POSTER: A Dual-Stream Transformer-Based Architecture for Overcoming Intrinsic Challenges in Facial Expression Recognition

Rawan Ahmed Ameen Hassan(SL22215001), Dong Lanfang

Abstract—This project addresses the intrinsic challenges in Facial Expression Recognition (FER), including high inter-class similarity, pronounced intra-class variability, and sensitivity to facial scale variations. The proposed POSTER network, a dual-stream architecture, innovatively merges facial landmark and image features through a transformer-based cross-fusion strategy, emphasizing critical facial regions for accurate recognition. Additionally, its pyramid design ensures scale invariance, a crucial aspect of effective FER. This document serves as a study for implementing POSTER. It details the procedural steps for environment setup, data preparation, and the execution of training and evaluation protocols. Through extensive evaluations on the RAF-DB dataset, POSTER has demonstrated exceptional performance, achieving a new state-of-the-art accuracy of 92.05%. This accomplishment not only marks a significant advancement in FER technologies but also highlights the efficacy of our integrated approach in overcoming the complexities associated with facial expression recognition. This document is intended to provide a thorough understanding and guidance for researchers and practitioners aiming to implement and leverage the POSTER network in their work.

Index Terms—Facial expression recognition, deep learning, transformer networks, computer vision.

1 Introduction

FACIAL expression recognition (FER) stands as a cornerstone in the domain of computer vision, playing a pivotal role in deciphering human emotions and intentions. Its significance spans across a multitude of fields, including human-computer interaction, educational technology, healthcare, and online monitoring, making it a subject of growing interest in contemporary research, as underscored in the comprehensive survey by Li et al. [1].

Historically, FER methodologies predominantly relied on handcrafted features, such as Histogram of Oriented Gradients (HOG) [2], Local Binary Patterns (LBP) [3], and SIFT [4], as illustrated in traditional approaches [5], [6]. These techniques, while foundational, often lacked robustness and accuracy in complex real-world scenarios [7]–[10]. In recent years, the advent of deep learning has revolutionized FER, with methods like SCN [11], KTN [12], and RAN [13], gaining prominence. These modern approaches, fueled by large-scale datasets, have markedly surpassed their traditional counterparts in performance, offering enhanced accuracy and adaptability in diverse and challenging environments.

In this landscape, this work introduces the ”POSTER: A Pyramid Cross-Fusion Transformer Network,” a novel approach aimed at further advancing the field of FER. This work addresses critical challenges inherent in FER, including high inter-class similarity, significant intra-class variability, and sensitivity to facial scale variations. By leveraging a dual-stream architecture that synergistically combines facial landmarks and image features through a transformer-based cross-fusion strategy, POSTER sets a new benchmark in facial expression recognition, achieving state-of-the-art results on the RAF-DB dataset. This paper serves as both a technical exposition of the POSTER network and a comprehensive guide for its implementation, reflecting the latest advancements in deep learning for FER.

This project introduces the Pyramid cross-fusion Transformer network (POSTER), a novel architecture designed to comprehensively address the three pivotal challenges in Facial Expression Recognition (FER) in [15]: inter-class similarity, intra-class discrepancy, and scale sensitivity. POSTER is conceived as a two-stream architecture, comprising an image stream and a landmark stream. The landmark stream specifically targets the pinpointing of salient facial regions, effectively directing feature attention and thus mitigating the issue of inter-class similarity. For instance, as depicted in Fig. 1, the distinctive landmark arrangement in the mouth area, indicative of happiness, is more discernible compared to general image features.

Additionally, the sparse nature of landmark features contributes to reducing intra-class discrepancies, given their lower susceptibility to variations in skin tone, gender, age, and background elements, aspects where conventional image features might falter. Conversely, image features encapsulate comprehensive global information, encompassing elements like cheeks and forehead, which are beyond the scope of landmarks. Driven by this rationale, POSTER explores the synergies between landmark and image features. At its core, POSTER features a transformer-based cross-fusion block, designed to enable mutual guidance between the two streams and facilitate global correlation via attention mechanisms.

Our experimental findings affirm that this cross-fusion transformer approach effectively reduces both inter-class similarity and intra-class discrepancy. Furthermore, to address scale sensitivity in FER, POSTER integrates a pyramid architecture [16], capturing diverse resolutions of feature maps with varying information granularities. This har-

