HUST

Introduction to Artificial Intelligence

Lecturer: **Prof. Dr. Hai Van Pham**

Class ID: 147838

Project Proposal
Facial Emotion Recognition

Member List

Hoang Quoc Hung	20226043	hung.hq226043@sis.hust.edu.vn
Le Dai Lam	20225982	lam.ld225982@sis.hust.edu.vn
Vu Hai Dang	20225962	dang.vh225962@sis.hust.edu.vn
Nguyen Minh Khoi	20229050	khoi.nm226050@sis.hust.edu.vn
Hoang Khai Manh	20225984	manh.hk225984@sis.hust.edu.vn

Facial Emotion Recognition

Abstract

Facial Emotion Recognition (FER) is instrumental in enhancing human-computer interactions and advancing psychological analysis. Despite its importance, the effectiveness of traditional FER methods is often undermined by high inter-class similarity and significant label noise. Addressing these challenges, this report details the application of the Erasing Attention Consistency (EAC) method within established FER systems. EAC improves the robustness and accuracy by promoting the consistency of relevant features in facial images, instead of merely minimizing loss. This method involves selectively erasing parts of the images and analyzing the impact on feature stability, effectively reducing the influence of noisy data. Our implementation of EAC in the ResNet50 and MobileNet_V2 architectures—known for their efficiency in deep learning and mobile optimization—demonstrates notable enhancements in model performance, particularly under noisy conditions. The adaptability of EAC across different architectures highlights its vast potential for broad implementation in real-world FER tasks, suggesting new directions for both research and practical deployments in the domain.

1 Introduction

In the fields of human-computer interaction and psychological research, Facial Emotion Recognition (FER) is a pivotal technology. The effectiveness of FER systems is often compromised by significant challenges such as high inter-class similarity and ambiguous labeling of facial expression data. Addressing these issues is crucial for improving the accuracy and reliability of FER applications. This report delves into the application of the Erasing Attention Consistency (EAC) method, a novel approach aimed at enhancing the robustness and accuracy of FER systems. EAC distinguishes itself by focusing on feature consistency, rather than solely on error minimization. By methodically erasing segments of facial images and assessing the consistency of the remaining features with their horizontally flipped counterparts, EAC encourages the model to prioritize stable and reliable facial features. This approach is particularly effective in mitigating the impact of noisy or incorrect labels commonly encountered in real-world datasets. The implementation of EAC within renowned neural network architectures like ResNet50 and MobileNet_V2 offers a strategic advantage. ResNet50, with its deep residual blocks, effectively prevents the vanishing gradient problem, enhancing learning in deep networks. Meanwhile, MobileNet_V2 is optimized for performance in resource-constrained environments such as mobile and embedded systems, making it an ideal choice for deploying FER applications that require efficiency and speed. This report will further outline the implementation details of the EAC method and present empirical evidence of its efficacy in improving the performance of FER systems. By applying EAC, we aim to advance the state of FER technology, making it more robust and effective in diverse and noise-prone real-world environments.

2 Related Work

Class Activation Mapping:
Class Activation Mapping (CAM) is an advanced technique used in the field of deep learning to enhance the interpretability of Convolutional Neural Networks (CNNs). CAM allows us to visualize the predicted class scores on given images by highlighting the discriminative regions detected by the CNN. Essentially, it provides insight into which parts of an image the model considers significant when making a prediction.

In a CNN designed for classification tasks, an attention map is generated as a weighted sum of the feature maps from the last convolutional layer, combined with the weights from a fully connected (FC) layer. This weighted combination results in a heatmap that indicates the areas of the image that are most relevant to the model's decision. By examining these attention maps, researchers and practitioners can gain a better understanding of the model's inner workings and the specific features it uses to distinguish between different classes.

The practical applications of CAM are numerous. For instance, in medical imaging, CAM can help

identify the regions of an X-ray or MRI that a model relies on to diagnose a condition, providing valuable insights for medical professionals. Similarly, in the field of autonomous driving, CAM can be used to ensure that self-driving cars are making decisions based on relevant aspects of their environment, such as road signs and obstacles, thereby enhancing safety and reliability.predictions

Erasing Attention Consistency:
Facial Expression Recognition (FER) is a challenging task, particularly in the presence of noisy labels. Noisy labels can arise due to the high inter-class similarity and annotation ambiguity that are common in facial expression images. Traditional methods for handling noisy labels in FER generally fall into two categories: sample selection and label ensembling. Sample selection involves identifying and potentially excluding mislabeled samples, while label ensembling combines multiple labels to reduce noise.

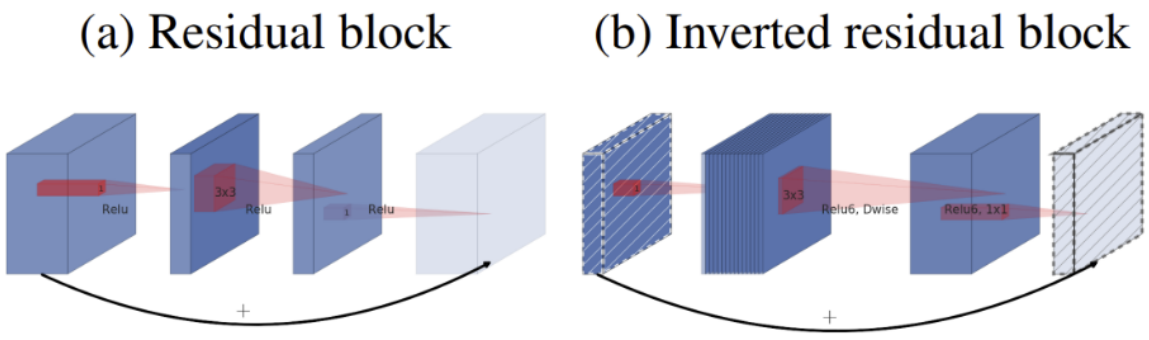
Our approach introduces a novel framework called Erasing Attention Consistency (EAC), which builds upon the principles of attention consistency to address the issue of noisy labels in FER. Unlike traditional methods that rely heavily on loss-based sample selection, EAC employs a feature-learning strategy to mitigate the effects of noisy labels.

The core idea behind EAC is to erase parts of input images and utilize flip semantic consistency, thereby compelling the model to focus on consistent features across both the original and flipped versions of the images. This process involves selectively removing (erasing) regions of the image during training and ensuring that the model's attention remains on features that are invariant to these transformations. By doing so, EAC prevents the model from overfitting to noisy samples, enhancing its robustness and generalization capabilities.

One of the significant advantages of EAC is that it does not require prior knowledge of the noise rate or the training of multiple models. This makes it a more efficient and scalable solution for dealing with noisy labels in FER. The EAC framework can be particularly beneficial in real-world applications where label noise is prevalent, such as in large-scale datasets collected from diverse and uncontrolled environments.

