DIANet: A Phase-Aware Dual-Stream Network for Micro-Expression Recognition via Dynamic Images

Vu Tram Anh Khuong, Luu Tu Nguyen, Thi Bich Phuong Man, Thanh Ha Le and Thi Duyen Ngo*

Faculty of Information Technology, VNU University of Engineering and Technology, Hanoi, Vietnam

Abstract

Micro-expressions are brief, involuntary facial movements that typically last less than half a second and often reveal genuine emotions. Accurately recognizing these subtle expressions is critical for applications in psychology, security, and behavioral analysis. However, micro-expression recognition (MER) remains a challenging task due to the subtle and transient nature of facial cues and the limited availability of annotated data. While dynamic image (DI) representations have been introduced to summarize temporal motion into a single frame, conventional DI-based methods often overlook the distinct characteristics of different temporal phases within a micro-expression. To address this issue, this paper proposes a novel dual-stream framework, DIANet, which leverages phase-aware dynamic images - one encoding the onset-to-apex phase and the other capturing the apex-to-offset phase. Each stream is processed by a dedicated convolutional neural network, and a cross-attention fusion module is employed to adaptively integrate features from both streams based on their contextual relevance. Extensive experiments conducted on three benchmark MER datasets (CASME-II, SAMM, and MMEW) demonstrate that the proposed method consistently outperforms conventional single-phase DI-based approaches. The results highlight the importance of modeling temporal phase information explicitly and suggest a promising direction for advancing MER.

Keywords

Micro-expression, micro-expression recognition, dynamic image, dual-stream network, deep learning

1. Introduction

Micro-expressions (MEs) are brief, involuntary facial movements that typically last less than 0.5 seconds and reveal genuine emotions that an individual may attempt to conceal. Unlike macro-expressions, MEs are subtle and transient, often involving only localized muscle activations. These characteristics make MEs difficult to detect by human observers and pose significant challenges for automatic micro-expression recognition (MER) systems. Nevertheless, accurate MER has substantial value in various real-world applications such as lie detection, security screening, psychotherapy, and human–computer interaction [1, 2, 3].

Early research in MER primarily relied on handcrafted features designed to capture spatial texture (e.g., Local Binary Patterns from Three Orthogonal Planes, LBP-TOP [4]) or optical strain [5], combined with temporal analysis through optical flow [6]. While effective in controlled settings, these methods often struggle with robustness under cross-subject variation or real-world noise and require extensive feature engineering. In recent years, the emergence of deep learning and motion-based representations has led to more compact and learnable frameworks for MER. Among them, the dynamic image (DI) representation has gained popularity. Originally proposed in action recognition, DI summarizes the temporal evolution of motion into a single image using rank pooling [7]. In the context of MER, DI-based models such as LEARNet [8] demonstrate that DI can effectively encode subtle motion signals in a suitable form for convolutional neural networks. Extensions of this idea include Active Image [9], which highlights salient motion regions, and Affective Motion Imaging [10], which emphasizes emotion-relevant motion features. These methods show that DI offers an efficient way to encode facial motion

4DMR@IJCAI25: International IJCAI Workshop on 1st Challenge and Workshop for 4D Micro-Expression Recognition for Mind Reading, August 29, 2025, Guangzhou, China

*Corresponding author

duyennt@vnu.edu.vn (T. D. Ngo)

into a single frame.

However, existing DI-based approaches typically treat the entire sequence as a whole, ignoring the distinct temporal phases of MEs (i.e., onset-apex and apex-offset). This holistic modeling overlooks the asymmetric rise-and-fall dynamics of facial motion. Incorporating phase-specific representations could better capture these patterns, yet remains underexplored in the literature. Phase-aware modeling has been explored in a few recent works. For example, Liong *et al.* [11] emphasized the onset–apex segment by computing optical flow between two key frames, while Zhang *et al.* [12] proposed apex-guided representation learning. However, these approaches typically rely on hand-crafted segmentation or optical flow and do not fully integrate phase modeling into a learnable deep framework. Moreover, none of them explicitly leverage DI to encode motion separately within each phase.

To address this gap, this paper proposes DIANet, the first dual-stream MER framework that uses phase-specific dynamic images to separately model the onset–apex and apex–offset intervals. By processing these phases in parallel, the model captures both the rising and falling motion patterns of MEs - subtle dynamics often overlooked in single-stream approaches. This phase-aware design enhances representation learning and leads to more accurate recognition of subtle and transient expressions. To the best of our knowledge, this is the first work to explicitly integrate phase-specific DI representations into an end-to-end framework for micro-expression recognition.