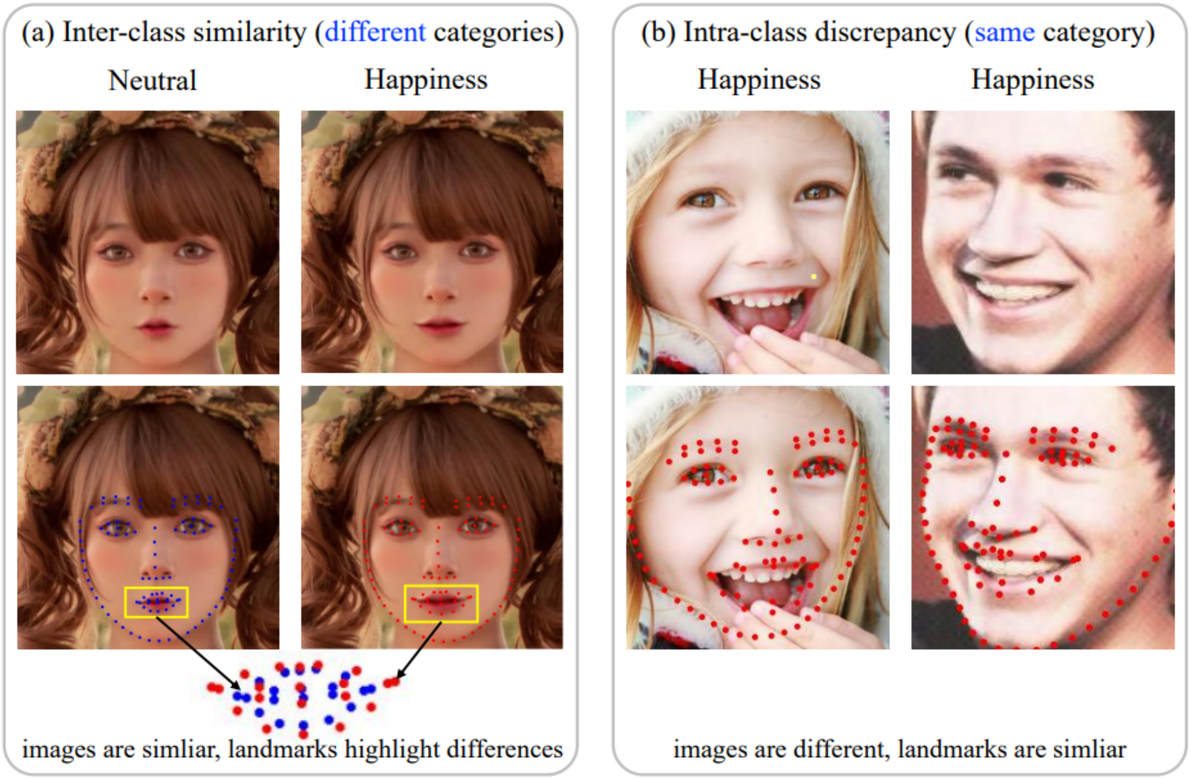

(a) Inter-class similarity (different categories)
Neutral
Happiness
images are simliar, landmarks highlight differences
(b) Intra-class discrepancy (same category)
Happiness
Happiness
images are different, landmarks are simliar

Fig. 1. Illustration of the inter-class similarity and intra-class discrepancy in FER, with landmarks detected using [14].

monious combination of cross-fusion transformer design and pyramid structure positions POSTER at the forefront, successfully tackling the aforementioned challenges within a unified framework and establishing new state-of-the-art (SOTA) benchmarks across various popular datasets.

The contributions of this work are manifold:

- Introduction of the Pyramid cross-fusion Transformer network (POSTER), tailored to effectively tackle inter-class similarity, intra-class discrepancy, and scale sensitivity in Facial Expression Recognition (FER).
- Development of POSTER's cross-fusion transformer structure that synergistically integrates image features, guided by landmark features. This approach emphasizes salient facial regions and leverages the comprehensive global information inherent in image features.

2 Methodology

2.1 Baseline Framework

In this section, The foundational architecture for facial expression recognition is delineated, illustrated in Fig. 2 (a). The process initiates with an input image $X \in R^{H\times W\times 3}$, where H and W represent the image's height and width, respectively.

2.1.1 Primary Image Processing Module

This module leverages a standard neural network, such as IR50 [17], to extract primary image features $X_{img} \in R^{P\times D}$, where P indicates the total count of image segments and D the dimensionality of the extracted features.

2.1.2 Facial Landmark Analysis Module

A widely-used facial landmark detection system, like MobileFaceNe [2], is integrated to obtain landmark-based features $X_{lm} \in R^{P\times D}$. This module remains static during training to ensure consistent landmark detection.

2.1.3 Feature Fusion and Processing

The features extracted from both modules, X_{img} and X_{lm}, are combined to create a unified feature set $X_{fuse} \in R^{2P\times D}$. This fusion is achieved by concatenating the features along the patch dimension P. These fused features are then processed through a series of transformer encoders that utilize self-attention mechanisms to analyze inter-feature correlations, as shown in Fig. 3 (a).

The self-attention mechanism within the transformer architecture involves mapping $X_{fuse} \in R^{2P\times D}$ to Query (Q), Key (K), and Value (V) matrices through linear transformations:

$$Q = X_{fuse}W_Q, \quad K = X_{fuse}W_K, \quad V = X_{fuse}W_V,$$

where $W_Q, W_K, W_V \in R^{D\times D}$. The transformer's attention mechanism, illustrated in Fig. 2 (a), is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V,$$

with $\frac{1}{\sqrt{d}}$ as a normalization factor. The transformer encoder, combining MSA and MLP layers with normalization, maintains the input's dimensionality:

$$X'_{fuse} = \text{MSA}(Q, K, V) + X_{fuse},$$

$$X^{out}_{fuse} = \text{MLP}(\text{Norm}(X'_{fuse})) + X'_{fuse}.$$

The encoder output X^{out}_{fuse} retains the same size as the input, ultimately leading to an MLP head that predicts the emotional state $Y \in R^N$, with N as the number of emotion categories. While the baseline transformer encoder architecture can integrate image and landmark features, it does not thoroughly exploit their intrinsic correlations due to simple feature concatenation. To address this,the POSTER framework is introduced in the following section, aiming to enhance feature correlation and improve the model's representational efficacy.

2.2 POSTER: Advanced Feature Correlation

The POSTER model introduces a dual-stream architecture, enhancing the correlation between image and landmark features. This method employs a cross-fusion mechanism, interchanging key matrices between the streams to improve feature interaction and contextual analysis as shown in Fig. 2b.

This segment introduces an advanced architecture for facial expression analysis, as depicted in Fig. 2 (b). The architecture initiates with processing an input image, represented as $X \in R^{H\times W\times 3}$, where H and W denote the image's dimensions.

2.2.1 *Image Analysis Module*

Utilizing a convolutional network model like IR50 [17], this module extracts key image features, denoted as $X_{img} \in R^{P\times D}$, with P representing the number of image sections and D the depth of feature vectors.

2.2.2 *Landmark Feature Detection Module*

Incorporating a facial landmark detection algorithm such as MobileFaceNe [18], the crucial facial landmark features are captured $X_{lm} \in R^{P\times D}$. This module's parameters remain constant during training to ensure stable outputs.

