A Novel Magnification-Robust Network with Sparse Self-Attention for Micro-expression Recognition

Mengting Wei, Wenming Zheng*, Xingxun Jiang, Yuan Zong*, Cheng Lu, Jiateng Liu
Key Laboratory of Child Development and Learning Science of Ministry of Education
School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China
{weimengting,wenming_zheng,jiangxingxun,xhzongyuan,cheng.lu,jiateng_liu}@seu.edu.cn

***Abstract*—Existing works for spontaneous Micro-Expression Recognition (MER) tend to encode Micro-Expression (ME) movements to get more discriminative features. However, MEs' low intensity makes the capture for motion extremely difficult, and the widely adopted unified-magnification strategy is prone to noise and lacks flexibility. To this end, this paper provides a new insight to encode ME motion and tackle magnification noise. Specifically, we reconstruct a new sequence via magnification techniques to make subtle ME movements more distinguishable. Afterward, Sparse Self-Attention (SSA) rectifies self-attention with Locality Sensitive Hashing (LSH), cutting the space into several hush buckets of related features. Only keys in the same bucket are operated in the attention term for every query feature. The resulting sparsity in the attention matrix prevents the network from attending features stemming from less-informative magnification degrees which could be regarded as noise, while retains the sequence modelling capability of standard self-attention. Extensive experiments on three public MER databases demonstrate our superiority against the state-of-the-art methods.**

I. Introduction

Micro-expression (ME) is a kind of facial expression which appears when people attempt to conceal their actual emotional states [1]. It is helpful to understand real human emotions and therefore has potential application prospects in many fields, e.g., lie detection, human-computer interaction, and national security [2]. However, ME is characterized with shorter temporal duration—nearly 1/25 to 1/3 seconds and with localized subtle variation [3], which makes correctly recognizing MEs more difficult. Targeting at this problem, more and more approaches have been proposed to seek a more precise representation for micro-expression recognition (MER).

Since facial movement is a dynamic variation, encoding motion features is indispensible to acquire a more comprehensive representation. In recent works, many deep learning-based methods specialized in capturing dependencies in long sequences, e.g., 3D Convolutional Neural Networks (3D CNN) [4], [5], Long Short-Term Memory Network (LSTM) [6] are employed to capture the motion of MEs. However, MEs are reflected in the local area with low intensity of muscle movements, resulting in the perception of motion variation being complex. To tackle this issue, a technique to magnify these subtle movements can be helpful to improve the performance of MER. The most commonly used magnification techniques, e.g., Eulerian Motion Magnification (EMM) [7], Global Lagrangian Motion Magnification (GLMM) [8], Learning-based Video Motion Magnification (LVMM) [9], have done a good job in magnifying subtle movements. Inspired by this, some works adopt magnification techniques to magnify ME motion, whose results confirm the effectiveness of ME magnification in promoting MER performance [10], [7], [8], [11].

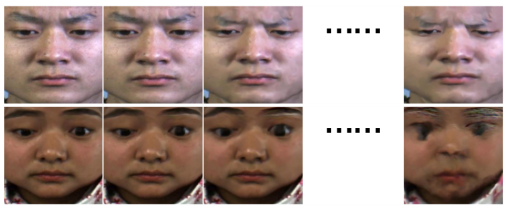

Fig. 1: A set of consistent magnification levels on two ME samples. The first, second row displays "disgust" and "surprise", respectively. A large magnification level is helpful for extracting discriminative features for "disgust", but can cause noise on "surprise".

Despite the progress, a potential problem in the magnification process is commonly neglected: physical techniques for magnifying MEs bring much noise and blurs when the magnification level is overlarge. As the magnification level grows, muscle movements become more intense but may also appear deformation in the face. More specifically, for some subjects, a larger magnification level is required to extract discriminative motion features. In contrast, for others, the same magnification level may induce severe deformation, resulting in the noise dominating the magnified image, as shown in Fig. 1. The deformation noise generated by the unified-magnification strategy is useless in the training process and may exert negative influence on subsequent tasks. Thus, it is imperative to devise a method that can effectively capture powerful motion representation while being robust to noise brought by magnification.