In summary, CAM and EAC represent sophisticated techniques that advance our ability to interpret and improve the performance of deep learning models in various applications. CAM enhances model transparency by visualizing the regions of interest, while EAC addresses the challenge of noisy labels in FER through innovative feature-learning strategies. Together, these methods contribute to the development of more reliable and interpretable AI systems.

Residual Block and Inverted Residual Block:



Feature	Residual Blocks	Inverted Residual Blocks
Main Purpose	Mitigate vanishing gradient	Enhance efficiency in mobile/embedded
Architecture	Shortcut connections with identity mappings and convolutional layers	Thin bottleneck layers with expansion and projection phases using depth-wise separable convolutions
Training Depth	Suitable for very deep networks	Optimized for shallower, efficient networks
Parameter Efficiency	More parameters due to deeper layers	Fewer parameters, highly efficient
Computational Cost	Higher due to deeper layers	Lower due to use of depth-wise separable convolutions
Performance	High accuracy in complex tasks	Maintains performance with lower computational resources
Typical Use Cases	Complex tasks requiring high capacity models	Mobile and embedded applications requiring efficiency
Key Advantage	Allows very deep networks without degradation	Balances performance and computational efficiency
Network Examples	ResNet, ResNeXt	MobileNetV2, MobileNetV3

Table 1: Comparison of Residual Blocks and Inverted Residual Blocks

3 Proposed Method

3.1 ResNet 50 Architecture

3.1.1 Residual Block

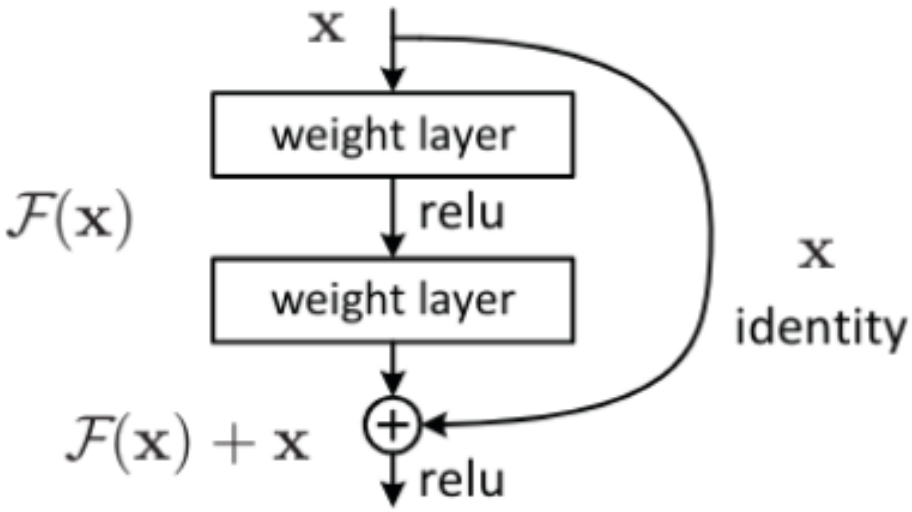


Figure 1: Residual learning: a building block

We consider a building block defined as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\}) + \mathbf{x}. \tag{1}$$

Where x and y are the input and output vectors of the layers considered. The function $\mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\})$ represents the residual mapping to be learned. Note that the dimension of x and F must be equal in Eq(1).

In case the dimension of input and output of the residual blocks are not the same we can perform a linear projection W_s by the shortcut connections to match the dimensions:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{\mathbf{W}_i\}) + \mathbf{W}_x\mathbf{x} \tag{2}$$

3.1.2 ResNet50 Architecture

Inspired by the philosophy of VGG networks, ResNet50 addresses the vanishing gradient problem that arises when building deeper plain networks by incorporating residual blocks. These residual blocks introduce skip connections between layers, enabling the model to learn residual functions with reference to the layer inputs, rather than learning unreferenced functions.

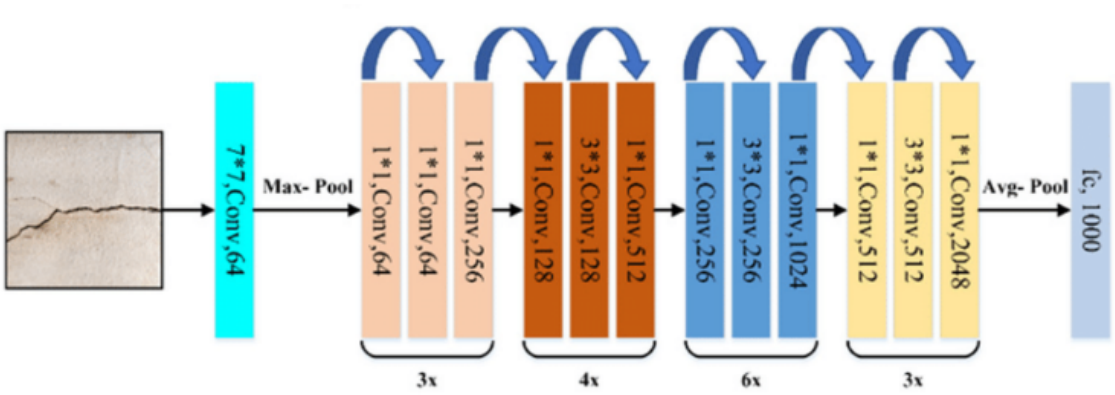


Figure 2: The architecture of ResNet50

The ResNet50 architecture is composed of several stages, each consisting of multiple residual blocks. The input image first undergoes a 7x7 convolution followed by max pooling, reducing the spatial dimensions and increasing the depth.

By using these skip connections, ResNet50 effectively mitigates the vanishing gradient problem, allowing the network to be trained with more layers while maintaining high accuracy. The architecture's ability to learn residuals helps in constructing very deep networks without degradation in performance, making ResNet50 a powerful model for various computer vision tasks.

3.2 Why Do Residual Networks Work?

Vanishing Gradient Problem

As neural networks become deeper, a significant challenge known as the vanishing gradient problem arises. This issue occurs when the gradients in the deeper layers of the network become extremely small during backpropagation, eventually approaching zero. As a result, these layers learn very slowly or not at all, hindering the effective training of the network.

Residual Networks and Shortcut Connections

Residual Networks (ResNets) address the vanishing gradient problem by introducing shortcut connections, also known as skip connections. These connections bypass one or more layers by adding the output of a previous layer directly to the output of a subsequent layer. Mathematically, if $\mathcal{F}(x)$ represents the residual mapping to be learned, the network actually learns $H(x) = \mathcal{F}(x) + x$, where x is the input. This formulation ensures that the network learns the residual function $\mathcal{F}(x)$, which is easier to optimize.

Benefits of Skip Connections

Gradient Flow Improvement: The skip connections create an uninterrupted path for the gradient to flow backward through the network. This mitigates the vanishing gradient problem by allowing the gradients to propagate without significant reduction, thus facilitating more effective training of the neural network.

Learning Minor Transformations: Instead of learning the entire transformation from input to output, the network only needs to learn the residual (or difference) between the input and the desired output. This simplifies the learning task, making it easier for the network to optimize.

