and Regional Selective Mixup

[1]Yu-Chen Lin, [1]Yi-Jing Chen, [1,*]Chih-Chang Yu and [2]Hsu-Yung Cheng

[1]Department of Information and Computer Engineering, Chung Yuan Christian University, Taoyuan, Taiwan

[2]Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan

E-mail: {libra031017, s11127133, chihchangyu}@cycu.org.tw Tel: +886-3-2654701

E-mail: chengsy@ncu.edu.tw Tel: +886-3-4222681

Abstract— **Current studies on micro-expression recognition still achieve limited performance, typically yielding less than 50% accuracy. Although utilizing recent powerful models, such as Mamba, improves performance, they usually require enormous data for training. Therefore, this study proposes a regional selective mixup (RSMix) strategy that relies on facial Action Unit (AU) information to augment the dataset. The proposed approach systematically fuses facial images from micro-expression-generating regions with corresponding expressionless regions to synthesize new micro-expression samples. RSMix ensures that salient features from micro-expression-generating regions are preserved within these synthetic samples, facilitating the model's capacity to focus on essential visual cues for distinguishing different micro-expressions. We applied the VideoMamba model as the backbone to test the efficiency of RSMix and other augmentation methods. Experimental results on the CAS(ME)³ dataset demonstrate that with RSMix, the VideoMamba-Ti model achieves an accuracy improvement of Unweighted F1-score (UF1) of 0.0138 and Unweighted average recall (UAR) of 0.0048 compared to baseline configurations without augmentation. When flip augmentation is combined with the RSMix, the VideoMamba-S model achieves UF1 of 0.5040 and UAR of 0.4941, superior to the Mamba-based approach. These results prove that the RSMix approach effectively enhances data diversity, which also helps VideoMamba for the micro-expression recognition task.**

I. INTRODUCTION

Micro-expressions are essential to human emotional expression and have great potential across multiple domains. In criminal investigations, they can reveal emotional fluctuations in suspects, aiding in assessing the credibility of testimony. In education, micro-expressions provide subconscious emotional feedback from students, allowing instructors to evaluate teaching effectiveness and adapt instructional strategies accordingly. In business and marketing, analyzing micro-expressions exhibited by consumers during product launches can help forecast sales performance, identify customer preferences, and guide subsequent product development.

However, due to their extremely short duration (typically ranging from 0.04 to 0.2 seconds), subtle characteristics, and difficulty of conscious perception, micro-expression recognition poses far greater challenges than conventional facial expression analysis, often failing to achieve satisfactory accuracy when applied to micro-expression scenarios. Several factors contribute to this difficulty: first, the subtle nature of micro-expressions results in low inter-class separability, making them typically classifiable into only positive, negative, and neutral categories; second, the limited number of annotated micro-expression samples hinders large-scale model training; and finally, the rapid temporal dynamics of micro-expressions require more powerful models to capture transient motion patterns effectively.

Earlier approaches for micro-expression analysis primarily relied on time-series models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) [1], which are capable of capturing short-term dependencies but often suffer from vanishing gradient and long-term dependency issues. More recently, Transformer-based architectures have achieved breakthroughs across multiple domains by leveraging self-attention mechanisms; however, their quadratic computational complexity limits scalability when processing long video sequences.

Given these shortcomings, an efficient, accurate, and generalizable micro-expression analysis framework becomes increasingly critical. This study aims to address this challenge by employing the VideoMamba architecture. In addition, we proposed a novel augmentation method to satisfy the sample requirements for such models, thereby providing a scalable and effective solution for micro-expression recognition.

II. RELATED WORK

A. Sequence Modeling Challenges and Introduction to Mamba

Sequence modeling capability is crucial in tasks with high temporal sensitivity, such as micro-expression recognition. Traditional Recurrent Neural Networks (RNNs) often suffer from gradient vanishing and difficulty capturing long-term dependencies. On the other hand, Transformer-based architecture offers global modeling capabilities; however, their attention mechanism incurs a quadratic computational cost

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

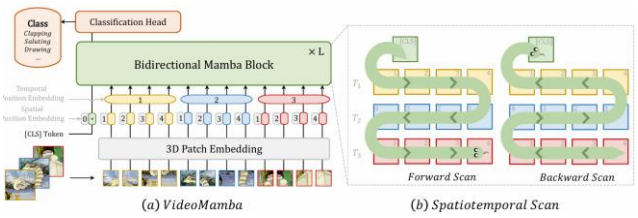

Fig. 1. Framework of VideoMamba [5].

proportional to sequence length, making them less efficient for processing long video sequences.

