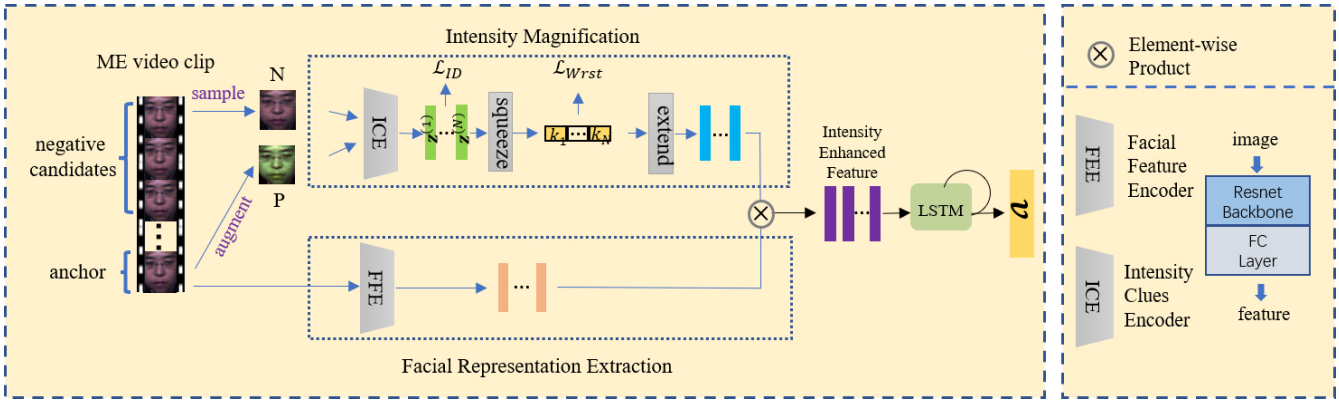


Figure 1: The framework contains two branches, where the lower is used to extract facial texture features, and the upper is for the magnification of intensity clues. We get the intensity enhanced features by conducting the element-wise product of facial texture features and intensity clues. Then, a LSTM layer deals with the enhanced features into a feature $\boldsymbol{v}$ to represent the entire ME video clip for classification. 'P' denotes the positive and 'N' denotes the negative. "squeeze" means mapping each feature vector into a single value and "extend" means mapping the single value to a vector again.

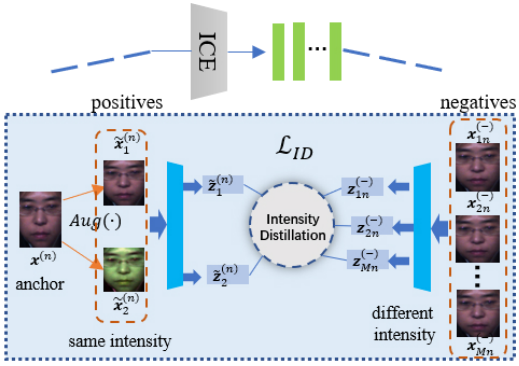


Figure 2: The intensity clues encoder (ICE).

Proposed Method

In this section, we will detail our method. The overall framework, presented in Fig. 1, consists of two branches, where the upper branch is the intensity magnification part which we will introduce elaborately. Our magnification strategy comes from two perspectives: first extract the intensity clues, and then ensure the consistency of variation tendency.

Intensity Enhancement

Extracting Intensity Clues We manage to extract the intensity clues through the intensity clues encoder (ICE), detailed in Fig. 2, which is based on the intensity discrepancy between ME frames. Consider in the same video clip, the only difference rests merely with the intensity clues presented by facial muscle movements, so we extract useful intensity information by seizing the difference.

Contrastive learning is a widely explored method serving as a feature extractor. It implements by modeling the difference between the positives and the negatives, then the difference clues are kept as the feature. Meanwhile, the features shared by the positives are discarded. Inspired by this, we propose to build the intensity as the difference, so the features learned are intensity clues. Specifically, we first unify the length of the videos into $N = 16$ where frames are denoted by $\boldsymbol{x}^{(n)}, n = 1, ..., N$. For a single frame $\boldsymbol{x}^{(n)}$ in a video, it's extended into two views $\tilde{\boldsymbol{x}}_1^{(n)}$ and $\tilde{\boldsymbol{x}}_2^{(n)}$. These two views share the same intensity clues and differs in other clues like color, which are used as the positives. The negatives, denoted as $\boldsymbol{x}_{1n}^{(-)}, ..., \boldsymbol{x}_{mn}^{(-)}, ..., \boldsymbol{x}_{Mn}^{(-)}$, are supposed to be different in intensity, so they are chosen from the remaining frames apart from the current anchor $\boldsymbol{x}^{(n)}$.