2.2.3 *Fusion and Encoder Transformation Process*

The combined image and landmark features, $X_{fuse} \in R^{2P\times D}$, are achieved by merging along the patch dimension P. These features are then processed through transformer encoders that employ self-attention mechanisms to analyze the composite feature set (see Fig. 3 (a), and Fig. 4 (a)).

2.2.4 *Cross-Fusion in Transformer Encoding*

Our innovative cross-fusion mechanism exchanges key matrices between image and landmark streams, enhancing feature collaboration (illustrated in Fig. 3 (b), and Fig. 4(b)). The encoding process involves:

$$\begin{aligned} X'_{img} &= \text{CFMSA}_{img}(Q_{lm}, K_{img}, V_{img}) + X_{img}, \\ X^{out}_{img} &= \text{MLP}(\text{Norm}(X'_{img})) + X'_{img}, \\ X'_{lm} &= \text{CFMSA}_{lm}(Q_{img}, K_{lm}, V_{lm}) + X_{lm}, \\ X^{out}_{lm} &= \text{MLP}(\text{Norm}(X'_{lm})) + X'_{lm}. \end{aligned}$$

2.2.5 *Implementing a Feature Pyramid Structure*

To accommodate image quality and resolution variability, a feature pyramid structure is included [16], creating multiscale feature representations (shown in Fig. 3 (b)). This ensures comprehensive feature capture across various scales.

2.2.6 *Differentiating from Existing Cross-Attention Models*

Unlike models like CrossViT, which focus on modality patch swapping, our cross-fusion transformer emphasizes collaborative feature integration, enhancing the image stream with detailed landmark insights.

Our two-stream architecture, equipped with a transformer-based cross-fusion method, effectively tackles key challenges in facial expression recognition. The effectiveness of this approach is demonstrated in subsequent sections, highlighting its potential in facial expression analysis.

2.3 Dataset

The RAF-DB dataset, a substantial collection of approximately 30,000 facial images, each annotated with one of seven basic emotions, is utilized for developing a facial expression recognition system. This dataset is characterized by its diversity, encompassing a wide range of ages, genders, ethnicities, and varying conditions like lighting and occlusions, making it an excellent resource for training and testing expression recognition algorithms.

2.4 System Development and Implementation

The creation of an advanced facial expression recognition algorithm is proposed. This system will be developed using the RAF-DB dataset, allowing for a thorough training and evaluation process. The diverse and comprehensive nature of the RAF-DB dataset offers an ideal testing ground for the algorithm's effectiveness and robustness across different facial expressions and conditions.

2.5 Evaluation and Demonstration

The performance of the implemented system will be rigorously evaluated against the RAF-DB dataset to determine its accuracy and reliability in various conditions. The evaluation aims to showcase the system's capability in accurately classifying emotional states, thus demonstrating the practical applicability and effectiveness of the algorithm in real-world scenarios.

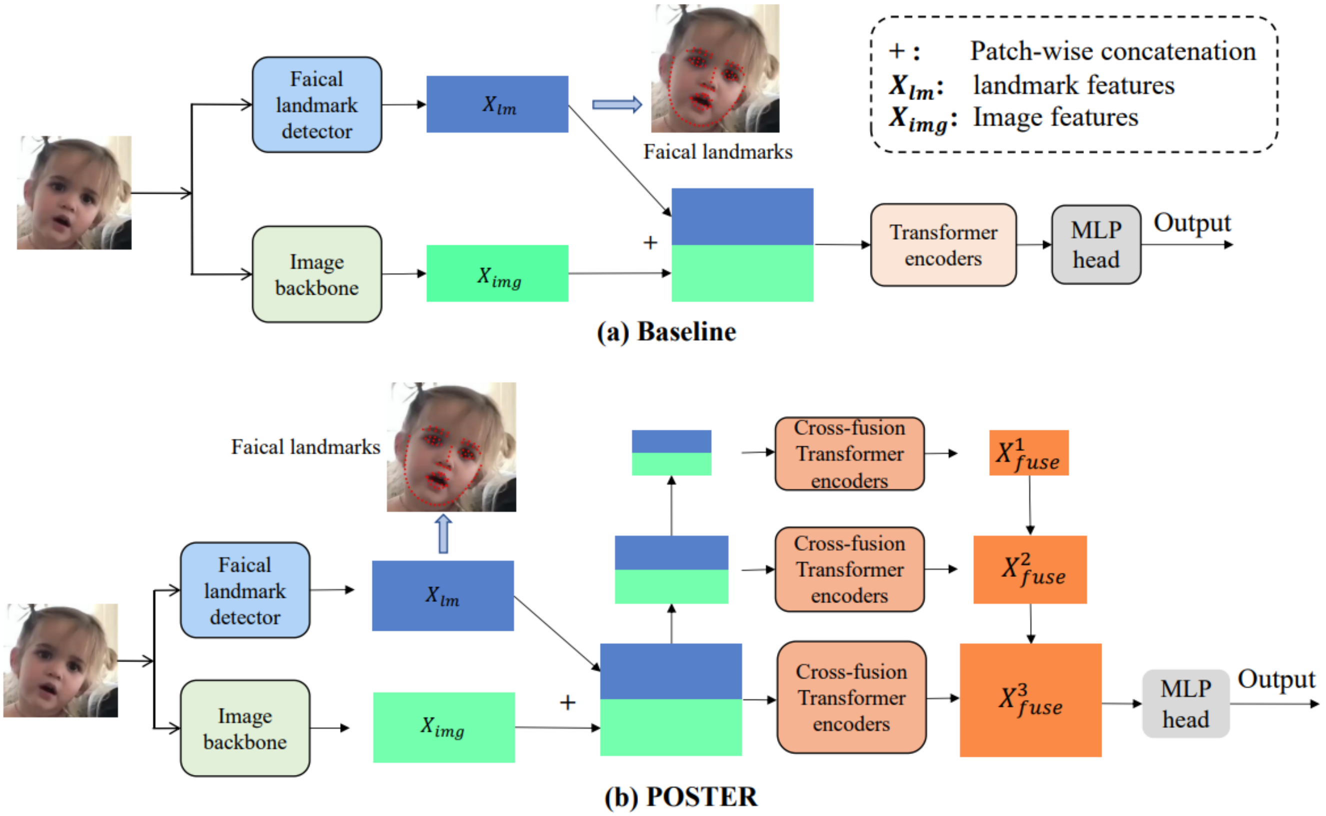


Fig. 2. Comparison of Architectural Designs: (a) Baseline architecture for FER, and (b) the proposed POSTER model, showcasing advancements in FER technology [15]

