HUST

Introduction to Artificial Intelligence

Lecturer: Prof. Dr. Hai Van Pham

Class ID: 147838

Project Proposal
Facial Emotion Recognition

Members:
Le Dai Lam - 20225982
Nguyen Minh Khoi -20226050
Vu Hai Dang -20225962
Hoang Khai Manh -20225984
Hoang Quoc Hung -20226043

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

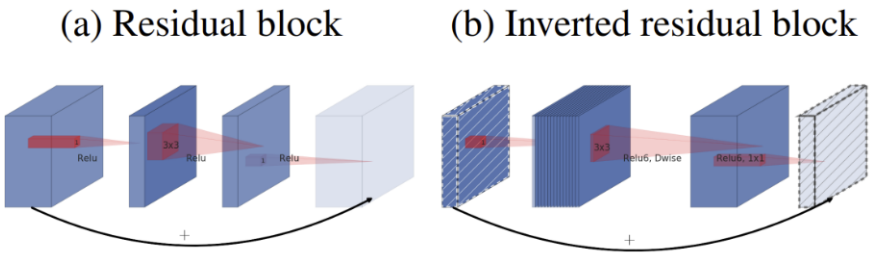
Class Activation Mapping: Class Activation Mapping (CAM) is an attention method, which allows us to visualize the predicted class scores on the given images, highlighting the discriminative parts detected by the CNN.

In the CNN trained for classification, an attention map is the weighted sum of the feature maps from the last convolutional layer with the weights from a fully connected (FC) layer. By viewing the attention maps, we can know what the model is based on to make the predictions

Erasing Attention Consistency: Facial Expression Recognition (FER) in the presence of noisy labels poses significant challenges due to the high inter-class similarity and annotation ambiguity inherent in facial expression images. Traditional methods for handling noisy labels in FER typically fall into two categories: sample selection and label ensembling.

Our approach builds upon the principles of attention consistency and introduces a novel framework called Erasing Attention Consistency (EAC). Unlike traditional methods that rely on loss-based sample selection, EAC uses feature-learning to mitigate the effects of noisy labels. By erasing parts of input images and utilizing flip semantic consistency, EAC forces the model to focus on consistent features across original and flipped images. This prevents the model from overfitting to noisy samples without the need to know the noise rate or train multiple models.

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

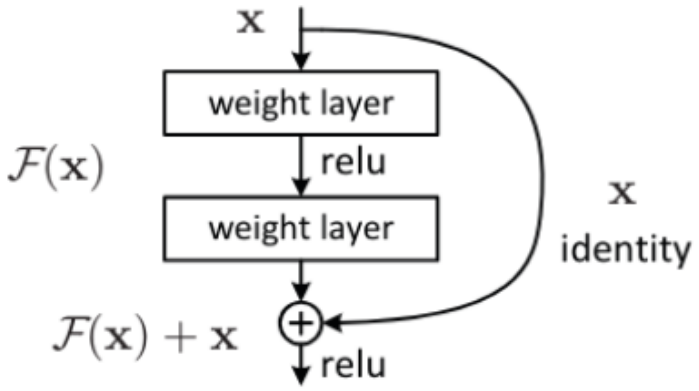


Figure 1: Residual learning: a building block

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 philosophy of VGG nets, the vanishing gradient problem when building a deeper plain network and the advantage of residual block, the authors insert skip connections between layers