Avoiding Performance Degradation: In traditional deep networks, adding more layers can sometimes degrade performance due to the increased difficulty in training. However, in ResNets, the skip connections ensure that adding more layers does not degrade performance. This enables the construction of very deep networks without the risk of reduced accuracy.

Practical Effectiveness of ResNets

Residual Networks have proven to be highly effective in practice. They have enabled the training of networks with hundreds or even thousands of layers, achieving superior performance across a wide range of applications. In particular, ResNets have shown remarkable success in computer vision tasks such as image classification, object detection, and semantic segmentation. Additionally, they have been effectively applied in natural language processing tasks, demonstrating their versatility and robustness.

In summary, the introduction of residual connections in ResNets solves the vanishing gradient problem, facilitates more efficient training, and enables the construction of very deep neural networks. These advantages have made ResNets a foundational architecture in the field of deep learning, contributing significantly to advancements in both computer vision and natural language processing.

3.3 MobileNet-V2 Architecture

3.3.1 Depthwise Separable Convolution

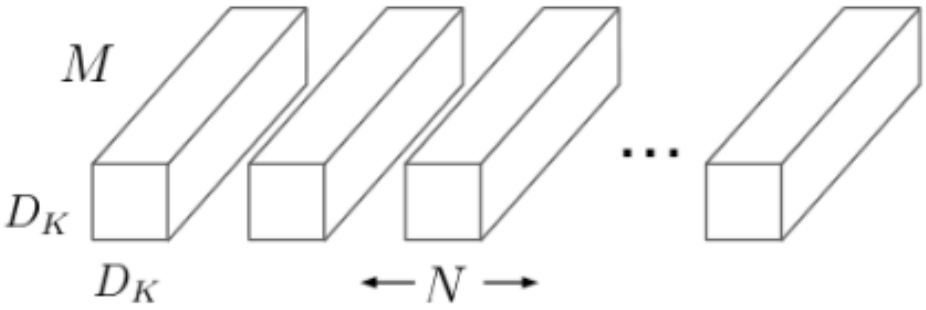


Figure 3: Standard Convolution Filters

Standard Convolutions have the computational cost of:

$$D_k \cdot M \cdot N \cdot D_F \cdot D_F \tag{3}$$

Where: M, N is the input and output channels. $D_K.D_K$ is the kernel size and $D_F.D_F$ is the size of the feature map.
By extracting Standard Convolutions into two parts. The first one is depthwise Convolution with computational cost $D_K \cdot D_K \cdot M \cdot D_F \cdot D_F$. And the second one is pointwise Convolution with the cost of computation $M \cdot N \cdot D_F \cdot D_F$. Combining two processes (Depthwise Separable Convolution), we get a

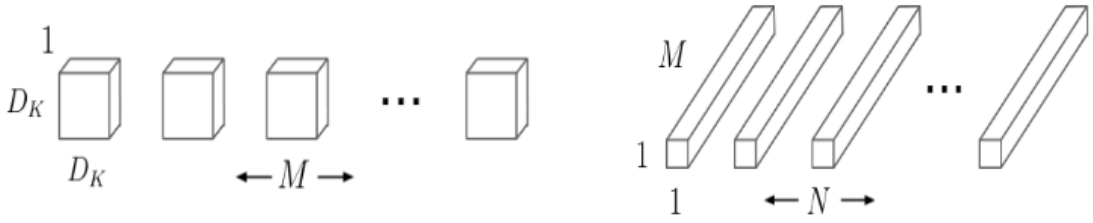


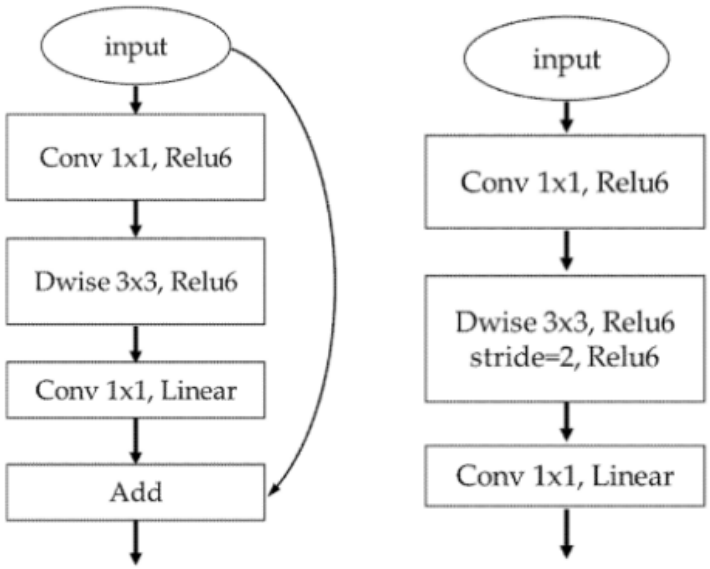
Figure 4: Pointwise Convolution

reduction in computation cost:

$$\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{4}$$

Depthwise convolution is more computationally efficient and has fewer parameters compared to normal convolution, making it suitable for resource-constrained environments like mobile devices. However, normal convolution tends to provide better accuracy as it captures more complex features by combining information across all input channels. The choice between the two depends on the application's requirements for performance versus efficiency. Often, a combination of both is used to balance these needs, such as in Depthwise Separable Convolutions.

3.3.2 Inverted Residual Block



(a) Stride=1 block (b) Stride=2 block

Figure 5: The inverted residual block in MobileNet-v2

The image depicts two types of inverted residual blocks in MobileNet-v2, tailored for mobile vision applications. The Stride-1 block (a) uses a 1x1 convolution with ReLU6 activation, a depthwise 3x3 convolution (also with ReLU6), and a 1x1 linear convolution, adding the output back to the input for residual connections. The Stride-2 block (b) follows a similar process but includes a stride of 2 in its depthwise convolution to downsample, bypassing the addition step due to size differences. These blocks utilize depthwise separable convolutions to optimize computational efficiency and maintain effective feature processing.

It has two structures. The first structure is the stride = 1 with the shortcut connections between bottlenecks when the dimension of input and output are same. The second one is stride = 2 for downsampling the size of the feature map.