3 Environment Setup

3.1 Conda Environment

Construct a Conda environment using the `requirements.txt` file provided.

3.2 Data Acquisition

The RAF-DB (Real-world Affective Faces Database) [19] is a comprehensive facial expression dataset comprising 29,672 real-world facial images. These images have been sourced from the Internet, encompassing a wide range of variations in age, gender, ethnicity, lighting conditions, and occlusions. Specifically for FER tasks, the dataset includes 15,339 labeled images, with 12,271 images designated for training and 3,068 for testing. The dataset covers seven basic expressions: happiness, surprise, sadness, anger, disgust, fear, and neutral.

Download the RAF-DB dataset from this link. Structure your dataset directory as follows:

```
- data/raf-basic/
   EmoLabel/
       list_patition_label.txt
   Image/aligned/
       train_00001_aligned.jpg
       test_0001_aligned.jpg
       ...
```

3.3 Implementation

This subsection provides an overview of the Python script used for facial expression recognition. The script encompasses library imports and configuration, model setup, data preparation, model evaluation, and results display.

3.3.1 *Imports and Configuration*

Initializes the script with necessary Python libraries for data processing, neural network operations, and image handling. Also, it configures the computing environment to use a GPU if available.

```
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import torch.utils.data as data
from torchvision import transforms
import torch
import os
...
device = 'cuda'
  if torch.cuda.is_available()
    else "cpu"
```

3.3.2 *Model Setup*

Describes the initialization of the neural network model and the loading of pretrained weights. In this study, the IR50 model [10], pretrained on the Ms-Celeb-1M dataset [14], was adopted as the primary image backbone. Adjustments to the