To address these limitations, recent developments in State Space Models (SSMs) [2] introduce an "input–state–output" update mechanism that effectively enhances long-term memory while enabling integration with modern neural networks through discretization and convolutional transformations. Building on this concept, the Mamba architecture [3] combines SSMs with convolutional layers and gated MLPs [4], residual connections, to form a modular, stackable design. Additionally, Mamba employs hardware-aware optimizations and selective scanning strategies to achieve efficient computation and fast inference, making it particularly well-suited for micro-expression recognition tasks involving subtle features and long sequence lengths.

B. Application of VideoMamba in Video-based Tasks

VideoMamba [5] is a Mamba-variant architecture designed for video sequence modeling. It divides the input video into spatiotemporal segments using 3D patch embedding, enabling simultaneous extraction of dynamic and spatial features. Its core component, the Bidirectional Mamba Block (B-Mamba) [6], models both forward and backward dependencies across temporal and spatial dimensions (see Fig. 1). Furthermore, VideoMamba supports multiple spatiotemporal scanning strategies, including Spatial-First, Temporal-First, and Spatiotemporal Bidirectional scanning, allowing the model to adapt to various task characteristics flexibly. These capabilities make VideoMamba highly suitable for micro-expression recognition, where facial movements are brief, subtle, and require high temporal precision.

C. Application of SSMs in Micro-Expression Analysis

Zou et al. [7] explored the applicability of utilizing the Mamba model for micro-expression analysis. They adopted a dual-path architecture that handles both spotting and recognition tasks. In the spotting path, the system detects whether a micro-expression occurs. Subsequently, it classifies the detected segments into corresponding micro-expression categories. Although the spotting accuracy is unsatisfactory, their system achieves an UF1 of 0.4754 and a UAR of 0.487 in the recognition task.

D. Summary

By leveraging state space modeling and hardware optimization strategies, Mamba achieves an optimal balance between computational efficiency and modeling accuracy, making it particularly well-suited for long sequence modeling. Building upon this foundation, VideoMamba incorporates bidirectional information flow and demonstrates superior performance in visual recognition tasks. Given its inherent capability to process sequential visual data efficiently, we consider VideoMamba exceptionally appropriate for the micro-expression recognition challenge addressed in this study. Therefore, our proposed methodology adopts the VideoMamba architecture, capitalizing on its strengths in spatiotemporal feature extraction to capture fine-grained micro-expression characteristics.

III. METHODOLOGY

A. Data Preprocessing I - Face Detection and Cropping

Systematic preprocessing is applied to raw images within the dataset to enhance the effectiveness and stability of micro-expression recognition. We employ dlib's facial detector [8] to identify facial landmarks and establish precise anatomical reference points. Facial images are aligned using the nasal bridge midline as the primary reference axis, followed by face cropping in each frame. The cropping dimensions are determined by calculating the vertical distance from the nasal bridge to the chin, ensuring consistent facial region coverage across all samples. An additional 10-pixel margin is extended outward from the calculated cropping boundary to preserve more comprehensive facial information. This alignment strategy enhances the spatial consistency of facial positions across temporal frames, providing more precise feature extraction and improved model convergence.

B. Data Preprocessing II - Regional Selective Mixup

Although conventional augmentation methods can easily increase the training samples, we found that VideoMamba receives little benefit from these augmentations in micro-expression recognition. This is because micro-expressions are subtle, meaning that most features in a facial image are not discriminative. To address this problem, this study proposed Regional Selective Mixup (RSMix). First, we established a mapping between the Action Unit (AU) [9] annotations provided by the dataset and the 68 facial landmarks defined by dlib (for this mapping, please refer to Appendix A). The facial area was then divided into 14×14 grids, and the landmarks contained within each grid were recorded. Given a micro-expression sequence called a source sequence, we extract the grids where the active AUs (i.e., AUs with micro-expressions) are located. The grids that contain active AUs in the source sequence are used to replace the corresponding grids in the target sequence. Each selected AU grid for replacement is expanded to a 3×3 AU region to mitigate possible