The remainder of the paper is organized as follows. Section 2 describes the proposed DIANet architecture in detail. Next, section 3 describes the experimental setup and reports results on CASME-II, SAMM, and MMEW. Finally, section 4 concludes the paper and discusses future directions.

2. Proposed method

This paper introduces a dual-stream dynamic imaging (DI) framework to address the limitations of conventional DI-based MER. By decomposing facial motion into two distinct temporal phases (i.e., onset-apex or apex-offset), this paper enhances motion representation while suppressing noise. The overall workflow is illustrated in Fig. 1.

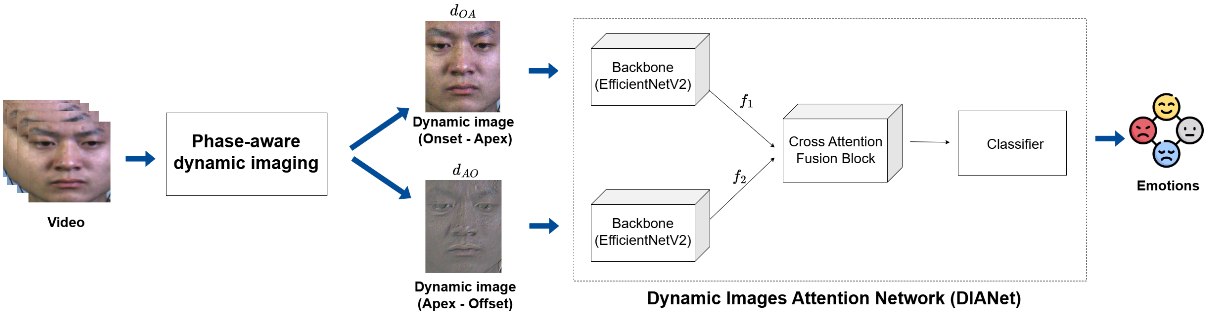
Video
Phase-aware dynamic imaging
d_{OA}
Dynamic image (Onset - Apex)
d_{AO}
Dynamic image (Apex - Offset)
Backbone (EfficientNetV2)
Backbone (EfficientNetV2)
f_1
f_2
Cross Attention Fusion Block
Classifier
Dynamic Images Attention Network (DIANet)
Emotions

Figure 1: The workflow of our proposed dual-stream dynamic imaging for micro-expression recognition

2.1. Phase-aware dynamic imaging

Dynamic image is a technique that summarizes a video sequence into a single image, encoding both spatial and temporal information into a static representation. It is commonly constructed using Approximate Rank Pooling (ARP) [7], where each frame in a sequence is projected into a feature space and combined with a weight proportional to its temporal position. Formally, given a sequence of T consecutive frames $F_1, F_2, \ldots, F_T$ and corresponding features $\psi(F_t)$, the dynamic image d^* is computed as follows:

$$d^* = \sum_{t=1}^{T} \alpha_t \psi(F_t), \quad \text{with} \quad \alpha_t = 2t - T - 1. \tag{1}$$

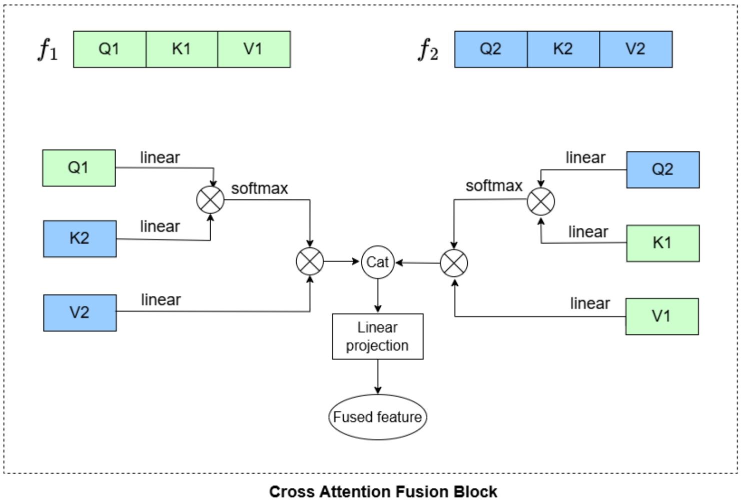

Figure 2: The cross-attention fusion block in DIANet. Inputs $f1$ and $f2$ are feature vectors obtained from the outputs of two EfficientNetV2 backbones corresponding to the onset–apex and apex–offset streams, respectively. Each is projected into query, key, and value spaces, and mutual attention is computed to produce a fused representation

where $\psi(F_t) = F_t$, since raw pixel values are directly employed as frame-level features in each color channel.