Considerating the motion variation is weak in the original video, as we expect to achieve better contrast result, we need to encourage intra-video separability. Here we compare three sampling patterns for the negatives, i.e., deterministic, random, probabilistic. The former two are extensively adopted in contrastive learning and the last is proposed specific to the subtle movement problem in our case.

Deterministic sampling is to choose all the remaining frames in the video. That is, for the current anchor frame $\boldsymbol{x}^{(n)}$, the other $M = 15$ frames except it are negatives, as shown in Fig. 3(a).

Random sampling is implemented by sampling $M = 15$ times randomly in the left frames. Each frame has the chance of $\frac{1}{M}$ to be chosen, as shown in Fig. 3(b).

Probabilistic sampling (Fig. 3(c)) is to sample the negatives based on their similarity with the anchor. That is, for the anchor $\boldsymbol{x}^{(n)}$ to be extended, the candidates sharing larger similarity with it are our priority. The sampling number is also 15.

To be specific, for an anchor $\boldsymbol{x}^{(n)}$, we compute its similarity with the remaining frames in the same video, achieved in a latent space. Then, a Softmax function maps their similarities with the anchor into different probabilities, based on which the negatives are sampled. In this way, the candidates sharing larger similarity with the anchor have larger probability to be chosen. This criterion requires stronger condition to classify $\boldsymbol{z}^{(n)}$, and produces a more rigorous boundary for

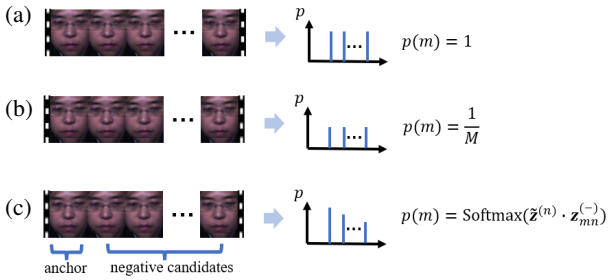


Figure 3: Three ways on sampling the negatives. $\boldsymbol{z}_{mn}^{(-)}$ denotes the feature of a negative candidate.

the positives and negatives, encouraging more discriminative learning of intensity features.

Once the positives and negatives are chosen, we use the ICE block to extract their features, where $\tilde{\boldsymbol{z}}_1^{(n)}$, $\tilde{\boldsymbol{z}}_2^{(n)}$ corresponds to $\tilde{\boldsymbol{x}}_1^{(n)}$, $\tilde{\boldsymbol{x}}_2^{(n)}$, and $\boldsymbol{z}_{mn}^{(-)}$ corresponds to $\boldsymbol{x}_{mn}^{(-)}$ $(m = 1, ..., M)$, respectively. An intensity distillation loss, aimming to distinguish the positives from the negatives, is enforced based on the features, formulated as

$$\mathcal{L}_{ID} = \frac{-1}{N}\sum_{n=1}^{N}\Big[\log\frac{\text{sim}(\tilde{\boldsymbol{z}}_1^{(n)}, \tilde{\boldsymbol{z}}_2^{(n)})}{\text{sim}(\tilde{\boldsymbol{z}}_1^{(n)}, \tilde{\boldsymbol{z}}_2^{(n)}) + \sum_{m=1}^{M}\text{sim}(\tilde{\boldsymbol{z}}_1^{(n)}, \boldsymbol{z}_{mn}^{(-)})} + \log\frac{\text{sim}(\tilde{\boldsymbol{z}}_1^{(n)}, \tilde{\boldsymbol{z}}_2^{(n)})}{\text{sim}(\tilde{\boldsymbol{z}}_1^{(n)}, \tilde{\boldsymbol{z}}_2^{(n)}) + \sum_{m=1}^{M}\text{sim}(\tilde{\boldsymbol{z}}_2^{(n)}, \boldsymbol{z}_{mn}^{(-)})}\Big], \tag{1}$$

where $\text{sim}(a, b) = \exp(a \cdot b/\tau)$. τ is a pre-defined temperature hyper-parameter, and n is the index of frame in a video.