In this paper, to address the issues above, we provide a new insight towards effectively encoding motorial features in ME video clips, namely Magnification-Robust Network with Sparse Self-Attention (MRN). To make the network perceive ME movements more easily, we use magnification techniques to rebuild a sequence which has more discriminative represen-

* Corresponding authors

tation for motion. To suppress the noise during magnification, we impose sparsity constraints via Locality Sensitive Hashing (LSH) into a self-attention block, which can adaptively reserve highly-correlated ME features and discard the uncorrelated ones. As a result, our network retains the global modeling ability of the standard self-attention while tackling defective magnification problems through sparse representation. Furthermore, our network is efficient in time by zeroing out irrelevant information. The main contributions are summarized as follows:

- A sequence more able to reflect ME movements is rebuilt via magnification techniques.
- To preserve as much as ME-related features and suppress magnification noise, we propose a sparse self-attention (SSA) block through enforcing sparsity in attention terms using Locality Sensitive Hashing (LSH).
- Extensive experiments conducted on three widely used databases manifest that our approach yields competitive results compared with state-of-the-art MER methods.

II. Proposed Method

A. The Framework

The framework of our proposed model is shown in Fig.2. To begin with, we magnify MEs using the onset and apex frame in a ME video with a set of consecutive amplification factors (AFs). As the AF grows, muscle movements are more apparent, resulting in the capture of movements more obtainable than the original sequence. Then, to adaptively reserve useful magnification levels, we zero out some irrelevant information in the sparse self-attention (SSA) block. At the end of the framework, we jointly concatenate the motion representation and initial spatial information for final classification.

B. ME Sequence Reconstruction

Considering that ME intensity variation is very subtle to perceive, instead of using the original video clip, we simulate this intensity variation through magnification techniques, as demonstrated in the left box in Fig.2. Different from the original one in which intensity grows with time in a non-linear manner, our reconstructed sequence discards the temporal meaning but focus on enhancing motorial representation by exploring their underlying linear forms.

Considering hand-crafted magnification techniques require more manual intervention, in our practice, to enable the network to operate in an end-to-end manner, we adopt transfer learning strategy to initialize the network. Our magnification technique is based on a deep network pre-trained on a large-scale database devised by Oh et al. [12]. Following Wadhwa et al.'s definition of motion magnification [13], for a frame in a ME video clip at position $P = (x, y)$, denoted as $I(P, t) = f(P+\sigma(P, t))$, where $\sigma(P, t)$ represents the motion field at P and time t, the magnified image is

$$\tilde{I} = f(P + (1 + \alpha)\sigma(P, t)), \tag{1}$$

where α is amplification factor. In our case, we use the onset frame I_{onset} and apex frame I_{apex} with certain motion offset in the same ME video to generate magnified images I_{mag}, denoted as

$$I_{mag} = f(I_{onset} + (1 + \alpha)|I_{apex} - I_{onset}|), \tag{2}$$

where $|\cdot|$ denotes the pixel-wise substraction. The magnified images are then arranged according to their corresponding AFs to construct a new sequence. Compared with the original, the new sequence has more powerful ability to reflect ME motion since it is built on the magnified difference between the apex and onset.

C. Magnification-Robust Network with Sparse Self-attention

To effectively capture ME dynamic motion in the reconstructed sequence, we employ self-attention for its capability in parallelly modelling long sequences. However, a prominent drawback in magnification process lies in the uncontrollable noise caused by overlarge magnification levels, which means that some feature vectors from the rebulit sequence is useless. To alleviate this problem, we zero out some irrelevant magnification information by imposing sparsity constraint.

1) Self-Attention: According to [14], a self-attention block integrates sequence information by enumerating all positions, which is mainly functioned by three learnable feature matrices, i.e., $\boldsymbol{Q} \in \mathbb{R}^{L\times d_q}$ to match others, $\boldsymbol{K} \in \mathbb{R}^{L\times d_k}$ to be matched and $\boldsymbol{V} \in \mathbb{R}^{L\times d_v}$ for the information to be extracted, formulated as

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V}, \tag{3}$$

where $d_q = d_k$ and d_q, d_k, d_v are the dimension of query $(\boldsymbol{q})$, key $(\boldsymbol{k})$ and value $(\boldsymbol{v})$, respectively. L represents the sequence length. Equ. 3. implements by operating each $\boldsymbol{q}_i$ to multiply the $\boldsymbol{k}_j$ one by one at every position of the sequence, where $i, j \in \{1, ..., L\}$. In our practice, we expect the features from excessive deformation images are not operated product but set as zero in the attention matrix, so we only operate attention on the locations with highly-correlated features. Therefore, for some queries in $\boldsymbol{Q}$ from magnified frame with over-large deformation, it's very likely that no keys in $\boldsymbol{K}$ share high correlation with them. These cases can actually be set to zero in the term $softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}})$.