3.3.3 MobileNet-V2 Structure

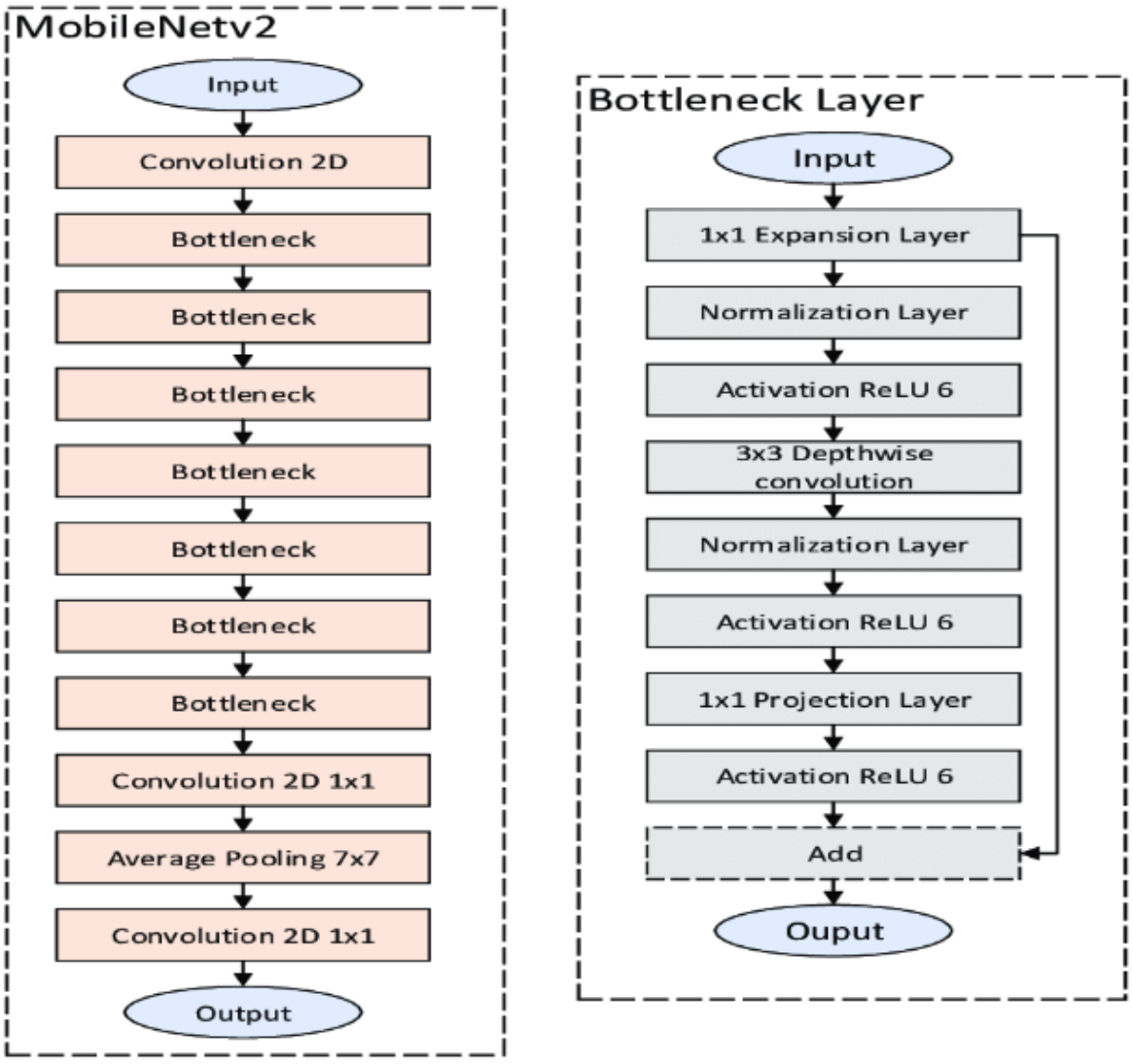


Figure 6: The architecture of MobileNet-V2

The model begins with an input layer followed by an initial 2D convolution, leading into multiple bottleneck layers, which are the core of the model. Each bottleneck layer consists of a 1x1 expansion layer, a 3x3 depthwise convolution, and a 1x1 projection layer, all interspersed with normalization and ReLU6 activations. Notably, many of these layers include residual connections to enhance gradient flow during training. The architecture concludes with a final 1x1 convolution, a 7x7 average pooling, and another 1x1 convolution to produce the output. This setup, leveraging depthwise separable convolutions and linear bottlenecks, significantly reduces computational demands while maintaining high efficiency, making MobileNet-V2 ideal for devices with limited computational resources.

3.4 FrameWork and Training

3.4.1 Framework

We use a Convolutional Neural Network (CNN) with images to extract the final layer. After that, we customize the final layers and modify the Loss function according to the Framework in Fig[7].