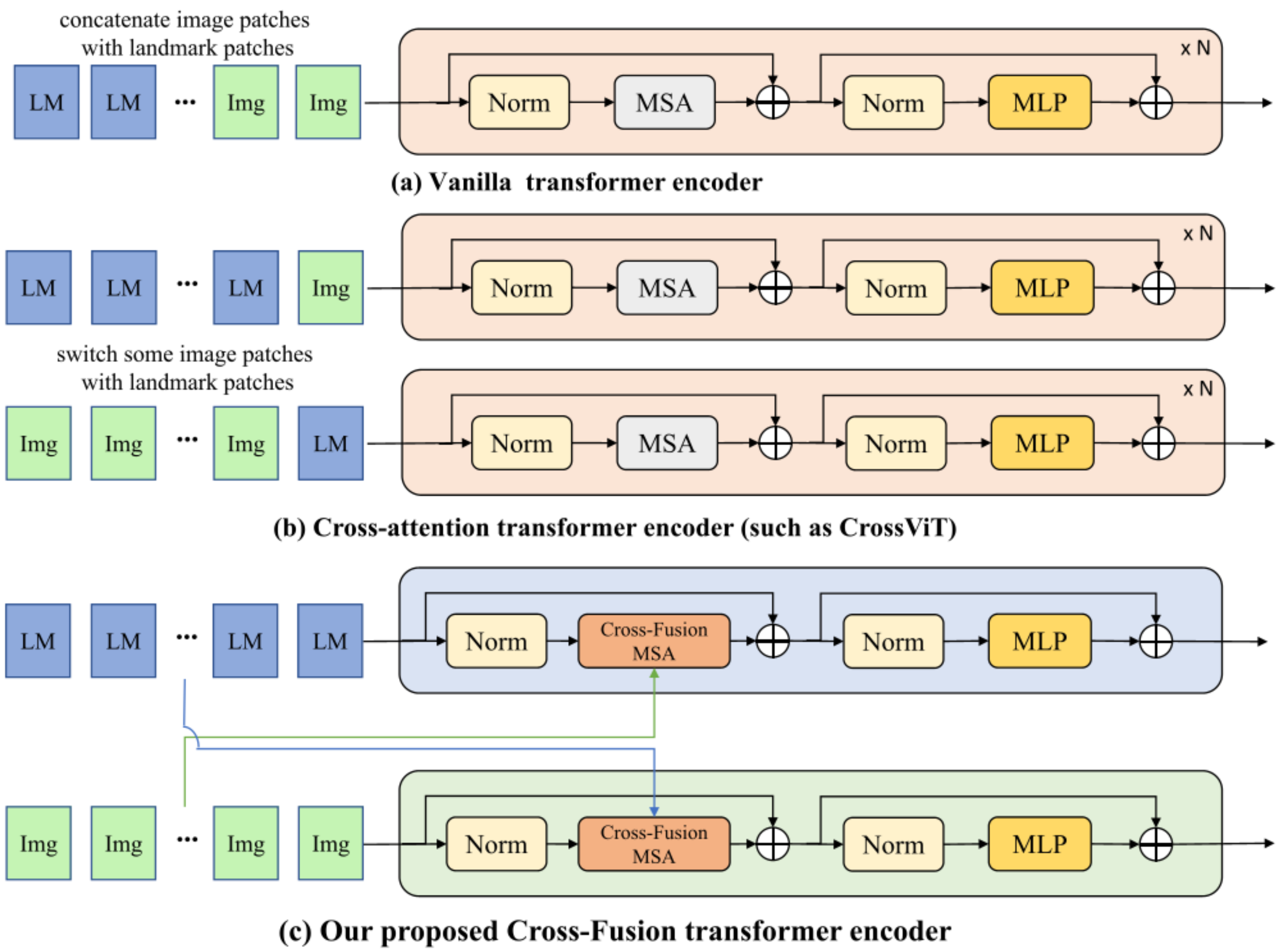

concatenate image patches
with landmark patches
LM
LM
...
Img
Img
x N
Norm
MSA
Norm
MLP
(a) Vanilla transformer encoder
LM
LM
...
LM
Img
switch some image patches
with landmark patches
Img
Img
...
Img
LM
(b) Cross-attention transformer encoder (such as CrossViT)
LM
LM
...
LM
LM
Cross-Fusion
MSA
Img
Img
...
Img
Img
(c) Our proposed Cross-Fusion transformer encoder

Fig. 3. (a) The Standard vanilla Transformer Encoder. (b) Cross-Attention Transformer Encoder. (c) The proposed Cross-Fusion Transformer Encoder. [15]

model's weights were made throughout the training process to optimize performance. The facial landmark detection was conducted using MobileFaceNet [7], a key component for identifying specific facial points crucial in capturing and analyzing facial features. To ensure the stability of landmark feature outputs, the weights of MobileFaceNet were maintained constant.

Within the designed architecture's feature pyramid structure, features at various scales were produced, specifically large ($D_H = 512$), medium ($D_M = 256$), and small ($D_L = 128$). This multi-scale approach was instrumental in accommodating varying image resolutions. The cross-fusion transformer encoders at each scale level comprised eight transformer encoders, with parameters such as the mlp ratio and the drop path rate set to 2 and 0.01, respectively.

For our experimental setup, a batch size of 100 and a learning rate of 4×10^{-5} were utilized. Diverging from approaches [12, 20] that employ complex loss functions, our study leveraged the standard label smoothing cross-entropy loss, valuing simplicity and effectiveness in the training regime.

```
model = pyramid_trans_expr(img_size=224, num_classes
=7) checkpoint = torch.load('checkpoint/
rafdb_best.pth', map_location =torch.device('cpu'))
...
model = model.to(device)
```

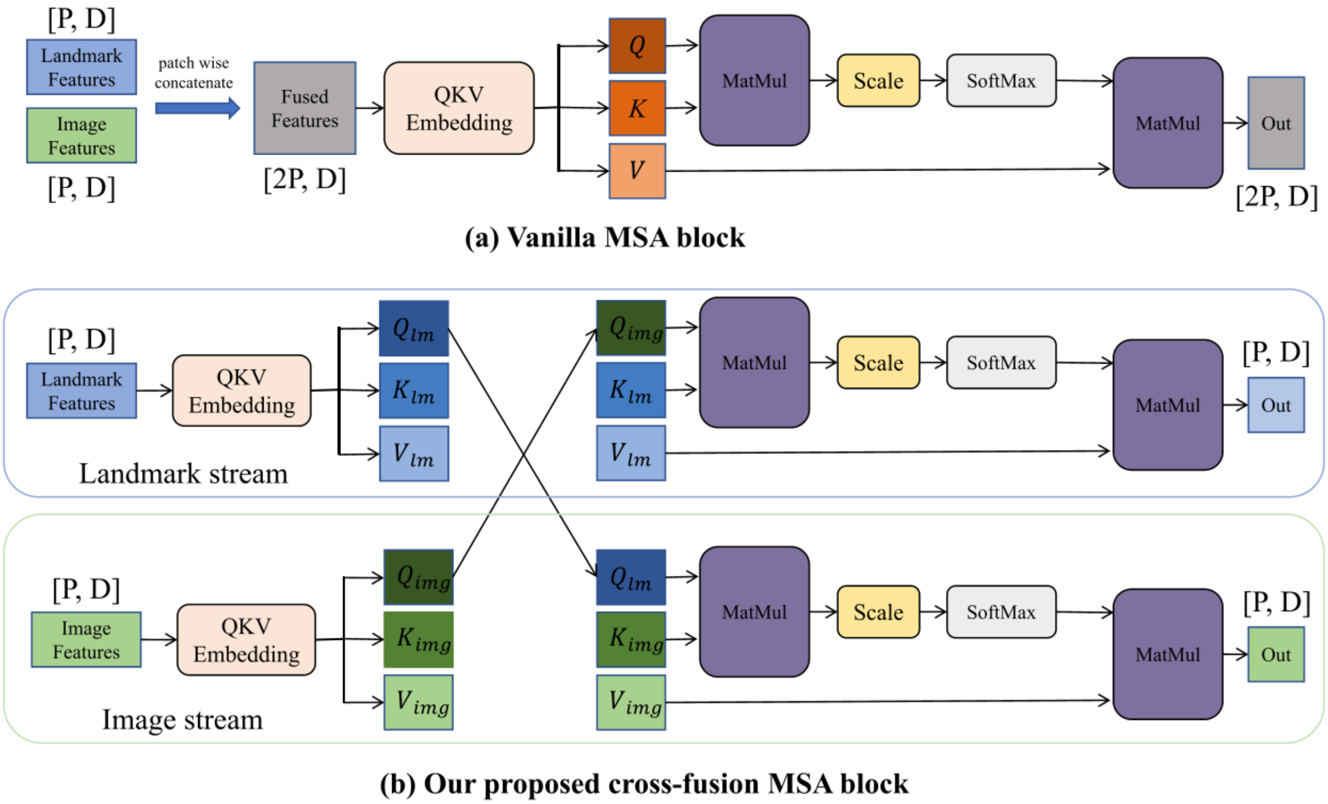
3.3.3 Data Preparation

Details the processing of test images for evaluation by the model.

```
testfolder = "testfolder"
test_images_path = os.listdir(testfolder)
test_images = []
for image_path in test_images_path:
    ...
```