2) Magnification-Robust Sparsity: The process is shown in the right dotted box of Fig.2. During the training, we want the attention matrix to keep the most relevant elements, a natural method is to sort the values and keep those larger. However, manually selecting thresholds is contrary to our requirement for "adaptive". We propose to adopt Locality Sensitive Hashing (LSH) to ensure the more relevant elements have higher probability to be operated attention. The way that LSH operates is to take the space and cut it into several separate regions with hyper-planes. Each region corresponds to a hash bucket. Hash function projects a vector to the space and if two vectors have higher correlation, they are likely to fall into the same hash bucket. In our case, feature vectors containing majority of noise will fall into different hash buckets, as demonstrated in Fig. 3. By refraining the attention

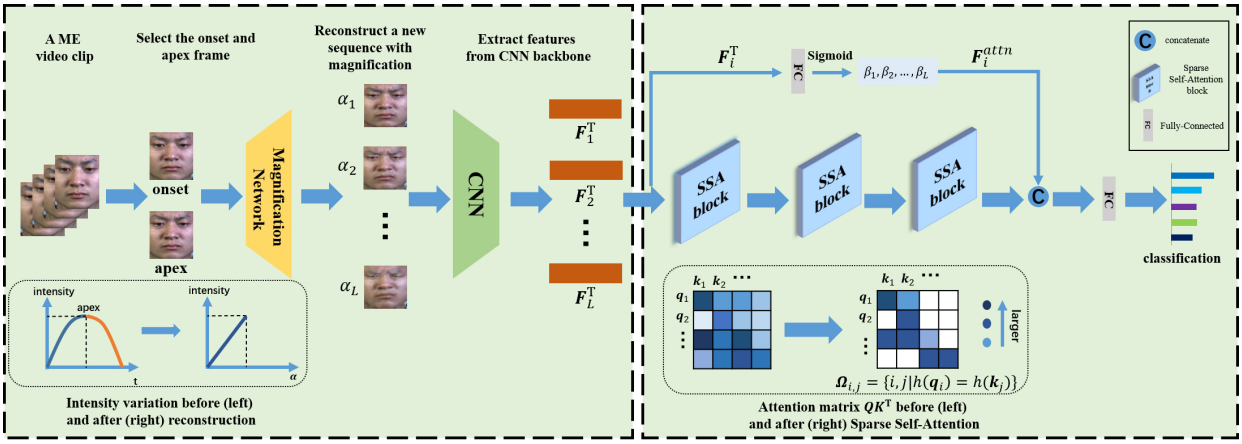

Fig. 2: Framework of our proposed method. First, we pick the onset and apex frame in a ME video and magnify the ME with different amplification factors to construct a new sequence. After that, we use Resnet-18 to extract features from the new sequence. Then, three Sparse Self-Attention (SSA) blocks are utilized to encode motion while suppress the noise generated by magnification. Feature vectors with different attention scores are concatenated with the motion features outputted from SSA at last. A fully-connected layer and the softmax layer is followed to calculate probabilities.

between noise and allowing that between highly-correlated ME features, our model filters the useful information for better ME representation.

Formally, suppose there are three hyper-planes randomly generated in the space, denoted as $\boldsymbol{h}_1$, $\boldsymbol{h}_2$ and $\boldsymbol{h}_3 \in \mathbb{R}^d$, they cut the space into $2^3 = 8$ regions. Each region is taken as a bucket for holding vectors. For a feature vector $\boldsymbol{x} \in \mathbb{R}^d$, if the product $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{h} > 0$, it is projected into one side of $\boldsymbol{h}$ with tag "0", otherwise "1". For instance, if we compute $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{h}_1 < 0, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{h}_2 < 0, \boldsymbol{x}^{\mathrm{T}}\boldsymbol{h}_3 > 0$, the vector $\boldsymbol{x}$ falls in the bucket tagged $h(x) =$ "001". All the buckets compose the set $H =$ {"001", "000", "010", "011", "101", "100", "111", "110"}. To form the attention matrix $\boldsymbol{QK}^{\mathrm{T}}$, we first hash all the queries and keys into the space and then manage to keep those in the same bucket. The result attention is operated merely by queries and keys in the index set:

$$\begin{aligned} \Omega_{i,j} =& \{(i,j)|h(\boldsymbol{q}_i) = h(\boldsymbol{k}_j)\}, \\ & \text{s.t.} \quad 0 < i, j < L. \end{aligned} \tag{4}$$