Tendency Consistency

The video clips in ME databases present with continuous variation of expressions, where intensity is minimal at the onset frame and reaches the peak at the apex frame, then starts to decline until the offset frame. It have been widely validated (Bai, Goecke, and Herath 2021) that employing its variation tendency is helpful for recognizing MEs. Therefore, we need to calibrate the extracted intensity clues to be arranged according to a certain tendency, where the tendency is built from the intensity variation of original video clip.

However, when calibrating the intensity clues, the network tends to converge them into specific values. This is contradictory to the limited condition that the truth intensity values in a ME video clip are not available. Thus, when building the variation, we focus on leading the network to perceive the overall intensity variation, rather than compelling the clues to converge into specific values. This is achieved by placing each intensity vacancy a Gaussian distribution, insteading of endowing a fixed value, as shown in Fig. 4(a).

Build the Intensity Tendency Prototype Specifically, we stipulate the intensity varies in the range $[\epsilon, 1]$ where $\epsilon > 0$,

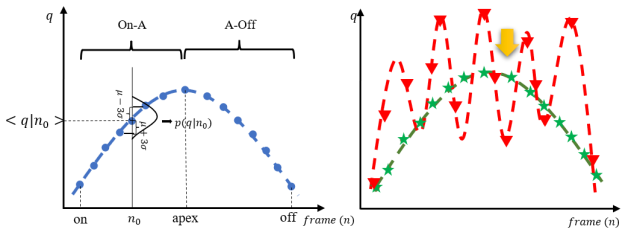


Figure 4: (a) is the intensity tendency in the original video. (b) is the calibration of the extracted intensity clues. Blue dots denote intensity vacancies in the original video. Red triangles denote the intensity clues without calibration. Green stars denote the intensity clues after calibration.

and devide the whole video into two segments with the apex as the boundary. The apex has the peak intensity with 1 as the mean, and the onset along with offset has the minimum intensity with ϵ as the mean. For simplicity, we flag the segment from the onset to the apex as On-A segment, and that from the apex to the offset as A-Off segment. The number of frames of the On-A segment (except the apex) is N_1, and that of the A-Off segment (except the apex) is N_2. $N_1 + N_2 + 1 = N$. To ensure the intensity sampled in both segments can take on a monotonic form, we adopt the 3σ principle, where the majority of sampled points fall in the range of $[\mu - 3\sigma, \mu + 3\sigma]$. Taking On-A segment as an example, in one side of the range, we have the gap between frames computed as $3\sigma = \frac{1-\epsilon}{N_1}$. Therefore, the variance is set as $\sigma = \frac{1-\epsilon}{3N_1}$, so the distribution is given as below:

$$I_{\text{On-A}} \sim \mathcal{N}(\epsilon + (n-1)\frac{1-\epsilon}{N_1}, \frac{1-\epsilon}{3N_1} \cdot \eta_n). \tag{2}$$

As the same way, we compute the intensity of frames in A-Off segments subjecting to:

$$I_{\text{A-Off}} \sim \mathcal{N}\left(1 - (n - N_1 - 1)\frac{1-\epsilon}{N_2}, \frac{1-\epsilon}{3N_2} \cdot \eta_n\right). \tag{3}$$

The intensity of the apex subjects to:

$$I_{\text{A}} \sim \mathcal{N}(1, \min(\frac{1-\epsilon}{3N_1}, \frac{1-\epsilon}{3N_2}) \cdot \eta_n), \tag{4}$$

where $0 < \eta_n < 1$ is a learnable variable to refrain outliners. Each specific intensity value q_n of frame $\boldsymbol{x}^{(n)}$ is sampled from its Gaussian distribution. Next, we use the modeled tendency as the prototype to guide the extracted intensity clues to optimize towards it.