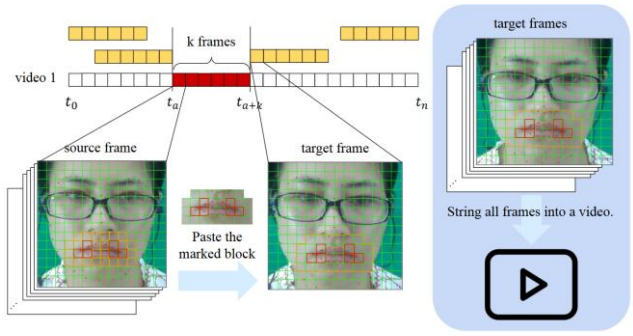

Fig. 2. Regional Selective Mixup (RSMix) method. The red grids represent the active AU grids. The orange area represents the AU region after expansion. We replace the AU regions in the source sequence (the red one) to the corresponding regions in the target sequence (the yellow one). For target sequence that is not sufficient long, we simply duplicate its last frame in order to increase its sequence length. Finally, we concatenate all frames to form a new video.

misalignments due to localization errors in the preprocessing step.

Next, we search within the same video for neutral segments (i.e., no micro-expressions), referred to as target sequences, from which continuous frames of the same length were extracted as the synthesis targets. The selected AU regions from the source sequence are then overlaid onto the corresponding positions of the target sequence to generate a new sequence. If no sufficiently long segment is found after traversing the entire video, the neutral segment following the source sequence that has not yet been used is selected instead. For any remaining frames, the last frame of the target sequence is duplicated to match the required length. The entire synthesis process is illustrated in Fig. 2.

Suppose there is a k-frame long micro-expression segment (source sequence) in a video that occurs between timestamps t_a and $t_{\{a+k\}}$, where a is the starting frame id. For each frame within this interval, the corresponding AU-active regions are marked. Specifically, the AU-active regions are identified via an AU-to-landmark mapping (see Appendix A). These regions, highlighted in red blocks, typically localize around expressive areas such as the mouth or eyebrows. To avoid missing micro-expressions due to incorrect face alignment, we slightly extend these regions from 1×1 to 3×3.

Next, a neutral segment (target sequence) with the same length k, the yellow blocks can be selected from another part of the same video. The target frames provide the facial background without any micro-expression. The AU-active regions extracted from the source frames are then copied to the corresponding grid locations in the target frames to synthesize new facial expressions while preserving the original neutral context.

IV. EXPERIMENTAL RESULTS

This study used the CAS(ME)³ dataset [10] as the experimental dataset. It contains approximately 80 hours of video, over 8 million frames, 1,109 manually annotated micro-expression samples, and 3,490 macro-expression samples across seven expression categories. After data cleaning, 3,887 samples were obtained, consisting of 639 micro-expression samples and 3,248 macro-expression samples. The duration of each micro-expression sequence ranges from 2 frames to 3455 frames. Subsequently, 11,246 samples were used for training after data augmentation. The dataset was split into training and testing sets following a 7:3 ratio. Importantly, the testing set exclusively comprises 273 original micro-expression samples without any data augmentation applied. All experiments are conducted on an RTX 4090 GPU. The batch size was set to 8, the learning rate to 0.0001, and the number of epochs to 1000. The training time with the VideoMamba-S model took approximately 11 days.

A. Comparison with State-of-the-art

To evaluate the effectiveness of the proposed method, we conducted a comparison with previous approaches. Considering that Zou et al. [7] conducted their experiments on predicting four emotion categories (positive, negative, surprise, and others), we followed their mapping strategy described in [7] to compare the results fairly. The seven emotion classes are mapped into the corresponding four categories: positive = {happy}, negative = {angry, disgust, fear, sad}, others = {others}, and surprise = {surprise}. The comparison results are summarized in Table 1.

During the training phase, we observed that using only micro-expression samples tends to overfit the model. Therefore, to enhance the model's generalization ability, we incorporated macro-expression samples into the training data to improve the stability and effectiveness of feature learning. Note that the test set contains only micro-expression samples. From Table 1, we can see that our method achieves the best results across all methods. This result demonstrates that the proposed method significantly benefits from the augmented samples, achieving

Table 1. Performance comparison between state-of-the-arts and our method

Methods	UF1	UAR
STSTNet[11]	0.3795	0.3792
RCN-A[12]	0.3928	0.3893
FeatRef[13]	0.3493	0.3413
AlexNet[14]	0.3001	0.2982
Zou et al.[7]	0.4754	0.4878
Ours (ME)	0.4765	0.4940
Ours (ME+MaE)	0.4903	**0.4988**
Ours (ME+MaE with VideoMamba-S)	**0.5040**	0.4941

Table 2. Comparison between different augmentation methods using VideoMamba-Ti (training with ME+MaE).