Considering that the hyper-planes are randomly generated in hashing vectors, there might be the case that highly-correlated elements are hashed into different buckets. To tackle this, we propose to hash those elements multi times with different sets of hyper-planes $\{H^{(1)}, H^{(2)}, ..., H^{(n_t)}\}$ and get the union of results, formulated as

$$\Omega_{i,j} = \bigcup_{r=1}^{n_t} \Omega_{i,j}^{(r)}, \tag{5}$$

where $\Omega_{i,j}^{(r)} = \{(i,j)|h^{(r)}(\boldsymbol{q}_i) = h^{(r)}(\boldsymbol{k}_j)\}$ and n_t is the number of times we generate hyper-planes. Therefore, original attention matrix shown in formula (3) can be rewritten as

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \sum_{(i,j)\in\Omega_{i,j}} softmax(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V}. \tag{6}$$

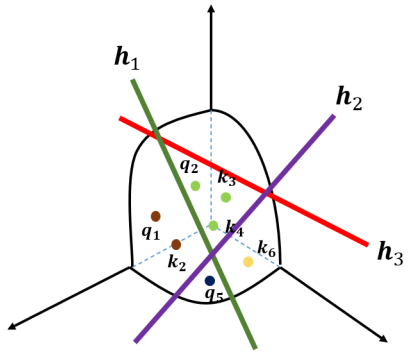

Fig. 3: An example of locality sensitive hashing (LSH). Three hyper-planes $\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3$ cut the space into eight buckets where nearby feature vectors are more likely to be projected into the same bucket, e.g., $(\boldsymbol{q}_1, \boldsymbol{k}_2)$, $(\boldsymbol{q}_2, \boldsymbol{k}_3, \boldsymbol{k}_4)$. Vectors from magnification noise are hashed into other buckets discretely, e.g., $\boldsymbol{q}_5, \boldsymbol{k}_6$.

Compared with Equ. 3, the sparse one retains useful magnification information while avoids the interference of useless noise. The pseudocode is shown in Algorithm 1.

D. Feature Fusion

Inspired by the effectiveness of residual connection [15], we concatenate the feature vectors of magnified MEs and motion features encoded by SSA blocks. To suppress the noise in the former branch, we assign different attention scores on the feature vectors. Specifically, for feature vectors corresponding to magnified frames in the reconstructed sequence, denoted by $[\boldsymbol{F}_1^{\mathrm{T}}, \boldsymbol{F}_2^{\mathrm{T}}, ..., \boldsymbol{F}_L^{\mathrm{T}}]$, we assign different scores by an attention unit composed of a linear fully-connected (FC) layer and a sigmoid activation function. Attention scores among those

Algorithm 1: Sparse Self-Attention with Locality Sensitive Hashing

Input: L queries ($\boldsymbol{q}$) and L keys ($\boldsymbol{k}$), number of hash table n_t, number of hyper-planes n_p

Output: Sparse attention matrix $\sum_{(i,j)\in\Omega_{i,j}} softmax(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V}$

for ***all*** $table \in \{1, ..., n_t\}$ **do**
  Randomly generate n_p vectors as hyper-planes, $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{n_p}$;
  Mark each bucket partitioned by these hyper-planes, one side of the plane is labeled '0' and the other side is labeled '1';
  Hash all the queries and keys to different buckets by computing their product with hyper-planes;
  Collate the key-query pairs falling into the same bucket.
end
for ***all*** *buckets* $\{h_p\}_{p=1}^{2^{n_p}}$ **do**
  Get the union of key-query pairs from all tables;
end
Only operate dot product for those queries and keys in the same bucket;

feature vectors from the same sequence can be calculated as:

$$\beta_i = \sigma(\boldsymbol{F}_i^{\mathrm{T}} \boldsymbol{q}^0), \tag{7}$$

where $\boldsymbol{q}^0$ is the parameter of FC layer, and $\sigma(\cdot)$ denotes sigmoid function. The new weighted feature representation of a single magnified frame can be formulated as:

$$\boldsymbol{F}_i^{attn} = \beta_i \boldsymbol{F}_i^{\mathrm{T}}. \tag{8}$$

In this way, magnified frames with excessive deformation could be adaptively assigned smaller weights during the training. After concatenation, a FC layer with Softmax function is employed to calculate the classification probabilities.

III. Experiments

A. Databases and Protocols

1) Databases: We use three representative databases to evaluate the performance of our method, i.e., CASME II, SAMM, and the subset HS of SMIC [16], [17], [18]. CASME II contains 255 video clips from 26 participants. We select five emotion types: happiness (32), disgust (63), repression (27), surprise (25) and other (99). SAMM contains contains 159 ME samples from 32 participants. We also select five categories from this database, which are anger (57), happiness (26), contempt (12), surprise (15) and other (26). SMIC-HS contains 164 spontaneous micro-expressions with samples divided into 3 classes: positive (51), negative (70), and surprise (43).We use all the samples for experimentation.