Calibrate the Intensity Clues The ICE block outputs the intensity features corresponding to frames, then the network deals with them into single intensity values, as shown in Fig. 1. Based on the modeled prototype of the intensity curve, we next calibrate the extracted intensity features to vary following the curve, achieved by a Wilcoxon rank sum test method. We adopt this test because our testing samples have small size ($N = 16$), and in the two statistics, i.e., modeled prototype and extracted intensity clues, the ranks among values matters more than their numerical significance. Following the custom of Rank Sum Test, we set

the null hypothesis as that the two sequences share the same distribution, and the significance level is $\alpha = 0.05$.

Let $S_1 = (q_1, ..., q_N)$, $S_2 = (k_1, ..., k_N)$ be the intensity values from the modeled prototype and the network respectively. q_n is a value sampled from each Gaussian distribution, k_n is a value mapped by a linear layer from the intensity feature corresponding to $\boldsymbol{z}^{(n)}$ (see Fig. 1). Rank q and k together and compute the rank of each value:

$$r_n^q = \text{rank of } q_n \text{ among } (q_1, ..., q_N, k_1, ..., k_N), \tag{5}$$

$$r_n^k = \text{rank of } k_n \text{ among } (q_1, ..., q_N, k_1, ..., k_N). \tag{6}$$

The rank sum over S_1 is calculated by summing over r_n^q, denoted as $T_1 = \sum_{n=1}^{N} r_n^q$. Similarly, the rank sum over S_2 is $T_2 = \sum_{n=1}^{N} r_n^k$. Since $N > 10$, the whole rank sum T is close to subjecting to a Gaussian distribution as below (Larson, Farber, and Farber 2009):

$$\begin{aligned} T \sim & \mathcal{N}(\mu_T, \sigma_T) \\ = & \mathcal{N}(\frac{N(2N+1)}{2}, \sqrt{\frac{N^2(2N+1)}{12}}). \end{aligned} \tag{7}$$

The test statistic is computed by the generalized difference test Equation $Z = \frac{T_2 - \mu_T}{\sigma_T}$. Then the Z is used to compute the probability on the standard normal distribution,

$$p = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}. \tag{8}$$

The p-value decides whether the two samples fit hypothesis. If $p > \alpha$, S_1 and S_2 accept the hypothesis so S_2 conforms to the tendency of S_1, otherwise it doesn't. Here we devise a Wilcoxon rank sum test loss to penalize the samples which fail to follow the built prototype with a margin:

$$\mathcal{L}_{Wrst} = \max(0, \xi - (p - \alpha)), \tag{9}$$

where ξ is a margin which can be a fixed hyper-parameter or a learnable variable.

Contrastive Magnification Network

As shown in Fig. 1, after the intensity clues are calibrated, they are conducted the element-wise product with the facial texture features outputted from the facial feature encoder (FFE). Here we obtain the facial features with enhanced intensity clues. For these features, a LSTM is used to capture the dependencies in the sequence and aggregates them into a vector $\boldsymbol{v}$, which is the final representation of a ME video clip used for recognition. Our optimization target is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ID} + \lambda_2 \mathcal{L}_{Wrst} + \lambda_3 \mathcal{L}_C, \tag{10}$$

with $\mathcal{L}_C = -\sum_{c=1}^{C} y_c \log(p_c)$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$, where C is the number of classes. λ_1, λ_2 and λ_3 are the hyper-parameters for weighting three losses respectively.

Experiments

Datasets and Preprocessing

Datasets We conducted experiments on three public spontaneous ME datasets : SMIC-HS (Li et al. 2013), CASME II (Yan et al. 2014), SAMM (Davison et al. 2016).

SMIC-HS consists of 164 ME video clips from 16 participants captured at 100 fps. We use all the samples in this subset with three categories : positive (51), negative (70) and surprise (43).

CASME II comprises of 246 samples out of 26 participants. For a fair comparison with state-of-the-art methods, we used five categories in these samples, including disgust (63), happiness (32), repression (27), surprise (25) and others (99).

SAMM contains 159 ME samples from 32 subjects. Likewise, we choose five categories from this dataset, which are anger (57), contempt (12), happiness (26), surprise (15) and others (26).

Preprocessing We use the tool proposed by (Bulat and Tzimiropoulos 2017) to locate the landmarks on the face, based on which the face area is cropped. Then we resize all the images into 224×224. For each ME video, we unify its length into $N = 16$ by extracing frames at regular intervals, where the interval is computed by N/N_0 with N_0 as the length of the original video clip. For the videos with number of frames less than 16, we use temporal interpolation (Bao et al. 2019) to uniform their length.