Augmentation settings			
Flip	RSMix	UF1	UAR
		0.4583	0.4744
✓		0.4745	0.4919
	✓	0.4721	0.4943
✓	✓	**0.4903**	**0.4988**

UF1 and UAR of 0.5040 and 0.4941, respectively. From Table 1, we can also conclude that including MaE samples leads to better performance than training with only ME samples, indicating that the proposed strategy mitigates overfitting and enriches training data diversity.

B. Ablation Study

A potential concern is whether the improvement in Table 1 should be attributed to the augmentation techniques or the model's capabilities. Therefore, we compared the performance of applying different augmentation methods to verify this concern. We experimented with the commonly used flip method, which horizontally flips the training videos to enlarge the dataset, as well as the proposed RSMix approach. Results are presented in Table 2.

As shown in Table 2, the number of augmented samples generated by the flip method is slightly higher than that of RSMix (3,887 vs. 3,472); however, the resulting performance is comparable. This proves that the RSMix strategy is effective. Moreover, we observed that combining flip and RSMix augmentations further improves the performance, reaching a UF1 of 0.4903 and a UAR of 0.4988.

C. Performance Validation of VideoMamba Models with Different Capacities

To further evaluate the capability of VideoMamba, we applied VideoMamba models with different capacities. The UF1 and UAR results are summarized in Table 3, and the Confusion matrix is shown in Fig. 3. By employing the larger-capacity VideoMamba-S model, the performance was further

Fig. 3. Confusion Matrix using VideoMamba-S.

Table 3. Comparison between VideoMamba-Ti and VideoMamba-S.

Model	UF1	UAR
VideoMamba-Ti	0.4903	**0.4988**
VideoMamba-S	**0.5040**	0.4941

improved, achieving a UF1 of 0.5040 and a comparable UAR of 0.4941, surpassing the results obtained using the VideoMamba-Ti model. Due to the hardware limitations, we cannot test other larger VideoMamba models; however, it can be concluded that with the proposed RSMix method, either the VideoMamba-Ti or VideoMamba-S model is sufficient for this micro-expression recognition task.

V. CONCLUSION

In this study, we proposed a proprietary mixup method for the micro-expression recognition task. By leveraging the VideoMamba architecture, the proposed method effectively addresses the challenges of subtle facial motion detection in micro-expression tasks. Experimental results on the CAS(ME)³ dataset demonstrate that our method outperforms existing state-of-the-art approaches, achieving UF1 and UAR scores of 0.5040 and 0.4941, respectively.

Furthermore, the ablation study confirms the effectiveness of the regional selective mixup (RSMix) technique, which performs similarly to conventional augmentation methods by preserving discriminative facial regions and enhancing the model's robustness under complex expression variations. When combining both augmentations, we obtain the best performance. Although a larger model performs better in this task, we did not find a significant difference between VideoMamba-Ti and VideoMamba-S. We plan to apply the RSMix method to other models, such as Mamba, in the future to justify its generalizability and effectiveness.

ACKNOWLEDGEMENT

The National Science and Technology Council, Taiwan, funded this research with grant number NSTC 114-2221-E-033-015-MY3.

REFERENCES

[1] M. Verburg and V. Menkovski, "Micro-expression detection in long videos using optical flow and recurrent neural networks," in Proc. IEEE/CVF Int. Conf. Computer Vision Workshops (ICCVW), Seoul, South Korea, 2019, pp. 2648–2651.

[2] A. Gu, *Modeling Sequences with Structured State Spaces*, Ph.D. dissertation, Dept. Computer Science, Stanford Univ., Stanford, CA, USA, 2023.

[3] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint arXiv:2312.00752, 2023.

[4] H. Liu, Z. Dai, D. R. So, and Q. V. Le, "Pay attention to MLPs," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 9204–9215, 2021

[5] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "VideoMamba: State space model for efficient video understanding," *arXiv preprint* arXiv:2403.06977, 2024.