As we only need onset and apex frame in the sequence reconstruction procedure, on the CASME II and SAMM, we directly use the annotation provided. Although the annotation for the apex on the SMIC-HS isn't available, our model doesn't rely on accurately locating the apex since we can obtain distinct magnification result only based on subtle movements, so we locate the middle frame as the apex.

2) Protocols: We adopt the leave-one-subject-out (LOSO) protocol to evaluate the performance of our approach, which is proved reliable and widely used in MER. The metric for calculating the accuracy rate is $acc = \frac{T}{N}$ where T is the total number of correct predictions and N is the total number of samples for test. To assess the ability towards unbalanced ME databases problem, we use the F1-score calculated as $F1 - score = \frac{1}{c}\sum_{c=1}^{C} \frac{2\times P_c \times R_c}{P_c + R_c}$, where P_c and R_c are the precision and recall of the c-th micro-expression, respectively, and C is the number of ME classes.

B. Implementation details

1) Preprocessing: In each ME sequence, we calculate 68 landmarks around the face on the onset frame according to [19] and crop the facial area by these coordinates. Then, the apex frame in the same video is aligned by the same coordinates as the onset. Finally, we resize all the images to 224×224.

During training, we use the first few layers of pre-trained Resnet-18 [15] to extract shallow features of the magnified MEs with vector size 1024. Resnet-18 is pre-trained on a macro-expression dataset: FER+ [20] which share similarity in some low-level features with MEs. For position embedding, we adopt sine and cosine functions of different frequencies:$PE_{(pos,2d)} = \sin(pos/10000^{2d/d_{model}})$, $PE_{(pos,2d+1)} = \cos(pos/10000^{2d/d_{model}})$, where pos is the position in the sequence and d is the specific dimension of d_{model}.

2) Training Details: For the sparse self-attention, we set the number of hyper-planes to $n_p = 3$ and the sequence length $L = 32$. The number of hash tables is $n_t = 4$ and attention heads is $n_h = 8$. Amplification factor range is $[2 : 1 : 13]$. We optimize the model by ADAM optimizer with learning rate = 2×10^{-4} reduced by 0.1 after every 10 epochs.

C. Experimental Results

TABLE I: Experimental Results (Accuracy/F1-score) on the CASME II with 5 classes under the LOSO protocal.

Methods	Accuracy(%)	F1-score(%)
LBP-TOP + AdaBoost (2014) [21]	43.78	33.37
STRBP (2017) [22]	64.37	N/A
HIGO+Mag (2018) [10]	67.21	N/A
ME-Booster (2019) [7]	70.85	N/A
TSCNN (2019) [23]	80.97	**80.70**
Graph-tcn (2020) [9]	73.98	72.46
AU-GCN (2021) [24]	74.27	70.47
Ours	**81.06**	78.42

In the comparison with other reported methods, we present some works using hand-crafted features like LBP-SIP [28], LBP-TOP + AdaBoost [21], STRBP [22], etc. We also present the state-of-the-art MER studies conducted on three databases, including those with techniques magnifying ME intensity, e.g., HIGO+Mag [10], ME-Booster [7], Graph-tcn [9], AU-GCN

TABLE II: Experimental Results (Accuracy/F1-score) on the SAMM with 5 classes under the LOSO protocal.

Methods	Accuracy(%)	F1-score(%)
DSSN (2019) [25]	57.35	46.44
TSCNN (2019) [23]	71.76	69.42
Graph-tcn (2020) [9]	75.00	69.85
MTMNet (2020) [26]	74.10	73.60
AU-GCN (2021) [24]	74.26	70.45
GEME (2021) [27]	65.44	54.67
MERSiamC3D (2021) [4]	64.03	60.00
Ours	**77.61**	**74.32**

TABLE III: Experimental Results (Accuracy/F1-score) on the SMIC-HS with 3 classes under the LOSO protocal.

Methods	Accuracy(%)	F1-score(%)
LBP-SIP (2014) [28]	62.80	N/A
STRBP (2017) [22]	60.98	N/A
Dual-Inception Network (2019) [29]	66.00	67.00
3D-CNNs (2019) [5]	66.30	N/A
TSCNN (2019) [23]	72.74	72.36
MTMNet (2020) [26]	76.80	74.40
GEME (2021) [27]	64.63	61.58
Ours	**79.88**	**75.50**

[24], and those deep models, e.g., TSCNN [23], 3D-CNNs [5], MERSiamC3D [4], DSSN [25], GEME [27]. Results are shown in TABLE I, II, III.