The standard ARP assigns higher weights to later frames, assuming motion accumulates over time - a valid assumption for full-length actions, but not always in the case of micro-expressions. Micro-expressions typically follow a bell-shaped intensity curve, with motion rising to a peak (apex) and then declining. Applying ARP over the full sequence skews the resulting dynamic image toward the offset phase, underrepresenting early-phase motion where key emotional cues often occur.

To overcome this limitation, this paper proposes a dual-phase dynamic imaging strategy that separately models the onset–apex and apex–offset phases using a tailored ARP formulation. Each segment is used to construct a dedicated phase-specific DI.

- **DI-Onset:** Captures the rising motion from onset to apex using the standard ARP with increasing weights, as shown in Eq.2. This emphasizes frames closer to the apex, where expressive intensity typically peaks.

$$d_{OA} = \sum_{t=1}^{T} \alpha_t \psi(F_t), \quad \text{with} \quad \alpha_t = 2t - T - 1. \tag{2}$$

  where t indexes the frames in the onset-apex segment of length T.
- **DI-Offset:** Encodes the declining motion from apex to offset using reversed ARP with decreasing weights defined as follows:

$$d_{AO} = \sum_{t=1}^{T} \alpha'_t \psi(F_t), \quad \text{with} \quad \alpha'_t = T + 1 - 2t. \tag{3}$$

  Here, t indexes the frames in the apex–offset segment of length T, and higher weights are assigned to frames near the apex to highlight the transition back to a neutral state.

This dual-phase formulation has two key advantages. First, it preserves the distinct motion characteristics of each phase, allowing the model to learn more fine-grained, phase-specific features. Second, it mitigates the temporal bias inherent in single-DI representations by balancing the representation across both phases. In summary, by applying ARP in a phase-aware manner, our dual-phase dynamic imaging method captures both the escalation and de-escalation of facial expressions, providing a more comprehensive and balanced motion representation for MER.

2.2. Dynamic Images Attention Network (DIANet)

To fully exploit the asymmetric and complementary motion patterns present in micro-expression sequences, this paper proposes a novel dual-stream classification architecture called Dynamic Images Attention Network (DIANet) (as illustrated in Fig 1). Unlike conventional approaches that rely on a single dynamic image over the entire sequence, DIANet introduces a phase-aware formulation by explicitly modeling the two key motion intervals: onset–apex (rising phase) and apex–offset (fading phase). Each phase is encoded into a dynamic image, referred to as DI-Onset and DI-Offset, which are processed independently in parallel streams.

Each stream in DIANet employs a shared backbone based on a modified EfficientNetV2[13] architecture. The classification head is removed to retain only high-level feature extraction, ensuring consistency and parameter efficiency across both branches. This backbone was selected based on its superior performance in the ablation study (see section 3.6), offering a strong balance between compactness and representational power - crucial for handling subtle motion in limited micro-expression data.

For each input stream, the backbone outputs a 512-dimensional feature vector: $f1$ for the onset–apex phase and $f2$ for the apex–offset phase. To bridge the semantic gap and promote information exchange between the two phases, DIANet introduces a novel cross-attention fusion module, as illustrated in Fig. 2. This module enables bidirectional interaction between DI-Onset and DI-Offset features through learnable query-key-value projections. Each feature vector attends to the other, allowing the model to focus on salient motion cues across both phases. The resulting attention-enhanced features are concatenated and passed through a projection layer to form a unified representation.

The fused feature is then passed through a lightweight multilayer perceptron (MLP) classifier, consisting of two fully connected layers with ReLU activation and dropout regularization, to predict the final emotion class. This architectural design balances expressiveness and efficiency, making it suitable for micro-expression recognition under limited data conditions.

2.3. Loss function

To enhance representational coherence between the two temporal phases, this paper introduces a consistency regularization term that encourages alignment between the features learned from the onset–apex and apex–offset streams. Specifically, a cosine-based consistency loss is applied to penalize discrepancies between the corresponding feature vectors. To the best of our knowledge, this is the first application of cross-phase regularization in the context of micro-expression recognition using dynamic image representations.