Experimental Details

Metrics and Protocals We utilize the Leave-One-Subject-Out (LOSO) cross-validation as the protocal to evaluate the performance of the proposed method. Specifically, for each database, there are totally W folds (W is the number of subjects) experiments. In each fold, the testing set collects the samples from one particular subject while the training set collects the samples from the remaining subjects. The performance is obtained by averaging over the performance on all folds. Two metrics, i.e., accuracy and F1-score, are adopted for evaluation. The accuracy is the ratio between the number of correct predictions B and that of testing samples D: $acc = \frac{B}{D} \times 100\%$. The F1-score is to assess the performance towards unbalanced classes, denoted as F1-score $= \frac{2 \cdot P \cdot R}{P + R}$, with $P = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c}$ and $R = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}$, where P is the precision, R is the recall, and TP_c, FP_c, FN_c denote the true positive, false positive and false negative for the c-th class, respectively.

Parameter Settings We utilize SGD optimizer with $\text{momentum} = 0.9$ and $\text{weight\_decay} = 1 \times 10^{-4}$. The learning rate is set to 0.001 and reduces by half every 50 epochs. The data augmentation strategy used for augmenting anchor samples are ColorJitter, RandomGrayscale and GaussianBlur. To keep the three losses on the same scale, we set $\lambda_1 = 0.07, \lambda_2 = 0.71, \lambda_3 = 0.22$ respectively. These parameters are chosen by implementing a grid search on the CASME II dataset, where the set of parameters with the highest performance are fixed and used when conducting on

Table 1: Comparison with other MER methods under the LOSO protocol in terms of Acc(%) and F1-score(%).

Methods	CASME II		SAMM		SMIC-HS	
	Acc	F1-score	Acc	F1-score	Acc	F1-score
EVM-MER (Le Ngo et al. 2016)	51.00	47.00	N/A	N/A	N/A	N/A
RCNN (Xia et al. 2018)	65.80	N/A	N/A	N/A	65.80	N/A
TSCNN (Song et al. 2019)	74.05	73.27	71.76	69.42	72.74	72.36
ME-Booster (Peng et al. 2019)	70.85	N/A	N/A	N/A	68.90	N/A
STRCN-G (Xia et al. 2019)	**80.30**	**74.70**	78.60	74.10	72.30	69.50
MTMNet (Xia et al. 2020)	75.60	71.10	74.10	73.60	76.80	74.40
Graph-tcn (Lei et al. 2020)	73.98	72.46	75.00	69.85	N/A	N/A
KFC-MER (Su et al. 2021)	72.76	73.75	63.24	57.09	65.85	66.38
AU-GCN (Lei et al. 2021)	74.27	70.47	74.26	70.45	N/A	N/A
ours	78.05	73.99	**78.68**	**77.19**	**79.27**	**80.11**

the other two datasets. The minimum mean ϵ is set as 0.2, and the margin in $\mathcal{L}_{Wrst}$ is a fixed parameter with $\xi = 0.01$.

Experimental Results and Analysis

Experimental Results To validate the efficacy of our method, Table 1 reports its experimental results together with that of the state-of-the-art approaches, mainly including those applying magnification techniques for MER. To give more comprehensive demonstration of our work's efficacy, we add some deep learning-based methods for comparison.

From the tables above, we can observe that our approach outperforms most MER methods with magnifying ME movements, e.g. EVM-MER (Le Ngo et al. 2016), ME-Booster (Peng et al. 2019), Graph-tcn (Lei et al. 2020), AU-GCN (Lei et al. 2021), STRCN-G (Xia et al. 2019), and yields comparative results as against those deep learning MER methods in terms of Acc and F1-score, indicating that the proposed method performs well not only in recognizing MEs correctly, but also in handling unbalanced classes.

Comparison with Magnification in the Image Space On the CASME II, our approach gets obvious improvement compared with most prior works adopting hand-designed filters for ME magnification like EVM-MER, ME-Booster and RCNN. Graph-tcn and AU-GCN adopt a more advanced magnification strategy, demonstrated to have less ringing artifacts and better anti-noise property, to magnify ME images. We improve their accuracy by 4.07%, 3.78%, and improve their F1-score by 1.53%, 3.52, respectively. On the SAMM, our model also exceeds Graph-tcn and AU-GCN by a large margin in terms of accuracy and F1-score.