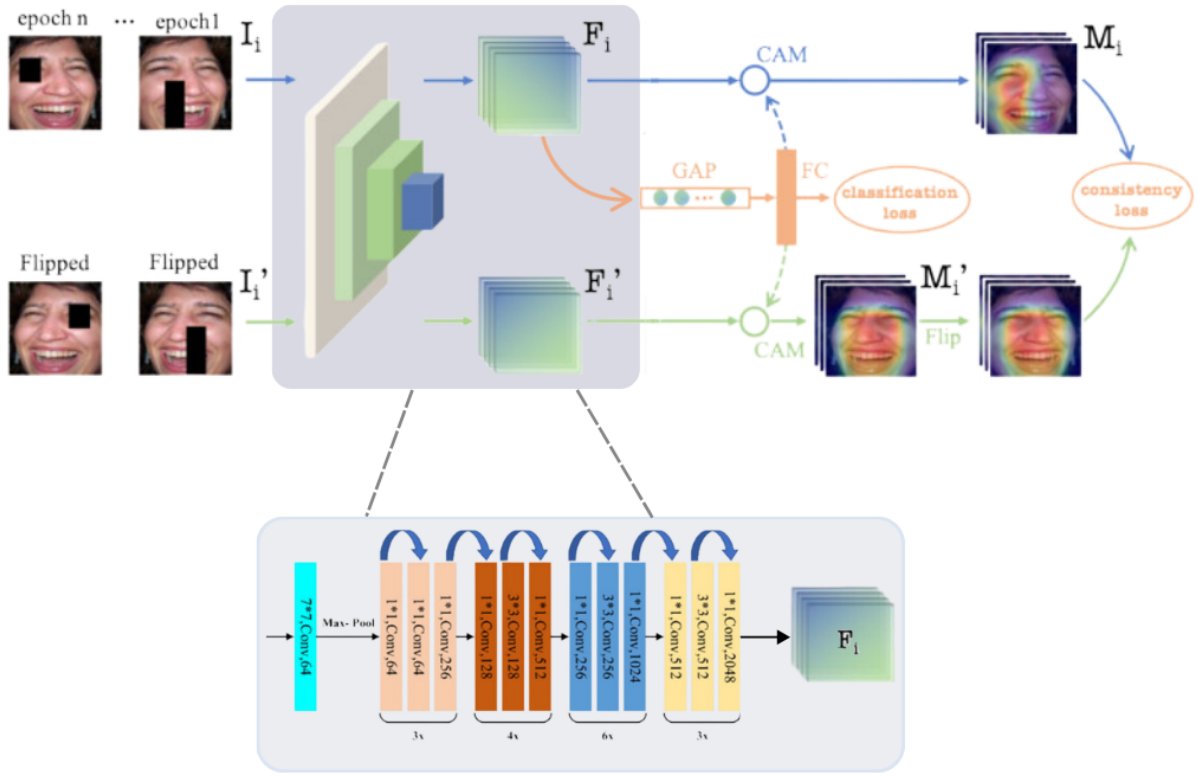


Figure 7: The architecture of the model using Erasing Attention Consistency

The Erasing Attention Consistency (EAC) framework processes both original images and their flipped counterparts through a Convolutional Neural Network (CNN) to extract feature maps. Only the original image features undergo Global Average Pooling (GAP) and pass through fully connected layers for classification. The model employs two key loss components: classification loss and consistency loss. The classification loss is computed only for the original images, using the output from the fully connected layers, guiding the model to correctly categorize inputs. The consistency loss is calculated between the original images and their flipped versions, preventing the model from memorizing noisy labels by encouraging consistent predictions across both versions. EAC incorporates a Class Activation Map (CAM) to generate attention maps for both image sets, highlighting important regions, and employs random erasing as a data augmentation technique. This comprehensive approach aims to improve the model's ability to focus on relevant image areas, handle noisy labels effectively, and enhance overall classification performance by leveraging the relationship between original and flipped images while incorporating attention mechanisms and robust loss strategies.

By random erasing data augmentation and getting **I**, we flip these images to get their flipped counterparts **I'**. **I** and **I'** are the input images. The feature maps are extracted from the last convolutional layer, denoted as $\mathbf{F} \in R^{N\times C\times H\times W}$ and $\mathbf{F}' \in R^{N\times C\times H\times W}$. We only input **F** through the global average pooling **GAP** layer to get features $\mathbf{f} \in R^{N\times C\times H\times W}$.

3.4.2 Loss Fuction

a, Classification Loss:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{W_{y_i}f_i}}{\sum_j e^{W_j f_i}}\right) \tag{5}$$

In above equation, f is the output of F through the global average pooling (GAP) and flatten to be the fully connected (FC) layer. W_{y_i} is the matrix of weight parameters corresponding to y_i in the FC layer. N, C, H, W respectively represent the number of images, channels, height and width of the feature map.

b, Consistency Loss:

To compute the Consistency Loss, we compare the two attention maps of the original image and the augmented image. So we can reduce the problem that the model only focuses on a few features in the image, which is particularly important when dealing with noisy images.

$$M_j(h,w) = \sum_{c=1}^{C} W_{(j,c)} \cdot F_c(h,w) \tag{6}$$

In above equation, $F \in R^{C\times H\times W}$ is the feature map extracted from the last convolutional layer. The weights of the FC layer are denoted as $W \in R^{L\times C}$, where L represents the number of classes. $M_j(h,w)$ is the attention value of location (h,w) for class j, which is the weighted sum of feature maps over different channels.

Note: We compute attention maps M and M' for I and I' according to Eq(6). We use consistency loss to minimize the distance between the feature maps M and $\mathrm{Flip}(M')$ as:

$$l_c = \frac{1}{NLHW}\sum_{i=1}^{N}\sum_{j=1}^{L}\|M_{ij} - \mathrm{Flip}(M'_{ij})\|^2 \tag{7}$$

c, Total Loss Function:
The total loss is computed as follows: $l_{\text{total}} = l_{cls} + \lambda l_c$

4 Experiments

4.1 Datasets

The RAF-DB dataset, annotated by 40 trained human coders, is a valuable resource in emotion recognition research. It includes annotations for both basic and compound facial expressions. Our experiments focused on a subset of this dataset, specifically images depicting seven fundamental expressions: neutral, happy, surprise, sad, angry, disgust, and fear. This subset comprises a total of 12,271 images used for training and an additional 3,068 images reserved for testing.

The annotations provided by the human coders ensure a diverse and reliable dataset, capturing nuanced variations in facial expressions essential for training and evaluating emotion recognition models. This structured approach not only supports the development of accurate algorithms but also facilitates the comparison and benchmarking of different methodologies in the field of computer vision and affective computing. The dataset's size and comprehensiveness make it particularly suitable for exploring various techniques aimed at enhancing the understanding and interpretation of human emotions through facial cues.

4.2 Implementations

In our approach, we leverage the robust RESNET-50 model, pre-trained on the ImageNet1K-V2 dataset, as the foundational backbone of our network architecture. This choice capitalizes on the rich feature representations learned from vast amounts of diverse image data, enhancing the model's capability to extract meaningful features from facial images.

To ensure uniformity and optimize input data quality, facial images are meticulously processed using three-point landmarks for alignment and cropping, followed by resizing to a standardized 224×224 pixel resolution. This preprocessing step aims to minimize variations in facial orientation and scale, thereby facilitating more consistent and effective feature extraction during training and evaluation.

To evaluate the efficacy of our proposed method, we employ a streamlined approach to data augmentation. Specifically, horizontal flipping and random erasing techniques are applied during training. Horizontal flipping enhances the model's ability to generalize by presenting mirrored versions of images, while random erasing introduces controlled noise reduction by randomly occluding parts of the input images. These augmentation strategies collectively enrich the diversity of the training data, helping the model generalize better to unseen facial expressions and variations.

During the training phase, we adopt a batch size of 64, which balances computational efficiency with gradient stability during backpropagation. The training process initiates with an initial learning rate of 0.0002, chosen to facilitate steady convergence towards optimal parameter values. To further optimize training dynamics, we employ the Adam optimizer with a weight decay of 0.0001, promoting regularization and preventing overfitting by penalizing large parameter values.

Additionally, an ExponentialLR learning rate scheduler with a decay factor (gamma) of 0.9 is utilized to systematically reduce the learning rate after each epoch. This adaptive learning rate adjustment strategy helps navigate the training process towards convergence, fine-tuning model parameters more effectively as training progresses over 40 epochs.

4.3 Evaluation Metrics

Accuracy is a popular evaluation metric in the field of machine learning, especially in classification tasks. It measures the model's accuracy by determining the ratio between the number of correct predictions and the total number of predictions. The formula for calculating accuracy is as follows:

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions} \tag{8}$$

In this experiment, the results are computed as the mean of the accuracy from the last 5 epochs:

$$MeanAccuracy = \frac{1}{5}\sum_{i=n-4}^{n} Accuracy_i \tag{9}$$

Where n is the total number of epochs, $Accuracy_i$ is the accuracy for the i_{th} epoch

4.4 Evaluation on RAF-DB

The noise rate of the label on image equal to 10%.