The consistency loss is defined as the average cosine dissimilarity across a mini-batch:

$$\mathcal{L}_{\text{cons}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \cos\left(f_1^{(i)}, f_2^{(i)}\right)\right), \tag{4}$$

where $f_1^{(i)}$ and $f_2^{(i)}$ denote the feature vectors from the DI-Onset and DI-Offset streams for the i-th sample, and N is the batch size.

The total training objective combines the standard cross-entropy classification loss $\mathcal{L}_{\text{CE}}$ with the proposed consistency loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{cons}}, \tag{5}$$

where λ is a regularization coefficient that balances classification accuracy and cross-phase feature alignment. This joint objective promotes both discriminative learning and temporal consistency, leading to more robust and generalizable micro-expression recognition.

3. Experiments and Results

3.1. Datasets

To evaluate the effectiveness of the proposed method, experiments are conducted on three publicly available benchmark datasets commonly used in micro-expression recognition:

- CASME-II[14]: This dataset consists of 255 spontaneous micro-expression samples from 26 subjects. Recordings were captured under laboratory conditions at a high frame rate of 200 fps, allowing precise temporal localization of facial movements. All samples were manually annotated with emotion labels and apex frames.
- SAMM[15]: The SAMM dataset contains 159 high-resolution micro-expression samples from 32 participants. Similar to CASME-II, the recordings were captured at 200 fps in a controlled environment. Emotion annotations were provided by expert coders based on FACS (Facial Action Coding System) criteria.
- MMEW [16]: The MMEW dataset includes 300 spontaneous micro-expression samples collected in more naturalistic settings. Videos were recorded at 90 fps and annotated with six emotion categories. MMEW introduces more variation in head pose and lighting, making it a challenging benchmark for generalization.

To ensure balanced evaluation, emotion classes with fewer than 10 samples are excluded. After filtering, CASME-II is reduced to five categories (i.e., disgust, happiness, surprise, repression, others), SAMM to five categories (i.e., happiness, surprise, anger, contempt, others) and MMEW to six categories (i.e., happiness, sadness, fear, others, surprise, disgust). To improve generalization and address data scarcity, standard data augmentation techniques are employed during training. These include horizontal flipping and random in-plane rotations within the range of ±10 degrees. All video sequences are temporally normalized and resized to a fixed spatial resolution before being processed into dynamic image representations.

3.2. Evaluation Protocol

The evaluation follows the Leave-One-Subject-Out (LOSO) cross-validation protocol, which is commonly used in the MER field to assess subject-independent generalization. In each iteration, the samples from one subject are reserved for testing, while the remaining data are used for training.

3.3. Evaluation Metric

Model performance is evaluated using overall Accuracy (Acc), which reflects the proportion of correctly classified samples over the total number of samples, as shown in Eq 6:

$$\text{Acc} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}} \tag{6}$$

3.4. Implementation Details

All experiments are conducted using PyTorch. Input dynamic images are resized to 224×224 and normalized. Models are trained for 50 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 32. Early stopping is applied based on validation loss. The training objective includes the consistency loss described in the section 2.3 to encourage alignment between phase-specific feature representations.

3.5. Results

Table 1 presents a comparative analysis of the proposed method (DIANet) against recent state-of-the-art approaches that utilize dynamic images or DI-inspired representations. The evaluation is conducted across three widely used micro-expression datasets: CASME-II, SAMM, and MMEW.

Our proposed method DIANet achieves the highest accuracy on two of the three datasets: 68.89% on SAMM and 64.24% on MMEW, demonstrating its robustness across both high-resolution and naturalistic micro-expression scenarios. These results highlight the advantage of phase-aware modeling in capturing localized motion patterns that are often overlooked in single-phase or holistic DI-based approaches. On CASME-II dataset, DIANet achieves a competitive accuracy of 70.00%, demonstrating a performance comparable to recent advanced methods such as GEME [17] (75.20%) and FDCN [18] (73.09%). GEME leverages gender information as an auxiliary input to capture identity-dependent dynamics, but at the cost of increased model complexity and reliance on demographic metadata, which may raise practical and ethical concerns. FDCN combines dynamic images with optical flow to enrich motion features, yet its multi-stream architecture introduces additional computational overhead and modality alignment challenges.implementation overhead. In contrast, DIANet relies solely on phase-specific dynamic images derived from the original video sequence, requiring no auxiliary labels or multi-modal fusion. By explicitly separating motion into the onset-apex and apex-offset phases, DIANet learns more fine-grained, phase-sensitive features that are often lost in conventional DI-based approaches. This design not only simplifies the model architecture but also improves interpretability and portability across datasets.