STRCN-G obtains superior performance than ours with 2.25% higher accuarcy and 0.71% higher F1-score on the CASME II. It is a deep model mainly including recurrent convolutional layers to encode facial appearance and ME movements spatiotemporally. To simulate the intensity variation, it uses Eulerian Video Magnification (EVM) to magnify frames anchored by the onset in a video, and the magnified frames are rearranged into a new sequence, based on which the temporal connectivity is encoded. This method considers the intensity tendency in encoding temporal clues, but acts magnification a bit casually. When magnifying ME intensity, it set an unified amplification factor for all images, so the images may present different results given that their original intensity can be strong or weak. For images with strong intensity, the magnified result may be contaminated with large deformation on the face, which may degrade the effect of recognition. Differently, our method also considers the temporal clues by calibrating intensity tendency, but manage to adjust the magnification degree adaptively through extracting intensity clues corresponding to frames. Thus, we get more improvement on the SAMM and SMIC-HS.

Comparison with Magnification in the Feature Space On the SAMM, SMIC-HS, we also improve the accuracy and F1-score by a large margin compared with most deep learning methods for MER, e.g. TSCNN (Song et al. 2019), RCNN (Xia et al. 2018).

Worth mentioning is the MTMNet, which obtaines much closer results with us. It provides a more roundabout idea, i.e, leveraging the macro-expression as a guidance to learn expression-related features, inspired by the fact that macro-expressions are presented with more intense movements than micro-expressions. In its network, a loss inequality regularization is imposed to calibrate the MicroNet. In this way, the pattern learned in the MacroNet can be helpful in learning ME features. However, this scheme neglects the features specific to MEs. Micro-expressions may hold some features distinct from macro-expressions considering the way it is generated, so referring simply to macro-expressions is inadequate to obtain ME-targeted representation. The intensity clues, extracted from macro-expressions to guide the training of the micro-expressions, is also ambiguous and not belongs to micro-expressions. While in our work, we obtain explicit representation of intensity by extracing the difference between frames, which is more credible.

Ablation Study

Sampling the Negatives in Intensity Distillation We compare the three sampling methods, i.e., deterministic, random, probabilistic, and compute their accuracy and F1-score on three datasets. Results are shown in Table 2. Moreover, we retrieve the intensity values of the built prototype and

that of the features under three methods.

From Table 2 and Fig. 5 (a), we can observe that the probabilistic manner demonstrates improvements than the other two, but with limited degree, and the other two sampling methods show no significant difference under this evaluation. In Fig. 5 (a), it can be found that all three ways perform well in following the prototype, but as for the values of intensity compared with the prototype, no large increment is shown. We suppose this is because of the inherent subtle intensity variation of the video, leading to subtle fluctuation of the similarity between negative candidates and anchor in the latent space. Therefore, the negatives sampled in probabilistic way may not change a lot compared with that in the other two ways.

Table 2: The Acc(%) and F1-score (%) of different sampling methods for the negatives implemented on three datasets.

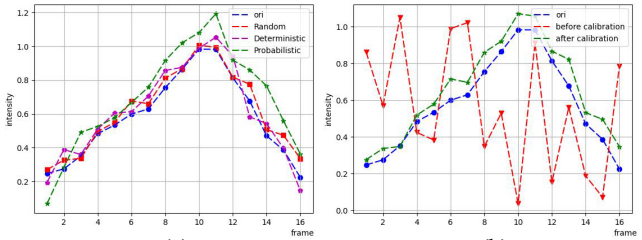
Sampling	CASME II		SAMM		SMIC-HS	
	Acc	F1-score	Acc	F1-score	Acc	F1-score
Deterministic	74.80	70.30	75.74	73.94	78.05	78.35
Random	75.61	71.01	76.47	74.41	76.22	76.61
Probabilistic	**78.05**	**73.99**	**78.68**	**77.19**	**79.27**	**80.11**

Table 3: The Acc(%) and F1-score (%) of different losses implemented on three datasets.