4.4.1 Backbone ResNet50

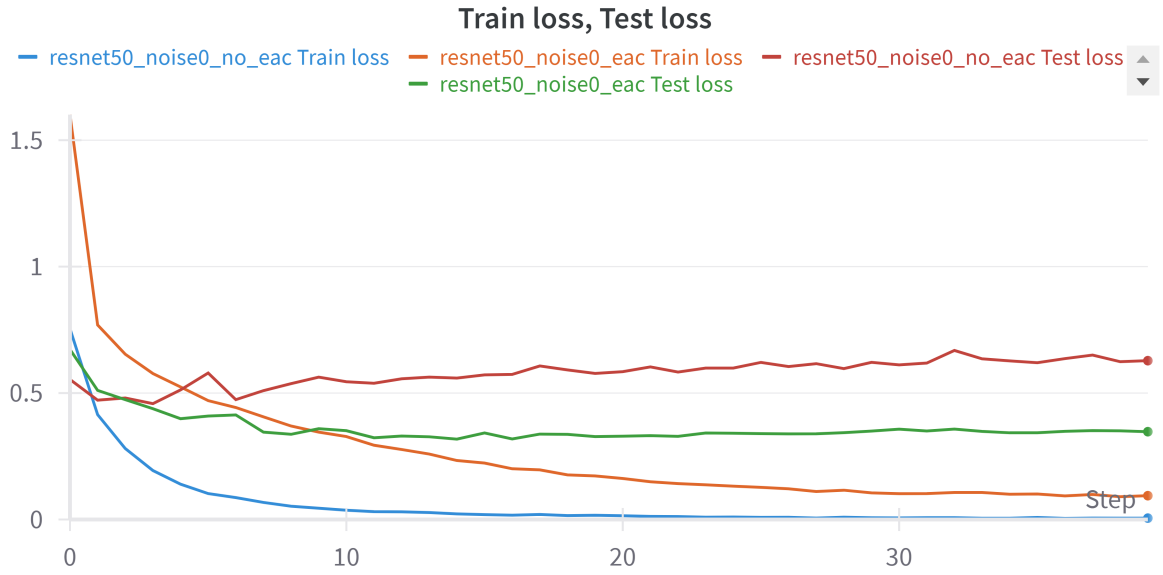


Figure 8: Train and Test Loss

Comparing the training and testing loss patterns of two versions of ResNet50—one with Erasing Attention Consistency (EAC) and one without—reveals significant differences. The original model without EAC experiences a rapid drop in training loss but subsequently suffers from evident overfitting. This is evident as the test loss increases after an initial improvement, indicating the model's difficulty in generalizing beyond the training data. The overfitting likely arises because the model memorizes noisy images with incorrect labels during training, leading to less accurate predictions on unseen test data.

In contrast, the EAC-enhanced model demonstrates a more balanced performance. It shows a slower but more consistent reduction in both training and test losses, suggesting improved generalization ability. While the EAC model may initially have a slightly higher training loss, it achieves notably lower test loss, highlighting its superior performance on unseen data. This suggests that EAC acts as an effective regularizer, helping the model avoid memorizing noisy labels and enhancing its overall robustness.

Furthermore, the improved consistency in performance between training and testing phases in the EAC model underscores its enhanced capability to generalize effectively from training examples to new, unseen instances. By encouraging the model to focus on more reliable features and reducing reliance on potentially incorrect labels, EAC proves to be a valuable technique for enhancing deep learning model performance, particularly in scenarios involving label noise in the training data.

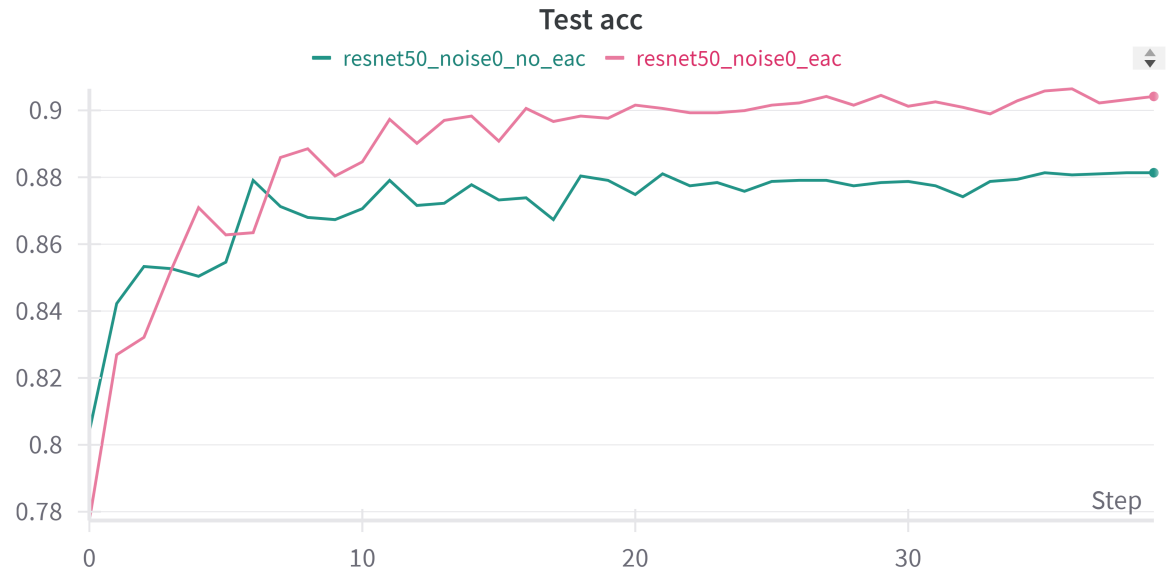


Figure 9: Test Accuracy of each model

Test accuracy curves for the two ResNet50 variants—namely, the original model without Erasing Attention Consistency (EAC) and the model enhanced with EAC—reveal compelling insights into their performance dynamics. Throughout the training process, the EAC-enhanced model consistently outperforms its counterpart in terms of test accuracy. Specifically, it achieves and maintains a higher accuracy level, reaching approximately 90.35%, compared to the original model's 88.27%. This substantial improvement in test accuracy aligns with our earlier observation of lower test loss for the EAC model.

The superior performance of the EAC model can be attributed to its ability to mitigate the memorization of noisy labels during training. By integrating Erasing Attention Consistency, the model appears to prioritize more reliable features and capture the latent truth embedded in the images. This approach contrasts with the original model, which shows signs of overfitting to potentially incorrect labels in the training data, as evidenced by its lower test accuracy and higher test loss.

Moreover, the stability and higher trajectory of the EAC model's accuracy curve further validate its enhanced generalization capability and robustness. In contrast, the accuracy curve of the original model, while maintaining high accuracy, plateaus at a lower level, indicating limitations in generalizing beyond the training dataset.

This comparison underscores the significant benefits of incorporating Erasing Attention Consistency into the ResNet50 architecture. It not only boosts performance on unseen test data but also serves as an effective strategy for addressing challenges posed by label noise in training datasets. The findings emphasize the potential of EAC as a valuable enhancement for deep learning models, particularly in contexts where maintaining robust performance across diverse datasets is crucial.

4.4.2 Backbone MobileNet_V2

The noise rate = 10% and the accuracy of the model achieved: 85.23%

Figure 10: Train and Test Loss

Comparison of the training and test loss curves between two variants of MobileNet_V2—one without Erasing Attention Consistency (EAC) and the other incorporating EAC—reveals nuanced insights into their performance dynamics. The variant without EAC initially demonstrates rapid convergence in training loss, suggesting effective learning on the training data. However, this early success is marred by a subsequent divergence where test loss increases noticeably after an initial period of improvement, indicative of significant overfitting. In contrast, the EAC-enhanced model exhibits a more tempered and consistent reduction in both training and test losses throughout the training process.