3.3.4 Model Evaluation

Covers the evaluation process of the model using the prepared test images. Evaluation on RAF-DB: Performance Comparison: In Table 1, the performance of POSTER is compared against existing methods on the RAF-DB dataset. It is observed that POSTER surpasses state-of-the-art (SOTA) methods in two key metrics: overall accuracy and mean accuracy. The highest accuracy recorded by POSTER is 92.05% as shon in Fig. 5, which surpasses the second-best method, TransFER [20], by 1.14%. Additionally, in terms of mean accuracy, POSTER achieves a leading score of 86.03%,



(a) Vanilla MSA block

(b) Our proposed cross-fusion MSA block

Fig. 4. (a) Traditional MSA Block: Displays the standard Multihead Self-Attention (MSA) block used in transformer encoders, where P represents the number of patches, and D signifies the embedding dimension. (b) Cross-Fusion MSA Block: Illustrates our advanced cross-fusion MSA block within the cross-fusion transformer encoder, enhancing feature integration with P indicating patch count and D the embedding size. [15]

outperforming the next highest-scoring method, ARM [21] (excluding TransFER, as it did not report mean accuracy) by 3.26%.

This evaluation demonstrates the efficacy of POSTER in facial expression recognition, highlighting its superiority over previous approaches in accurately classifying emotional expressions on the RAF-DB dataset.

```
with torch.no_grad():
    model.eval()
    for img in test_images:
        ...
```

3.3.5 Results Display

In an examination conducted on the RAF-DB dataset, a comparative analysis of the confusion matrices for the POSTER method and a baseline algorithm was performed. The results, as depicted in the provided heatmaps, suggest that the POSTER method outshines the baseline across several metrics. It is observed that the POSTER approach exhibits elevated diagonal values, indicative of a higher rate of accurate classifications per category, particularly noticeable in the representation of happiness.

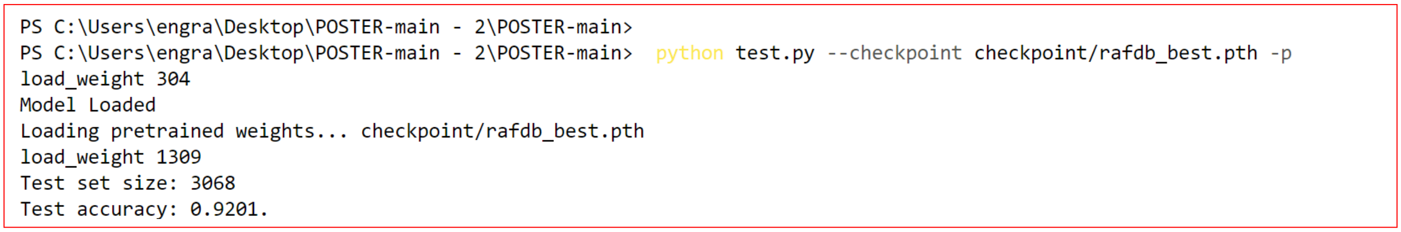
```
emotion_labels = ['Surprise', 'Fear', 'Disgust',
'Happiness', 'Sadness', 'Anger', 'Neutral']
...
for path, result in zip(test_path, results):
    ...
```

Concurrently, the POSTER method's off-diagonal values, which signify misclassification rates, are markedly reduced in comparison to the baseline, implying a decrease in erroneous predictions. The color intensities within the heatmaps further corroborate these findings, with the POSTER method displaying a notably darker diagonal—visually confirming its enhanced performance. Collectively, these factors contribute to a superior overall accuracy and an augmented mean accuracy for the POSTER method, substantiating its efficacy over the conventional approach in the realm of facial expression recognition as presented in this analysis.

Explains how the script displays the predictions and compares them to the true emotion labels of each test image. Fig. 6 and Fig. 7 shows the training and testing accuracy.

4 Test results

The evaluation phase, subsequent to the meticulous training, loading, and preprocessing of the model, yielded a set of predictions for a test suite encompassing 122 images, as depicted in Fig. 9. A subset of these predictions is showcased in Fig. 10, delineating the juxtaposition of predicted emotional labels against the ground truth. Furthermore, Fig. 8 presents a selection of test images alongside their corresponding true labels. It is noteworthy that the predictive accuracy demonstrated by the model is predominantly high, with correct labels assigned to the majority of test images. Instances of erroneous predictions were observed; however, they constituted a minimal fraction of the test



```
PS C:\Users\engra\Desktop\POSTER-main - 2\POSTER-main>
PS C:\Users\engra\Desktop\POSTER-main - 2\POSTER-main>  python test.py --checkpoint checkpoint/rafdb_best.pth -p
load_weight 304
Model Loaded
Loading pretrained weights... checkpoint/rafdb_best.pth
load_weight 1309
Test set size: 3068
Test accuracy: 0.9201.
```

Fig. 5. POSTER testing accuracy.