Table 1
Comparison with DI-based state-of-the-art MER methods (accuracy %)

Year	Input	Method	CASME-II	SAMM	MMEW
2019	DI	LEARNet* [19]	56.67	59.26	48.96
2020	Active image	OrigiNet [9]	62.09	34.89	-
2021	Affective-motion image	AffectiveNet [10]	46.49	58.12	-
2021	DI + gender	GEME [17]	**75.20**	55.88	-
2023	DI + Optical flow	FDCN [18]	73.09	58.07	-
-	**DI-Onset + DI-Offset**	**DIANet (ours)**	70.00	**68.89**	**64.24**

* This results are from our re-implementation with the model provided by the author

Compared to LEARNet[19], a widely used DI-based baseline, DIANet yields substantial improvements: +11.22% on CASME-II, +9.63% on SAMM, and +15.28% on MMEW. These gains confirm that explicitly modeling phase transitions enhances the model's ability to capture expressive facial motion. Other DI-motivated methods, such as OrigiNet[9] and AffectiveNet [10], also fall behind in generalization, with particularly low performance on SAMM (e.g., 34.89%), suggesting that simple enhancements over DI are insufficient for cross-dataset robustness.

In summary, the consistent performance of DIANet across diverse datasets without reliance on handcrafted rules, demographic information, or multi-modal fusion, demonstrates its effectiveness and generalization. The results underscore the importance of phase-aware modeling as a lightweight yet powerful strategy for advancing micro-expression recognition in both constrained and real-world environments.

3.6. Ablation study

3.6.1. Backbone selection for DIANet

The performance of DIANet is influenced not only by its dual-stream, phase-aware design but also by the feature extraction backbone used in each stream. To investigate the impact of backbone architecture, we evaluate four representative convolutional backbones: ResNet18 [20], EfficientNetV2 [13], ConvNeXt [21], and MobileViT [22]. These models are selected based on their widespread use, architectural diversity, and relevance to tasks involving subtle motion recognition under data constraints. Table 2 illustrates the performance of DIANet using different backbone architectures across three benchmark datasets.

Table 2
Comparison of backbone performance (accuracy %) on DIANet across three datasets

Backbone	CASME-II	SAMM	MMEW
Resnet18 [20]	61.67	64.44	**64.94**
EfficientNetV2 [13]	**70.00**	**68.89**	64.24
ConvNeXt [21]	59.17	58.52	60.45
MobileViT [22]	62.08	60.74	62.08

All backbones are integrated into the same DIANet framework and trained under identical experimental settings. Each model processes a pair of dynamic images corresponding to the onset–apex and apex–offset phases, and outputs are fused via the same cross-attention mechanism. The goal is to isolate the effect of the backbone itself on performance across three micro-expression datasets.

- ResNet18 serves as a strong and widely adopted baseline. It provides a good trade-off between depth and computational cost, making it suitable for MER tasks with limited training data. On MMEW, it surprisingly achieves the highest accuracy (64.94%), possibly due to its robustness to noisy and unconstrained environments.
- EfficientNetV2 is a recent lightweight architecture that incorporates compound scaling and advanced training optimizations. It achieves the best performance on both CASME-II (70.00%) and SAMM (68.89%), highlighting its superior capacity to extract discriminative features from phase-specific DIs. Its consistent results across datasets suggest that it offers a strong balance between efficiency and representational power.
- ConvNeXt adapts design elements from vision transformers into a CNN framework. Although it has shown promising results in large-scale classification tasks, its relatively lower performance here (e.g., 59.17% on CASME-II) indicates that deeper and heavier models may not always generalize well in low-data, fine-grained settings such as MER.
- MobileViT combines the strengths of CNNs and transformers in a compact form, making it attractive for lightweight applications. However, its performance lags behind other models across all datasets, possibly due to underfitting or difficulties in learning temporal-localized features from DIs.