1) Comparison Results to Methods Magnifying ME Intensity: From the tables above, we can observe that we improve the accuracy by 10.25% compared with ME-Booster and by 13.25% compared with HIGO+Mag on the CASME II, both of which adopt Eulerian Motion Magnification (EMM) for magnification. Moreover, on the SAMM, our method exceeds other state-of-the-art MER methods, i.e., Graph-tcn, AU-GCN by 2.61% and 3.35%, respectively, which adopt the same magnification technique as ours. These MER methods adopt magnification techniques by setting a unified AF for all ME samples, which is not suitable and can bring noise on some samples. While in our framework, we use a multi-magnification strategy and meanwhile suppress the noise via SSA blocks to make learned features more discriminative, which is shown to be more effective.

2) Comparison Results to Hand-Crafted Features and Deep Models: Our method also surpasses early hand-crafted features by a large margin, e.g., LBP-TOP + AdaBoost, LBP-TOP, which demonstrates the superiority of deep networks in extracting ME-specific features. Moreover, we also achieve better results on all three databases than most state-of-the-art deep models, e.g., TSCNN, 3D-CNNs, MERSiamC3D, DSSN, GEME. Merely on the CASME II, TSCNN presents higher F1-score than ours by 2.28%. We have carefully checked its ablation experiments and found that dynamic-temporal and static-spatial information are two major modules having more discriminative ability to represent a ME video clip. Similar with TSCNN, our method also focuses on the dynamic-temporal and static-spatial information. However, instead of directly encoding these useful clues, ours designs ME sequence reconstruction to enhance both clues. Thus, MRN can achieve better performance on the other two databases.

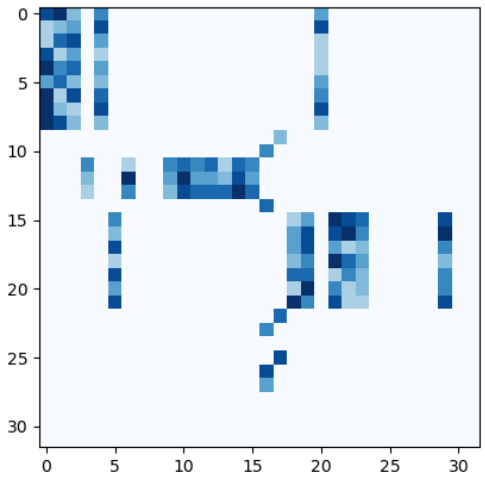

Fig. 4: An example ($sub02/EP03\_02f$ on the CASME II) of the sparse attention matrix $\sum_{(i,j)\in\Omega_{i,j}} softmax(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d_k}})$ where darker color denotes larger values. The rows represent the index of $\boldsymbol{q}$ and the columns represent the index of $\boldsymbol{k}$.

D. The Effectiveness of Locality Sensitive Hashing

In order to validate that the magnified images with excessive noise can be discarded through sparse representation of the attention matrix, we retrieve the feature vectors $\boldsymbol{q}$ and $\boldsymbol{k}$ and plot the result $\sum_{(i,j)\in\Omega_{i,j}} softmax(\frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V}$ by only operating product of $\boldsymbol{q}$ and $\boldsymbol{k}$ in the same hush bucket. Then we verify them with the frames from reconstructed sequence. Large indexes of $\boldsymbol{q}$ and $\boldsymbol{k}$ denote features from magnified frames with large amplification factors. As shown in Fig. 4.

It's clear to see that in general, non-zero elements are concentrated near the diagonal of the matrix, which demonstrates higher correlation between adjacent frames than distant frames. Besides, the lower right corner of the matrix where $\boldsymbol{q}$ and $\boldsymbol{k}$ are with large indexes (corresponding to images with large AF), is almost entirely zero values, indicating the effectiveness LSH for removing distorted images in the reconstructed sequence. There are exceptions that some feature vectors sharing adjacent indexes are hushed into different buckets, e.g., $\boldsymbol{k}_{20}, \boldsymbol{q}_9, \boldsymbol{k}_{17}$. We speculate that when the hyperplanes are partitioned, some two vectors close to each other may fall into two adjacent buckets, even if their distance are closer than features in the same bucket. Nonetheless, its impact is limited since we employ multi-head attention to extract multiple levels of information, thus the experimental results are not seriously depraved.