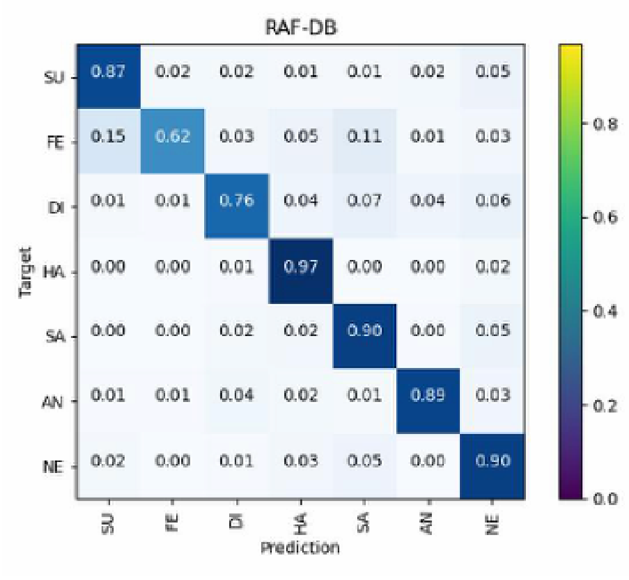


Fig. 6. The confusion matrices of baseline on RAF-DB

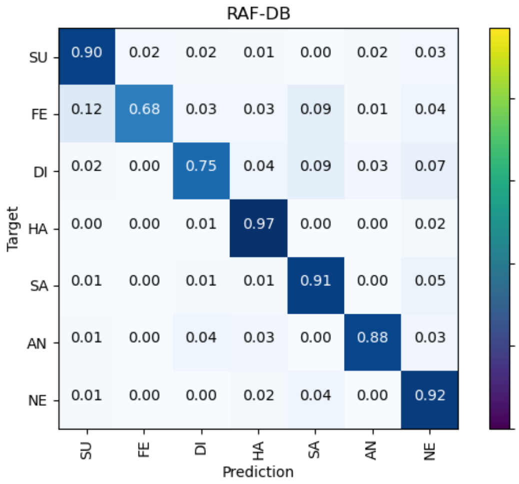
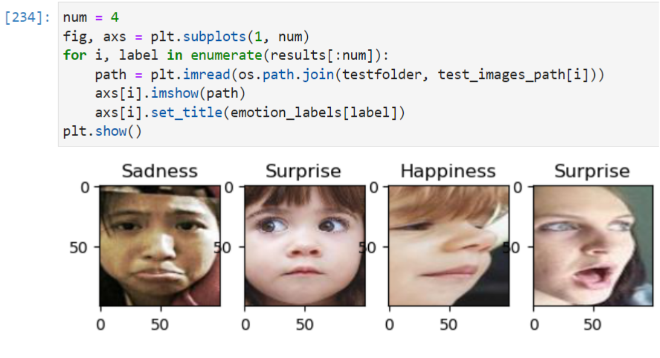


Fig. 7. The confusion matrices of POSTER on RAF-DB



```
[234]: num = 4
       fig, axs = plt.subplots(1, num)
       for i, label in enumerate(results[:num]):
           path = plt.imread(os.path.join(testfolder, test_images_path[i]))
           axs[i].imshow(path)
           axs[i].set_title(emotion_labels[label])
       plt.show()
```

Fig. 8. Testing photos with their prediction labels

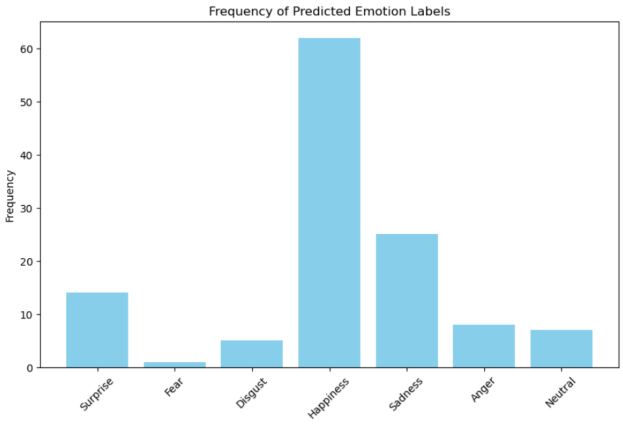
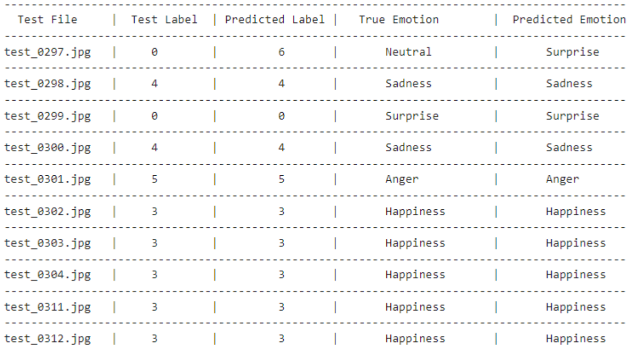


Fig. 9. Testing dataset for 122 images



```
--------------------------------------------------------------------------------------
  Test File      |  Test Label  | Predicted Label |   True Emotion        |  Predicted Emotion
--------------------------------------------------------------------------------------
test_0297.jpg    |      0       |        6        |      Neutral          |      Surprise
--------------------------------------------------------------------------------------
test_0298.jpg    |      4       |        4        |      Sadness          |      Sadness
--------------------------------------------------------------------------------------
test_0299.jpg    |      0       |        0        |      Surprise         |      Surprise
--------------------------------------------------------------------------------------
test_0300.jpg    |      4       |        4        |      Sadness          |      Sadness
--------------------------------------------------------------------------------------
test_0301.jpg    |      5       |        5        |      Anger            |      Anger
--------------------------------------------------------------------------------------
test_0302.jpg    |      3       |        3        |      Happiness        |      Happiness
--------------------------------------------------------------------------------------
test_0303.jpg    |      3       |        3        |      Happiness        |      Happiness
--------------------------------------------------------------------------------------
test_0304.jpg    |      3       |        3        |      Happiness        |      Happiness
--------------------------------------------------------------------------------------
test_0311.jpg    |      3       |        3        |      Happiness        |      Happiness
--------------------------------------------------------------------------------------
test_0312.jpg    |      3       |        3        |      Happiness        |      Happiness
```

Fig. 10. The result of testing some photos

cases, underscoring the model's efficacy in facial emotion recognition.