Overall, EfficientNetV2 provides the best trade-off between accuracy and efficiency. Its strong performance on both controlled (CASME-II, SAMM) and in-the-wild (MMEW) datasets suggests it is better suited for capturing the nuanced spatiotemporal features present in dynamic images of micro-expressions. In contrast, larger or transformer-based backbones like ConvNeXt and MobileViT may require more data or different training strategies to be effective in this domain. These findings support the use of modern, lightweight CNNs as backbones in MER, especially when paired with task-specific representations like phase-separated DIs.

3.7. Attention block selection for DIANet

To assess the effectiveness of our proposed Cross Attention Fusion Block, we compare it with a simplified variant named Simple Attention Block, which uses a basic dot-product attention and residual fusion.

As shown in Table 3, replacing the Cross Attention Fusion Block with Simple Attention Block results in a notable performance drop across all datasets: from 70.00% to 66.42% on CASME-II, from 68.89% to 63.16% on SAMM, and from 64.24% to 50.08% on MMEW. The performance degradation is especially pronounced on MMEW, which contains greater variability in pose and lighting. This suggests that simple attention fails to capture the nuanced interactions between onset–apex and apex–offset streams under challenging conditions. In contrast, the Cross Attention Fusion Block allows each stream to dynamically attend to salient features in the other via learnable query-key-value projections, promoting rich bidirectional interaction and context-aware fusion. This leads to more discriminative and phase-sensitive feature representations, which are critical in recognizing subtle micro-expressions.

These findings demonstrate that while simple attention mechanisms offer computational simplicity, they are insufficient for modeling the asymmetric and complementary dynamics of micro-expression phases. The proposed cross-attention strategy, albeit slightly more complex, substantially improves recognition performance and justifies its integration into DIANet.

Table 3
Comparison of attention block performance (accuracy %) on DIANet across three datasets

Attention	CASME-II	SAMM	MMEW
Simple Attention Block	66.42	63.16	50.08
Cross Attention Fusion Block (ours)	**70.00**	**68.89**	**64.24**

3.7.1. Effect of phase-wise dynamic images

To assess the contribution of phase-specific motion modeling, we conduct an ablation study comparing different input configurations within the DIANet framework, as summarized in Table 4. Specifically, we evaluate three types of inputs: (1) standard dynamic images generated from the entire sequence (as used in LEARNet [19]), (2) dynamic images computed separately for each temporal phase (DI-Onset and DI-Offset), and (3) a dual-stream setting where both DI-Onset and DI-Offset are processed in parallel.

Table 4
Performance comparison of phase-wise dynamic image on three datasets (accuracy %)

	Input	Method	CASME-II	SAMM	MMEW
1-stream	Dynamic image	LearNet* [19]	56.67	59.26	48.96
	DI-Onset	LearNet* [19]	**60.00**	54.07	**52.78**
	DI-Offset	LearNet* [19]	57.08	**61.48**	50.35
2-stream	Dynamic image	DIANet	58.33	59.26	51.04
	DI-Onset + DI-Offset	**DIANet**	**70.00**	**68.89**	**64.24**

* This results are from our re-implementation with the model provided by the author

In the single-stream setting, using phase-specific DIs (either DI-Onset or DI-Offset) yields better performance than using a holistic dynamic image. For instance, on CASME-II, DI-Onset improves accuracy from 56.67% (standard DI) to 60.00%, while on MMEW, it raises accuracy from 48.96% to 52.78%. Similarly, DI-Offset provides the best single-phase result on SAMM (61.48%), outperforming both the standard DI and DI-Onset. These improvements indicate that motion restricted to a specific phase, either the rising (onset–apex) or falling (apex–offset) interval, contains more discriminative features for recognizing micro-expressions than aggregating motion over the entire sequence. This supports our initial hypothesis that the dynamics of MEs follow an asymmetric temporal pattern, and treating the sequence holistically may dilute critical motion cues.

In the dual-stream configuration, DIANet processes both DI-Onset and DI-Offset simultaneously, with a cross-attention fusion module learning to integrate complementary information from the two phases. This setup achieves the best performance across all datasets: 70.00% on CASME-II, 68.89% on SAMM, and 64.24% on MMEW. Notably, the dual-stream DIANet outperforms all single-stream variants, including both phase-specific and standard DI inputs.

These results provide strong empirical evidence for the effectiveness of phase-aware modeling in MER. The dual-stream input not only leverages the unique temporal properties of each phase but also enables the model to learn richer and more balanced representations of facial motion. The significant performance gap between the single-phase and dual-phase settings further validates the design choice of using DI-Onset and DI-Offset as complementary inputs within the DIANet architecture.