[6] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient visual representation learning with bidirectional state space model," arXiv preprint arXiv:2401.09417, 2024.

[7] B. Zou, Z. Guo, W. Qin, X. Li, K. Wang, and H. Ma, "Synergistic spotting and recognition of micro-expression via temporal state transition," *arXiv preprint* arXiv:2409.09707, 2024.

[8] D. E. King, "Dlib-ml: A machine learning toolkit," J. Mach. Learn. Res., vol. 10, pp. 1755–1758, 2009.

[9] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[10] Li, J., Dong, Z., Lu, S., Wang, S. J.*, Yan, W. J., Ma, Y., Liu, Y., Huang, C., & Fu, X.* (2022). 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. doi: 10.1109/TPAMI.2022.3174895.

[11] S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition," in *Proc. 14th IEEE Int. Conf. Automat. Face Gesture Recognition (FG)*, 2019, pp. 1–5.

[12] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," IEEE Trans. Image Processing, vol. 29, pp. 8590–8605, 2020.

[13] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, "Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition," Pattern Recognition, vol. 122, p. 108275, 2022.

[14] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, and X. Fu, “Cas(me)3: A third generation facial spontaneous micro expression database with depth information and high ecological validity,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 3, pp. 2782–2800, 2023.

Appendix A

expression	AU	dlib point index	expression	AU	dlib point index
Inner Brow Raiser	1	[21, 22, 23, 24]	Left Dimpler	L14	[63-67, 55]
Left Inner Brow Raiser	L1	[21, 22]	Right Dimpler	R14	[49, 61-63, 67, 68]
Right Inner Brow Raiser	R1	[23, 24]	Lip Corner Depressor	15	[49, 59, 61, 65]
Outer Brow Raiser	2	[18, 19, 26, 27]	Lower Lip Depressor	16	[56-60, 66-68]
Left Outer Brow Raiser	L2	[18, 19]	Chin Raiser	17	[57-59]
Right Outer Brow Raiser	R2	[26, 27]	Lip Puckerer	18	[49, 50, 61, 60, 54-56, 65]
Brow Lowerer	4	[21, 22, 23, 24]	Lip stretcher	20	[49-60, 8, 9, 10]
Upper Lid Raiser	5	[38, 39, 44, 45]	Left Lip stretcher	L20	[52-58, 9, 10]
Cheek Raiser	6	[32, 36-49, 55]	Right Lip stretcher	R20	[49-52, 58-60, 8, 9]
Left Cheek Raiser	L6	[36, 43-48, 55]	neck tightener	21	[7-11]
Right Cheek Raiser	R6	[32, 49, 37-42]	Lip Funneler	22	[49-68]
Lid Tightener	7	[37-48]	Top Lip Funneler	T22	[50-54]
Left Lid Tightener	L7	[43-48]	Bottom Lip Funneler	B22	[56-60]
Nose Wrinkler	9	[28-36]	Lip Tightener	23	[49-68]
Left Nose Wrinkler	L9	[28-31, 34-36]	Top Lip Tightener	T23	[50-54]
Right Nose Wrinkler	R9	[28-34]	Bottom Lip Tightener	B23	[56-60]
Upper Lip Raiser	10	[32, 36, 49-55, 61-65]	Lip Pressor	24	[50-68]
Left Upper Lip Raiser	L10	[36, 52-55, 63-65]	Top Lip Pressor	T24	[50-54, 62-64]
Right Upper Lip Raiser	R10	[32, 49-55, 61-63]	Lips part	25	[6-12, 49-68]
Left Nasolabial Deepener	L11	[36, 54, 55, 65]	Jaw Drop	26	[55-61, 65-68, 8-10]
Right Nasolabial Deepener	R11	[49, 50, 61, 32]	Mouth Stretch	27	[49-68, 7-11]
Lip Corner Puller	12	[49-51, 59-61, 53-57, 65]	Lip Suck	28	[50-54, 56-60]
Left Lip Corner Puller	L12	[53-57, 65]	Top Lip Suck	T28	[50-54]
Right Lip Corner Puller	R12	[49-51, 59-61]	Bottom Lip Suck	B28	[56-60]
Dimpler	14	[49, 55, 61-68]	nostril dilator	38	[32, 33, 35, 36]