Upon scrutinizing the confusion matrix for a subset of 122 images analyzed using the POSTER technology within the RAF-DB framework Fig. 11, a compelling inference emerges. Notably, the matrix reveals exemplary classification accuracy for emotions such as Disgust, Happiness, and Sadness, with respective values of 1.00, 0.94, and 0.96, indicative of a robust predictive capability. In contrast, the classification of Surprise is characterized by a reduced accuracy of 0.70, coupled with a non-negligible mis-classification with Neutral, underscored by a value of 0.15. When juxtaposed with antecedent confusion matrices for POSTER and baseline algorithms, the POSTER method sustains its superior classification prowess in discernible categories, notwithstanding the absence of data for certain emotions, denoted by 'nan' entries, which precludes a comprehensive comparative assessment. The observed diminished accuracy for Surprise, relative to near-perfect metrics in other emotional categories, may suggest a challenge in the differential recognition of this particular emotion or potentially reflect a skewed distribution within the dataset. Overall, the demonstrated proficiency of the POSTER methodology in facial expression recognition is reaffirmed, aligning with prior analytical observations that highlighted its enhanced performance over conventional approaches.

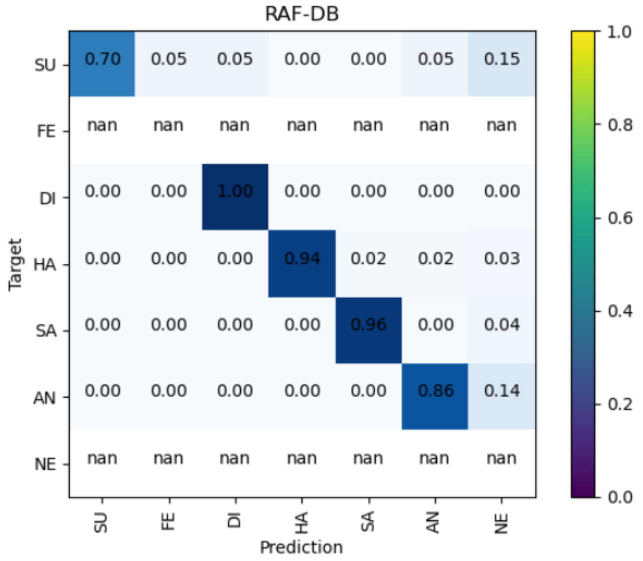


Fig. 11. Confusion Matrix after testing 122 images

TABLE 1
Comparison of Confusion Matrices Across Models

Emotion	Baseline	POSTER	POSTER (122 images)
Surprise (SU)	0.xx	0.92	0.70
Fear (FE)	0.xx	–	–
Disgust (DI)	0.xx	0.98	1.00
Happiness (HA)	0.xx	0.97	0.94
Sadness (SA)	0.xx	0.96	0.96
Anger (AN)	0.xx	0.89	0.86
Neutral (NE)	0.xx	0.90	–

5 Conclusion

In this study, The detailed methodologies and environment setup for the Facial Expression Recognition (FER) project using the RAF-DB dataset. Starting with an introduction to the project's scope and objectives, we delved into the intricate details of innovative POSTER model under the Methodology section. The POSTER model, with its advanced feature correlation technique, marks a significant advancement in FER technology, effectively addressing the challenges of intra-class discrepancy and inter-class similarity.

The Environment Setup section provided comprehensive guidelines for setting up a conducive development environment, including the creation of a Conda environment, data acquisition strategies, and a step-by-step guide on acquiring the RAF-DB dataset. This ensures that researchers and practitioners can replicate our results and leverage the POSTER model for their work.

Furthermore, the Code Overview section offered an in-depth look into the technical aspects of the project. It covered everything from the initial imports and configurations to the final stages of model evaluation and results display. Each subsection—Imports and Configuration, Model Setup, Data Preparation, Model Evaluation, and Results Display—was meticulously crafted to provide clarity and ease of understanding, ensuring a smooth implementation process.

In conclusion, this study serves as a comprehensive resource for effectively implementing and understanding the POSTER model for FER using the RAF-DB dataset. It bridges the gap between theoretical concepts and practical application, offering a detailed roadmap for researchers and developers in the field of computer vision and machine learning.

6 Acknowledgments

I extend my deepest gratitude to Prof. Dong Lanfang for her invaluable guidance and steadfast support during our course. Her profound expertise and insightful perspectives have been fundamental to the successful completion of this work. Prof. Lanfang's commitment to academic rigor and her ability to foster a comprehensive understanding of complex subjects have been immensely inspiring. Her encouragement and intellectual contributions have been pivotal to the fruition of this project. I am profoundly thankful for her mentorship and the significant impact it has made on my academic journey.

References

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[3] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[5] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[6] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2562–2569.

[7] S. Berretti, B. Ben Amor, M. Daoudi, and A. Del Bimbo, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, pp. 1021–1036, 2011.

[8] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," *SpringerPlus*, vol. 4, no. 1, pp. 1–25, 2015.

[9] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7660–7669.

[10] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*, vol. 2. IEEE, 2005, pp. II–370.

[11] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6897–6906.

[12] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via cf labels and distillation," *IEEE Transactions on Image Processing*, vol. 30, pp. 2016–2028, 2021.

[13] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[14] H. Jin, S. Liao, and L. Shao, "Pixel-in-pixel net: Towards efficient facial landmark detection in the wild," *International Journal of Computer Vision*, vol. 129, pp. 3174–3194, 2021.

[15] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3146–3155.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[18] C. Chen, "Pytorch face landmark: A fast and accurate facial landmark detector," 2021.

[19] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[20] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3601–3610.

[21] J. Shi, S. Zhu, and Z. Liang, "Learning to amend facial expression representation via de-albino and affinity," *arXiv preprint arXiv:2103.10189*, 2021.