Despite exhibiting slightly higher training loss during early epochs, the EAC model consistently maintains lower test loss across subsequent epochs, showcasing its ability to generalize better to unseen data. This phenomenon underscores the role of EAC as a robust regularizer, effectively curbing overfitting tendencies that can compromise model performance on real-world datasets. By promoting a more balanced learning trajectory, EAC enhances the model's adaptability to diverse data distributions and scenarios, thereby bolstering its overall robustness and reliability in practical applications.

Furthermore, the closer alignment between the training and test loss curves in the EAC model signifies its improved generalization capabilities. This alignment suggests that the model's performance on the training data reliably extends to new, unseen examples, affirming EAC's efficacy in enhancing the model's capacity to learn meaningful features and patterns without excessively fitting noise in the training set. Consequently, the EAC-enhanced MobileNet_V2 variant not only mitigates overfitting but also fosters greater confidence in its predictive performance across different domains and deployment environments.

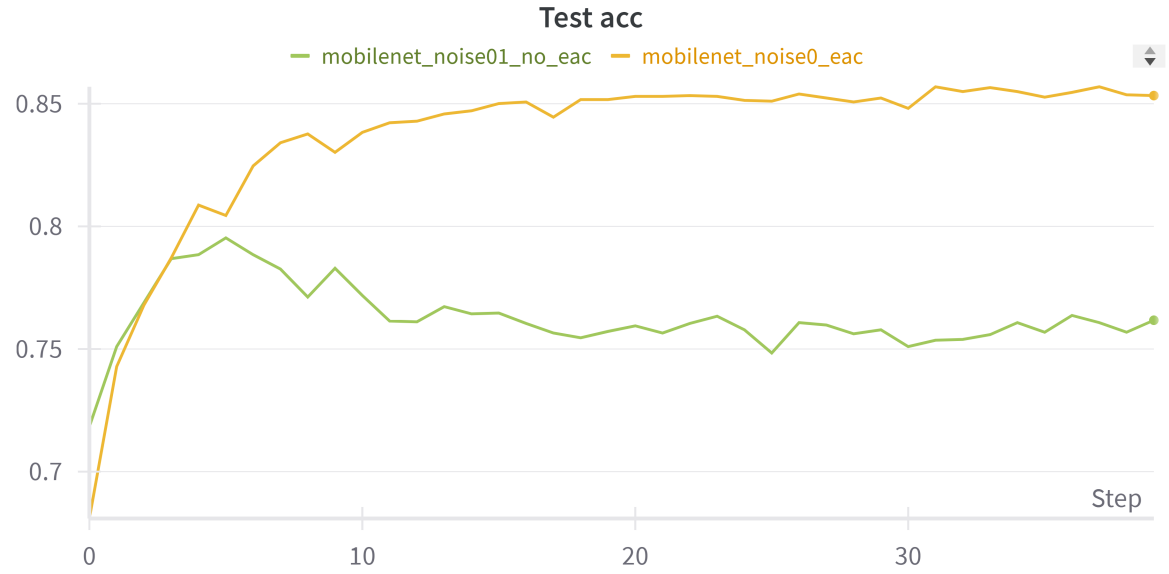


Figure 11: Test Accuracy of each model

Test accuracy curves for the two MobileNet_V2 variants—specifically, the original model without Erasing Attention Consistency (EAC) (mobilenet_noise01_no_eac) and the model enhanced with EAC (mobilenet_noise01_eac)—underscore significant performance differences. The graph not only corroborates but also extends our earlier analysis based on loss curves.

Throughout the training process, the EAC-enhanced model (represented by the orange line) consistently outperforms the original model (green line) in terms of test accuracy. It achieves and maintains a notably higher accuracy level, peaking at 85.47% compared to the original model's 76.40%. This substantial improvement aligns closely with our previous findings of lower test loss for the EAC model, indicating its enhanced ability to generalize from training to test data.

The stability and upward trajectory of the EAC model's accuracy curve further validate its superior generalization capability and robustness. In contrast, the accuracy curve of the original model plateaus at a lower level, underscoring issues of overfitting that were evident in the loss curves.

Overall, this comparison strongly supports the conclusion that incorporating Erasing Attention Consistency significantly enhances the MobileNet_V2 architecture's performance on unseen data. It not only boosts accuracy but also mitigates the impact of potential label noise in the training dataset, making EAC a valuable addition for improving the model's reliability and effectiveness across diverse real-world applications.

4.4.3 General Evaluation

Erasing Attention Consistency (EAC) has demonstrated robustness across various deep learning architectures beyond MobileNet_V2, highlighting its effectiveness as a regularization technique in improving model performance. By examining its impact on different models such as ResNet50 and others, EAC consistently enhances generalization capabilities and mitigates overfitting tendencies.Across different models, EAC consistently delivers several key benefits.

Improved Generalization: EAC helps models generalize better from training to test data by encouraging them to focus on more reliable features and reducing reliance on potentially noisy or incorrect labels in the training set. This results in higher accuracy and lower test loss, as observed in experiments with ResNet50 and MobileNet_V2.

Stability in Training: Models enhanced with EAC often exhibit smoother training curves with less variance in performance metrics such as accuracy and loss. This stability indicates that EAC effectively regularizes the learning process, leading to more robust models.

Reduced Overfitting: EAC mitigates overfitting, a common issue in deep learning where models memorize noise or irrelevant details from the training data. By incorporating attention erasing mechanisms, EAC encourages models to learn more meaningful representations, thereby improving their ability to generalize to unseen data.

Consistent Performance Gains: Experimental results consistently show that models with EAC outperform their counterparts without EAC across various evaluation metrics. This consistency underscores EAC's reliability and its potential as a standard regularization technique in deep learning applications.

Overall, experimental results consistently demonstrate that integrating Erasing Attention Consistency leads to performance gains across a range of deep learning architectures. This consistency underscores EAC's reliability and its potential as a standard technique for improving model robustness and performance in various practical applications of machine learning and computer vision.