E. Ablation Study

1) Settings on Amplification Factor Range: In sequence reconstruction procedure, when we operate magnification on a ME sample with a set of AFs, the movements of facial muscles

get more apparent as the AF grows. If we set AF too small, the magnified ME movements may be undiscriminative for extracting class-specific features. Therefore, to make all MEs are sufficiently magnified, and simultaneously to reduce the chance that images are completely deformed due to excessive magnification occupy the majority of the whole sequence, we conducted a set of experiments on setting the range of amplification factors. In our practice, we fix the left bound of factor range as 2 and dedicate to find the right bound with best performance. Results are shown in TABLE IV.

TABLE IV: Settings for Magnification Range

Acc(%) \ AF / Database	11	12	**13**	14	15	16
CASME II	76.79	78.22	**81.06**	80.21	79.55	79.55
SAMM	72.35	76.65	**77.61**	77.52	75.59	75.12
SMIC-HS	78.46	79.27	**79.88**	78.91	78.42	75.25

From the chart above, the performance first grows with amplification factors and then falls. We speculate that when the AF is overlarge, the noise dominates the reconstructed sequence, resulting in the attention matrix so sparse that the network fails to extract enough motion clues. Therefore, we uniformly set AFs range $[2:1:13]$ in our experiments.

2) Hyper-Planes n_p and Multi-Hashtables n_t: The sparsity of SSA is partly controlled by the number of hyper-planes cut in the space, or rather, by the number of hash buckets. If noise dominates the sequence while the number of buckets is relatively small, SSA would barely separate them into discrete buckets. If the hyper-planes are redundant in the space, the chance of hashing queries and keys with high correlation into different buckets would increase. On the contrary, multi-hashtables could reduce that chance, but at a price of increasing the computational cost linearly. Therefore, in order to find a trade-off between two hyper-parameters, we conduct experiments with different combination. Specifically, we set $n_p = \{1, 2, 3, 4, 5, 6\}$ and $n_t = \{1, 2, 4, 8\}$, respectively. Results are shown in TABLE V.

TABLE V: Ablation study on number of hyper-planes n_p and hashtables n_t. Only results on the CASME II are shown.

n_t \ n_p	1	2	**3**	4	5	6
1	75.22	76.49	78.25	76.55	74.12	73.35
2	76.79	77.51	79.68	77.37	76.25	73.21
4	78.15	79.02	**81.06**	77.45	76.25	73.74
8	78.57	79.47	80.48	78.36	76.78	74.36

As shown in TABLE V, the performance of SSA peaks at $n_p = 3$ and then deteriorates as the increment of n_p. When the number of buckets and hashtables are both set to 1, the result is approximate to that of the full attention, indicating that exiguous hash buckets are insufficient to distinguish information effectively. Meanwhile, the performance presents a clear downward trend when $n_p \geq 3$ even with higher n_t. This is mainly because redundant sub-regions in the space are liable to hash highly correlated features into separate buckets. On the other hand, increasing the number of n_t does facilitate performance improvement but not play a decisive role.

3) Full Attention versus Sparse Attention: To testify the proposed sparse self-attention could effectively zero out magnification noise, we conducted a set of comparative experiments on three databases. Moreover, we compare the SSA with standard self-attention in terms of computational efficiency. The number of hash buckets is exponential to the number of hyper-planes, denoted as $w = 2^{n_p}$, thus the average size of a hush bucket is $\frac{L}{n_p}$. For the sake of simplicity, we only focus on the result of hashing one time, so the maximum time complexity of SSA is calculated as $w(\frac{L}{2n_p})^2$. While the full-attention requires more cost since computational complexity is quadratic to sequence length, especially when the length is large. Our results are demonstrated in TABLE VI.

TABLE VI: Effect of Sparsity Constraints.

	Time Complexity O(N)	Acc(%)		
		CASME II	SAMM	SMIC-HS
Full-Attention	bn_hL^2	78.45	75.57	75.02
SSA	$bn_hn_tw(\frac{L}{2w})^2$	**81.06**	**77.61**	**79.88**

As shown in TABLE VI, the SSA yields better performance on three databases compared with the standard one. By imposing sparsity constraints, the network could retain the ability to model dependencies in long sequences while zero out some defective magnification information. When operating full-attention, although irrelevant noise could be assigned smaller scores, the performance still suffers especially when the noise dominates the whole sequence, which indicates that knowing where to attend is more important than attending all.