4. Conclusion

This paper has proposed DIANet, the first dual-stream micro-expression recognition framework that leverages phase-aware dynamic images to model facial motion with greater temporal precision. By explicitly separating the onset–apex and apex–offset phases, the proposed approach captures complementary motion dynamics that are often overlooked in holistic or single-phase representations. The two streams are integrated through a cross-attention mechanism and guided by a consistency objective, enabling the network to learn more discriminative and temporally aligned features. Extensive experiments conducted on three benchmark datasets (i.e., CASME-II, SAMM, and MMEW) demonstrate the effectiveness of the proposed method. DIANet achieves 70.00%, 68.89%, and 64.24% accuracy on these datasets, respectively, outperforming or matching state-of-the-art approaches without requiring auxiliary modalities or complex fusion schemes. These results underscore the importance of phase-aware modeling in micro-expression recognition. The proposed framework offers a lightweight yet powerful solution that generalizes well across both controlled and unconstrained settings. Future work may explore extending phase-aware modeling to other motion representations and incorporating temporal uncertainty in apex estimation.

Acknowledgement

This work has been supported by HORIZON-MSCA-SE-2022 PhySU-Net 241 project ACMod (grant 101130271).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Bhushan, A. Dhall, R. Goecke, Spotting facial micro-expressions in long videos using spatio-temporal recurrent convolutional networks, in: International Conference on Automatic Face and Gesture Recognition (FG), 2015, pp. 1–6.

[2] W. Yan, Q. Wu, Z. Liang, G. Chen, X. Fu, Fast facial expression recognition using local binary patterns, directional map features and a boosted classifier, in: IEEE Transactions on Systems, Man, and Cybernetics, 2013.

[3] S. Polikovsky, Y. Kameda, Y. Ohta, Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor, in: 3rd International Conference on Crime Detection and Prevention (ICDP), 2009, pp. 1–6.

[4] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikainen, Spontaneous micro-expression recognition using lbp-top, IEEE Transactions on Affective Computing 4 (2013) 146–158.

[5] S. Liong, J. See, K. Wong, R. Phan, Optical strain based recognition of subtle emotions, Neurocomputing 175 (2016) 830–838.

[6] T. Pfister, X. Li, G. Zhao, M. Pietikäinen, Recognising spontaneous facial micro-expressions, in: International Conference on Computer Vision (ICCV), 2011, pp. 1449–1456.

[7] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, Dynamic image networks for action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 3034–3042.

[8] X. Li, Q. Wu, G. Zhao, X. Hong, Deep learning for micro-expression recognition: A survey, IEEE Transactions on Affective Computing (2022).

[9] G. Verma, A. Dhall, A non-ordinal approach to micro-expression recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2436–2445.

[10] G. Verma, A. Dhall, Affectivenet: Affective-motion feature learning for micro-expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 2644–2651.

[11] S. Liong, J. See, K. Wong, R. Phan, N. Cheung, Less is more: Micro-expression recognition with apex frame, Image and Vision Computing 59 (2017) 29–39.

[12] K. Zhang, F. Wang, G. Liu, H. Li, Y. Fu, Facial micro-expression recognition based on apex frame-guided cross-attention network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14361–14370.

[13] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR, 2021.

[14] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, Casme ii: An improved spontaneous micro-expression database and the baseline evaluation, PloS one 9 (2014) e86041.

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

[16] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, Y.-J. Liu, Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, IEEE transactions on pattern analysis and machine intelligence 44 (2021) 5826–5846.

[17] X. Nie, M. A. Takalkar, M. Duan, H. Zhang, M. Xu, Geme: Dual-stream multi-task gender-based micro-expression recognition, Neurocomputing 427 (2021) 13–28.

[18] J. Tang, L. Li, M. Tang, J. Xie, A novel micro-expression recognition algorithm using dual-stream combining optical flow and dynamic image convolutional neural networks, Signal, Image and Video Processing 17 (2023) 769–776.

[19] M. Verma, S. K. Vipparthi, G. Singh, S. Murala, Learnet: Dynamic imaging network for micro expression recognition, IEEE Transactions on Image Processing 29 (2019) 1618–1627.

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

[21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986.

[22] S. Mehta, M. Rastegari, Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer, in: International Conference on Learning Representations (ICLR), 2022.