4.5 Compare to the State of the Art

Model	Accuracy	Details	Year
S2D	92.57	From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos	2023
ARBEx	92.47	ARBEx: Attentive Feature Extraction with Reliability Balancing for Robust Facial Expression Learning	2023
DDAMFN++	92.34	A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition	2024
EAC (ResNet-50)	**90.35**	**Learn From All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition**	**2022**
Ada-DF	90.04	A Dual-Branch Adaptive Distribution Fusion Framework for Real-World Facial Expression Recognition	2023
C-EXPR-NET	87.5	Multi-Label Compound Expression Recognition: C-EXPR Database & Network	2023
C MT PSR	84.8	Distribution Matching for Multi-Task Learning of Classification Tasks: a Large-Scale Study on Faces & Beyond	2024
C MT VGGFACE	81.4	Distribution Matching for Multi-Task Learning of Classification Tasks: a Large-Scale Study on Faces & Beyond	2024

Figure 12: Compare to State of Art

This table highlights the accuracy and relevant details of some leading models in facial expression recognition from recent years. An overview of the operation mechanism of the model and a comparison to the EAC model would show the following:

EAC (OURS): This method combines the features from a face recognition dataset with deep learning

techniques to enhance the performance of FER. Specifically, the study introduces an efficient convolutional neural network (CNN) that uses an affinity convolution module to reduce computational overhead. Additionally, it employs a deep facial clustering approach to generate expression labels from a face recognition dataset. The CNN is fine-tuned using these labels, resulting in improved accuracy in recognizing facial expressions. The results of this approach on the RAF-DB dataset show significant improvements in FER accuracy. The proposed method achieved a high accuracy rate, demonstrating its effectiveness compared to other existing methods.

C-EXPR-NET: This model is designed for multi-label compound expression recognition. It leverages the C-EXPR Database & Network to handle multiple facial expressions simultaneously, addressing the complexity of compound facial expressions. Although it can analyze multiple emotions and detect moods, EAC offers a more comprehensive solution for applications requiring dynamic interaction and user engagement. The ability of EAC to adapt its conversational strategies based on emotional context makes it a more suitable choice for creating empathetic and engaging user experiences.

C MT VGGFACE: This model employs Distribution Matching for Multi-Task Learning of Classification Tasks, integrating multi-task learning methodologies to improve classification performance. This model aims to enhance the efficiency and accuracy of classification tasks across diverse datasets. Despite its innovative approach to multi-task learning, its performance on RAF-DB indicates potential limitations in handling single-task scenarios compared to more specialized models. The C MT VGGFACE model, despite its innovative multi-task learning approach, may require further refinement to match the single-task performance of models like EAC (ResNet-50) on datasets such as RAF-DB.

DDAMFN++: The framework reduces end-to-end latency by 82-84% with less than a 1% accuracy loss by minimizing feature bias and reducing communication overhead through task-oriented asymmetric feature coding. Compared to the Erasing Attention Consistency (EAC) model, which handles noisy labels in facial expression recognition, this new framework focuses on optimizing model splitting and compression to better utilize limited edge resources. While EAC improves robustness against noisy labels by ensuring attention consistency between original and transformed images, the cloud-edge collaborative method excels in reducing latency and maintaining accuracy, making it more effective for real-time edge deployments.

S2D: Static-to-Dynamic model is designed to enhance facial expression recognition by transitioning knowledge from static (SFER) to dynamic contexts (DFER). It employs static expression features integrated with dynamically encoded information through facial landmark-aware features. The S2D model integrates a Vision Transformer (ViT) with Multi-View Complementary Prompters (MCP) and Temporal-Modeling Adapters (TMAs), which significantly improves performance by merging static expression features with landmark-aware features and modeling dynamic relationships of expression changes over time. This innovative approach allows the model to efficiently extend static image models to video models without retraining all parameters, thus saving computational costs. Achieving an accuracy of 92.21% on the RAF-DB dataset, the S2D model demonstrates superior performance compared to traditional models like EAC-ResNet50, which typically focus on static image analysis. The S2D's method of integrating dynamic and static features provides a more robust framework for handling facial expressions in dynamic scenarios, significantly enhancing its effectiveness and addressing the challenges of limited DFER data while effectively utilizing abundant SFER data.

ARBEx: This model is described as an advanced framework for facial expression learning that employs a Transformer with reliability balancing to enhance label prediction accuracy. Unlike other models like EAC Resnet50, which focus on augmenting performance through enhanced augmentation techniques and deeper ResNet architectures, ARBEx utilizes multi-level feature extraction based on multi-head attention mechanisms and learnable anchor points in the embedding space. This approach not only concentrates on robust feature extraction but also on enhancing the reliability of label predictions. ARBEx's unique focus on reliability balancing helps mitigate bias and uncertainty, allowing it to handle uneven and ambiguous data more effectively. The framework has demonstrated superior performance across various databases, significantly outperforming EAC Resnet50 and other advanced methods due to its ability to deal with diverse and challenging datasets effectively.

5 Conclusion

This project has explored the application of the Erasing Attention Consistency (EAC) method within the field of Facial Emotion Recognition (FER), demonstrating significant strides in addressing the persistent challenges of noisy data and high inter-class similarity. Through our detailed implementation of EAC in both ResNet50 and MobileNet_V2 architectures, we have evidenced the method's ability to enhance model robustness and accuracy by promoting the consistency of important facial features across manipulated image inputs.

The results from our empirical studies confirm that EAC effectively mitigates the adverse effects of noisy labels and improves overall FER accuracy. By comparing the performance of ResNet50 and MobileNet_V2 under the EAC framework, it is clear that this approach not only boosts performance in typical environments but is also robust in noise-prone settings.

In conclusion, the Erasing Attention Consistency approach marks a notable advancement in the development of facial emotion recognition systems. Its application can be seen as a valuable contribution to both the academic and practical aspects of FER, providing a reliable tool for enhancing human-computer interaction and supporting psychological studies. Future work will focus on further refining this technique and exploring its integration with other computational models to broaden its applicability and effectiveness in real-world scenarios.

References

[Che+23] Yin Chen et al. *From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos*. 2023. arXiv: 2312.05447 [cs.CV].

[He+15] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[Kol23] Dimitrios Kollias. "Multi-Label Compound Expression Recognition: C-EXPR Database & Network". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 5589–5598.

[KSZ24] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. *Distribution Matching for Multi-Task Learning of Classification Tasks: a Large-Scale Study on Faces & Beyond*. 2024. arXiv: 2401.01219 [cs.CV].

[LDD17] Shan Li, Weihong Deng, and JunPing Du. "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2584–2593. DOI: 10.1109/CVPR.2017.277.

[Liu+23] Shu Liu et al. "A Dual-Branch Adaptive Distribution Fusion Framework for Real-World Facial Expression Recognition". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097033.

[San+19] Mark Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: 1801.04381 [cs.CV].

[Was+23] Azmine Toushik Wasi et al. *ARBEx: Attentive Feature Extraction with Reliability Balancing for Robust Facial Expression Learning*. 2023. arXiv: 2305.01486 [cs.CV].

[Zha+22] Yuhang Zhang et al. *Learn From All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition*. 2022. arXiv: 2207.10299 [cs.CV].

[Zha+23] Saining Zhang et al. "A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition". In: *Electronics* 12.17 (2023). ISSN: 2079-9292. DOI: 10.3390/electronics12173595. URL: https://www.mdpi.com/2079-9292/12/17/3595.

[Zho+15] Bolei Zhou et al. *Learning Deep Features for Discriminative Localization*. 2015. arXiv: 1512.04150 [cs.CV].