	CASME II		SAMM		SMIC-HS	
	Acc	F1-score	Acc	F1-score	Acc	F1-score
$\mathcal{L}_C$	63.82	56.77	62.50	58.79	65.85	65.89
$\mathcal{L}_C$+$\mathcal{L}_{Wrst}$	65.45	60.37	63.97	60.22	65.85	66.05
$\mathcal{L}_C$+$\mathcal{L}_{ID}$	73.58	68.84	71.32	68.29	74.39	75.20
$\mathcal{L}_C$+$\mathcal{L}_{ID}$+$\mathcal{L}_{Wrst}$	**78.05**	**73.99**	**78.68**	**77.19**	**79.27**	**80.11**

Efficacy of $\mathcal{L}_{Wrst}$ and $\mathcal{L}_{ID}$ We explore the impact of the intensity enhancement as well as tendency consistency separately. Concretely, we set the weight coefficient corresponding to a loss to 0, and devote to explore the best performance using the other, during which the classifier along with its weight coefficient is retained. Experiment with merely the $\mathcal{L}_C$ is the baseline.

From Table 3, we can observe the performance deteriorates when the $\mathcal{L}_{Wrst}$ is removed, but not as much as when we remove the $\mathcal{L}_{ID}$. When we remove the both, the framework will degrade to a simple $\mathrm{Resnet}(\cdot)$ for obtaining spatial clues and a $\mathrm{LSTM}(\cdot)$ for obtaining temporal clues, and it performs worst. Enforcing $\mathcal{L}_{Wrst}$ alone yields little improvement. This is straightforward to interpret since the $\mathcal{L}_{Wrst}$ serves to calibrate the intensity clues extracted by $\mathcal{L}_{ID}$, so it fails to work when there's no intensity clues. On the other hand, the $\mathcal{L}_{ID}$, serving as the intensity extractor, is more helpful to boost the performance, even if the $\mathcal{L}_{Wrst}$ is absent. We speculate that the network can learn something on how to use the extracted intensity clues without the guidance from $\mathcal{L}_{Wrst}$, but may perform worse than it is under explict guidance to vary towards the trend of the original video. This also suggests that enhancing intensity is crucial for recognizing MEs correctly.



(a) (b)

Figure 5: (a) is an example for comparison among three sampling methods. (b) is an example to demonstrate the efficacy of tendency calibration.

Efficacy of Tendency Calibration To show more instinctive effect of the proposed method, we plot the intensity tendency of the prototype, and that of the intensity clues before and after calibration, as shown in Fig. 5(b).

Before calibration, the intensity values fluctuate with large scope, so there is no obvious tendency presented. After calibration, those values are adjusted to follow the trend of the built prototype, and their magnitude are also changed. We can also find the intensity values after tendency calibration are not much larger than that in the prototype. The reason of this lies in the $\mathcal{L}_{Wrst}$, during the optimization of which the intensity clues are encouraged to be allocated into the same distribution with the prototype's intensity values. Therefore, on the whole, the intensity values are much larger, but won't deviate from the prototype values by a large margin. Note that this situation has no influence on the recognition, since the prototype's intensity values are hypothetic and do not represent the truth intensity. As we mentioned before, the purpose of this prototype is to guide the network to learn the overall tendency, where the values are insignificant. Thus, as long as the extracted intensity clues follow its tendency, the network can be considered to fulfill tendency consistency along the time axis.

Conclusion

In this paper, we provide a new insight towards emphasizing micro-expression's (ME) intensity. Our strategy comes from two perspectives: intensity enhancement and tendency consistency. We manage to extract explicit intensity representation by leveraging the difference between frames. We achieve the intensity variation consistency with the original video clip. Experimental results conducted on three public ME databases validate the efficacy of the proposed magnification strategy.

Acknowledgement

This work was supported in part by the NSFC under the Grants U2003207, 61921004, and 61902064, in part by the Jiangsu Frontier Technology Basic Research Project under the Grant BK20192004, and in part by the Zhishan Young Scholarship of Southeast University. We thank the support of China Scholarship Council.

References

Bai, M.; Goecke, R.; and Herath, D. 2021. Micro-Expression Recognition Based On Video Motion Magnification And Pre-Trained Neural Network. In *2021 IEEE International Conference on Image Processing (ICIP)*, 549–553. IEEE.

Bao, W.; Lai, W.-S.; Ma, C.; Zhang, X.; Gao, Z.; and Yang, M.-H. 2019. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3703–3712.