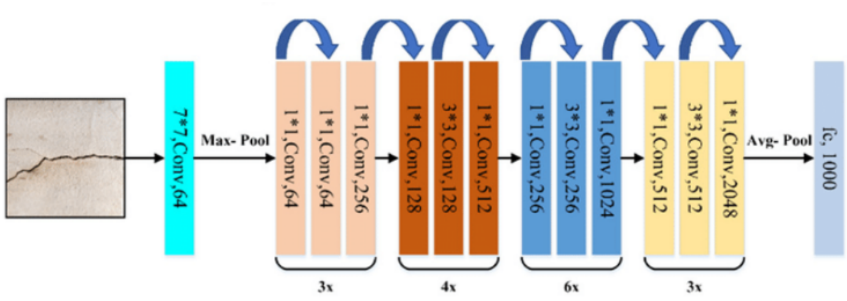


Figure 2: The architecture of ResNet50

3.1.3 Why do Residual Networks work?

When constructing increasingly deeper neural networks, the gradients in the deeper layers can become extremely small, eventually approaching zero. This phenomenon is known as the vanishing gradient problem.

By using shortcut connections that add the output of a previous layer to the activation function of a subsequent layer. We can prevent the vanishing gradient problem, ensuring more effective training of the neural network and enhancing the model accuracy.

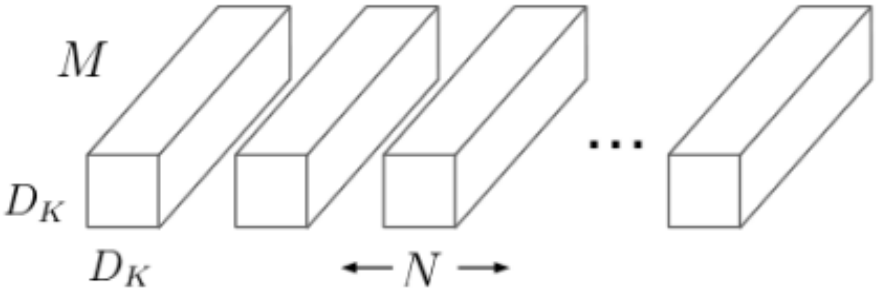


Figure 3: Standard Convolution Filters

3.2 MobileNet-V2 Architecture

3.2.1 Depthwise Separable Convolution

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.

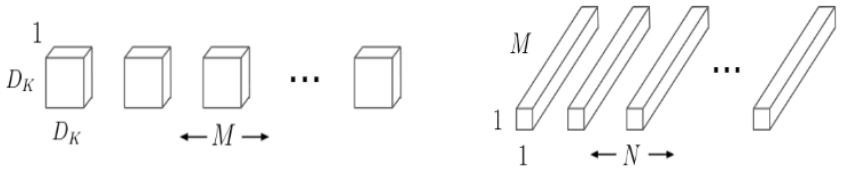


Figure 4: Pointwise Convolution

By extracting Standard Convolutions into:

- Depthwise Convolution with computational cost:

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F \quad (4)$$

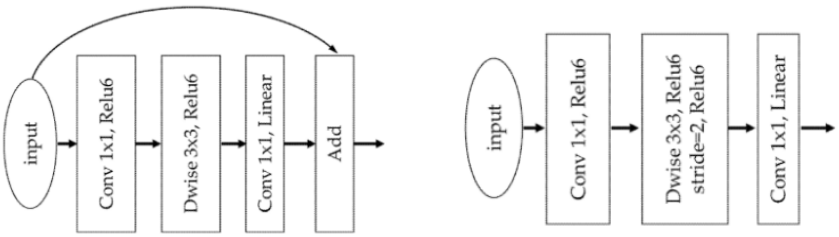
- And Pointwise Convolution with the cost of computation:

$$M \cdot N \cdot D_F \cdot D_F \quad (5)$$

Combining two processes (Depthwise Separable Convolution) we get a reduction in computation cost:

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

3.2.2 Inverted Residual Block



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

Figure 5: The inverted residual block in MobileNet-v2

It has two structures:

- Stride = 1 with the shortcut connections between bottlenecks when the dimension of input and output are same
- Stride = 2 for downsampling the size of the feature map.