IV. Conclusion

In this work, a magnification-robust network (MRN) is proposed to tackle the "magnification noise" problem which is scarcely noticed in MER. To be specific, in order to effectively extract motion features from a ME video, a substitutive sequence with more powerful ability to encode facial movements is reconstructed via a magnification technique. Subsequently, we impose sparsity constraints into standard self-attention using Locality Sensitive Hashing (LSH) to zero-out noise brought by magnification process. SSA globally attends the highly-correlated clues and disregards noise, resulting in a more robust operation. While preserving the ability of modeling long sequences, SSA also reduces the computational complexity. Extensive experiments implemented on three public databases, i.e., CASME II, SMIC-HS and SAMM demonstrate the effectiveness and superiority of our framework.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (NSFC) under the Grants U2003207, 61921004, and 61902064, in part by the Fundamental Research Funds for the Central Universities under

Grant 2242022k30036, and in part by the Zhishan Young Scholarship of Southeast University.

REFERENCES

[1] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013. 1

[2] Q. WU, X.-B. SHENG, and X.-L. FU, "Micro-expression and its applications," *Advances in Psychological Science*, vol. 18, no. 09, p. 1359, 2010. 1

[3] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," 2009. 1

[4] S. Zhao, H. Tao, Y. Zhang, T. Xu, K. Zhang, Z. Hao, and E. Chen, "A two-stage 3d cnn based learning method for spontaneous micro-expression recognition," *Neurocomputing*, vol. 448, pp. 276–289, 2021. 1, 5

[5] R. Zhi, H. Xu, M. Wan, and T. Li, "Combining 3d convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition," *IEICE Transactions on Information and Systems*, vol. 102, no. 5, pp. 1054–1064, 2019. 1, 5

[6] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proc. ACM MM*, 2016, pp. 382–386. 1

[7] W. Peng, X. Hong, Y. Xu, and G. Zhao, "A boost in revealing subtle facial expressions: A consolidated eulerian framework," in *Proc. FG*. IEEE, 2019, pp. 1–5. 1, 4

[8] A. C. Le Ngo, A. Johnston, R. C.-W. Phan, and J. See, "Micro-expression motion magnification: Global lagrangian vs. local eulerian approaches," in *Proc. FG*. IEEE, 2018, pp. 650–656. 1

[9] L. Lei, J. Li, T. Chen, and S. Li, "A novel graph-tcn with a graph structured representation for micro-expression recognition," in *Proc. ACM MM*, 2020, pp. 2237–2245. 1, 4, 5

[10] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 563–577, 2017. 1, 4

[11] M. Wei, W. Zheng, Y. Zong, X. Jiang, C. Lu, and J. Liu, "A novel micro-expression recognition approach using attention-based magnification-adaptive networks," in *Proc. ICASSP*. IEEE, 2022, pp. 2420–2424. 1

[12] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W. T. Freeman, and W. Matusik, "Learning-based video motion magnification," in *Proc. ECCV*, 2018, pp. 633–648. 2

[13] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-based video motion processing," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, pp. 1–10, 2013. 2

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778. 3, 4

[16] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014. 4

[17] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, 2016. 4

[18] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Proc. FG*. IEEE, 2013, pp. 1–6. 4

[19] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *Proc. ICCV*, 2017. 4

[20] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. ICONIP*. Springer, 2013, pp. 117–124. 4

[21] A. C. Le Ngo, R. C.-W. Phan, and J. See, "Spontaneous subtle expression recognition: Imbalanced databases and solutions," in *Proc. ACCV*. Springer, 2014, pp. 33–48. 4

[22] X. Huang and G. Zhao, "Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern," in *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*. IEEE, 2017, pp. 159–164. 4, 5

[23] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, and L. Zhao, "Recognizing spontaneous micro-expression using a three-stream convolutional neural network," *IEEE Access*, vol. 7, pp. 184 537–184 551, 2019. 4, 5

[24] L. Lei, T. Chen, S. Li, and J. Li, "Micro-expression recognition based on facial graph representation learning and facial action unit fusion," in *Proc. CVPR Workshop*, 2021, pp. 1571–1580. 4, 5

[25] H.-Q. Khor, J. See, S.-T. Liong, R. C. Phan, and W. Lin, "Dual-stream shallow networks for facial micro-expression recognition," in *Proc. ICIP*. IEEE, 2019, pp. 36–40. 5

[26] B. Xia, W. Wang, S. Wang, and E. Chen, "Learning from macro-expression: a micro-expression recognition framework," in *Proc. ACM MM*, 2020, pp. 2936–2944. 5

[27] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, and M. Xu, "Geme: Dual-stream multi-task gender-based micro-expression recognition," *Neurocomputing*, vol. 427, pp. 13–28, 2021. 5

[28] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *Proc. ACCV*. Springer, 2014, pp. 525–537. 4, 5

[29] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *Proc. FG*. IEEE, 2019, pp. 1–5. 5