Bluman, A. 2014. *Elementary Statistics: A step by step approach 9e*. McGraw Hill.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

Davison, A. K.; Lansley, C.; Costen, N.; Tan, K.; and Yap, M. H. 2016. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing*, 9(1): 116–129.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673.

Kim, D. H.; Baddar, W. J.; and Ro, Y. M. 2016. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In *Proceedings of the 24th ACM international conference on Multimedia*, 382–386.

Larson, R.; Farber, E.; and Farber, E. 2009. *Elementary statistics: Picturing the world*. Pearson Prentice Hall.

Le Ngo, A. C.; Johnston, A.; Phan, R. C.-W.; and See, J. 2018. Micro-expression motion magnification: Global lagrangian vs. local eulerian approaches. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 650–656. IEEE.

Le Ngo, A. C.; Oh, Y.-H.; Phan, R. C.-W.; and See, J. 2016. Eulerian emotion magnification for subtle expression recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1243–1247. IEEE.

Lei, L.; Chen, T.; Li, S.; and 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*, 1571–1580.

Lei, L.; Li, J.; Chen, T.; and 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*, 2237–2245.

Li, X.; Pfister, T.; Huang, X.; Zhao, G.; and 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)*, 1–6. IEEE.

Li, Y.; Huang, X.; and Zhao, G. 2018. Can micro-expression be recognized based on single apex frame? In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 3094–3098. IEEE.

Liu, J.; Zheng, W.; and Zong, Y. 2020. SMA-STN: Segmented Movement-Attending Spatiotemporal Network forMicro-Expression Recognition. *arXiv preprint arXiv:2010.09342*.

Liu, J.; Zong, Y.; and Zheng, W. 2022. Cross-database micro-expression recognition based on transfer double sparse learning. *Multimedia Tools and Applications*, 1–18.

Mann, H. B.; and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Peng, W.; Hong, X.; Xu, Y.; and Zhao, G. 2019. A boost in revealing subtle facial expressions: A consolidated eulerian framework. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–5. IEEE.

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

Su, Y.; Zhang, J.; Liu, J.; and Zhai, G. 2021. Key Facial Components Guided Micro-Expression Recognition Based on First & Second-Order Motion. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33: 6827–6839.

Wang, S.-J.; Li, B.-J.; Liu, Y.-J.; Yan, W.-J.; Ou, X.; Huang, X.; Xu, F.; and Fu, X. 2018. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing*, 312: 251–262.

Wei, M.; Zheng, W.; Zong, Y.; Jiang, X.; Lu, C.; and Liu, J. 2022a. A Novel Micro-Expression Recognition Approach Using Attention-Based Magnification-Adaptive Networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2420–2424. IEEE.

Wei, M.; Zong, Y.; Jiang, X.; Lu, C.; and Liu, J. 2022b. Micro-Expression Recognition Using Uncertainty-Aware Magnification-Robust Networks. *Entropy*, 24(9): 1271.

Wilcoxon, F. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, 196–202. Springer.

Xia, B.; Wang, W.; Wang, S.; and Chen, E. 2020. Learning from macro-expression: A micro-expression recognition framework. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2936–2944.

Xia, Z.; Feng, X.; Hong, X.; and Zhao, G. 2018. Spontaneous facial micro-expression recognition via deep convolutional network. In *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. IEEE.

Xia, Z.; Hong, X.; Gao, X.; Feng, X.; and Zhao, G. 2019. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Transactions on Multimedia*, 22(3): 626–640.

Yan, W.-J.; Li, X.; Wang, S.-J.; Zhao, G.; Liu, Y.-J.; Chen, Y.-H.; and Fu, X. 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1): e86041.

Yan, W.-J.; Wu, Q.; Liang, J.; Chen, Y.-H.; and Fu, X. 2013. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, 37(4): 217–230.

Zhang, T.; Zong, Y.; Zheng, W.; Chen, C. P.; Hong, X.; Tang, C.; Cui, Z.; and Zhao, G. 2020. Cross-database micro-expression recognition: A benchmark. *IEEE Transactions on Knowledge and Data Engineering*.

Zong, Y.; Huang, X.; Zheng, W.; Cui, Z.; and Zhao, G. 2018. Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. *IEEE Transactions on Multimedia*, 20(11): 3160–3172.