3.2.3 MobileNet-V2 Structure

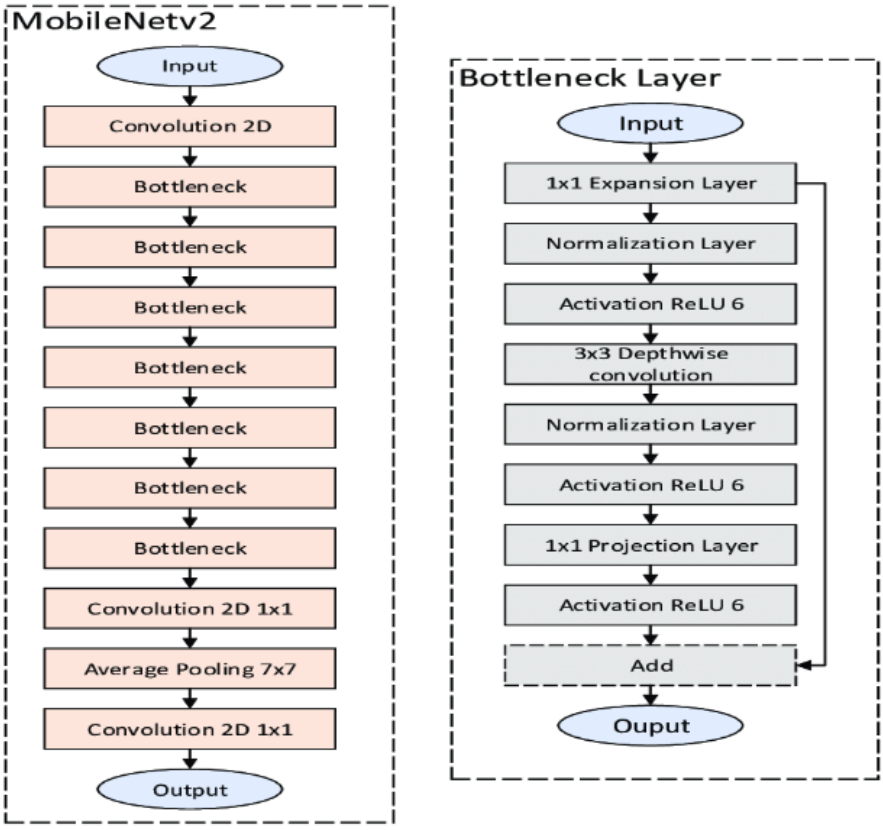


Figure 6: The architecture of MobileNet-V2

3.3 FrameWork and Training

3.3.1 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]. Initially, EAC randomly erases input images and then gets their flipped counterparts.In the next step, EAC only computes the classification loss with the original images. EAC uses the consistency loss between the original images and their flipped counterparts to prevent the model from remembering noisy labels.

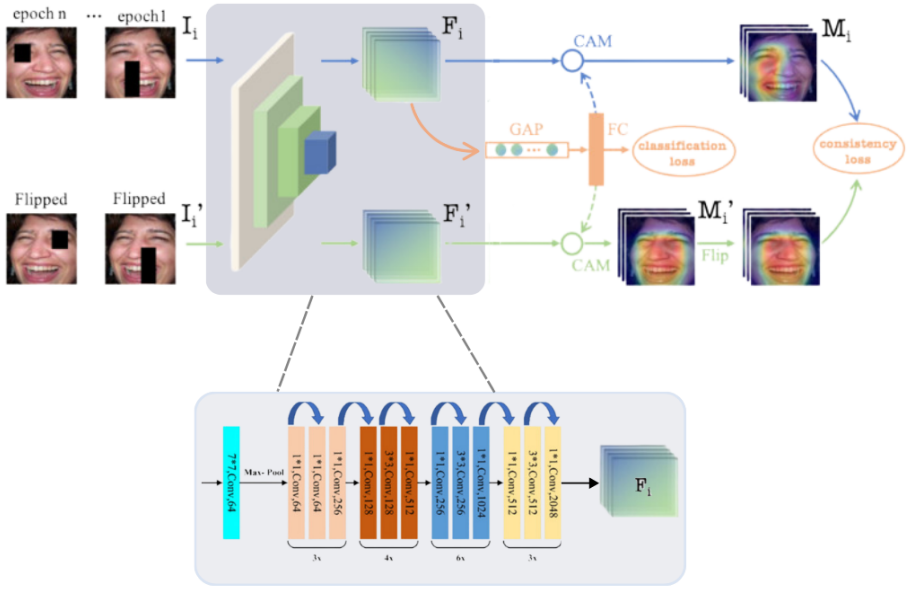


Figure 7: The architecture of the model using Erasing Attention Consistency

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 $\mathbf{f} \in R^{N\times C\times H\times W}$.

3.3.2 Loss Fuction

a, Classification Loss:

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

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.

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

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. (1). We use consistency loss to minimize the distance between the feature maps M and $\text{Flip}(M')$ as:

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

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 is annotated with basic and compound expressions by 40 trained human coders. For our experiments, we used images depicting seven basic expressions: neutral, happy, surprise, sad, angry, disgust, and fear. This includes 12,271 images for training and 3,068 images for testing

4.2 Implementations

We use the model RESNET-50 pretrained on ImageNet1K-V2 as the backbone of our network. The facial images are aligned and cropped using three landmarks and resized to 224×224 pixels. To assess the effectiveness of our proposed method, we only apply horizontal flipping and random erasing for data augmentation. During training, we use a batch size of 64 and start with an initial

learning rate of 0.0002. The Adam optimizer, with a weight decay of 0.0001, is utilized alongside an ExponentialLR learning rate scheduler with a gamma of 0.9 to reduce the learning rate after each epoch. The training process concludes at epoch 40.

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:

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions} \tag{4}$$

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

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

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

4.4 Evaluation on RAF-DB

Consider the noise rate of the label on image equal to 10%

4.4.1 Backbone ResNet50

Figure 8: Train and Test Loss

In the training process the RESNET-50 model remember the noisy image with the wrong label and do not use the latent truth of the image. Then it makes the prediction of the model on the test set less accurate.

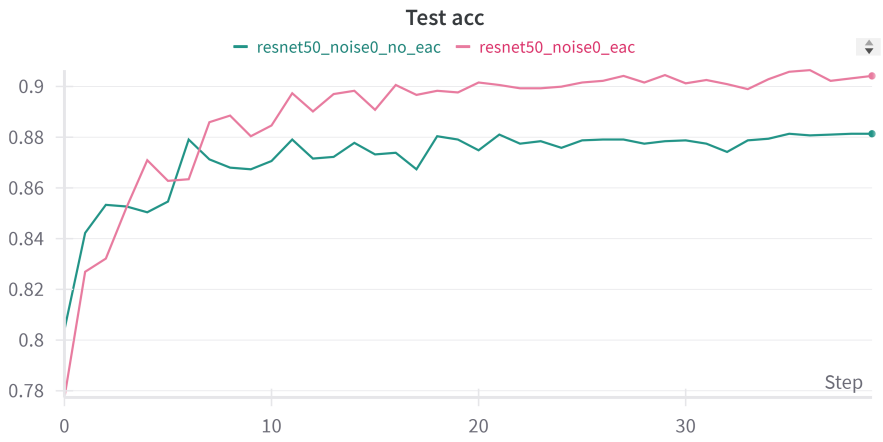


Figure 9: Test Accuracy of each model

Compare to our method - EAC, the model are looking on the image by a comprehensive view, using more features to make the prediction on the image, especially the noisy images (since the difference between the CAM added to the loss function).

4.4.2 Backbone MobileNet_V2

The noise rate = 10% and the accuracy of the model achieved: 88.23%

Figure 10: Train and Test Loss

The overfitting on noisy image in the train dataset in origin model MobileNet-V2 that lead to low accuracy on test dataset.

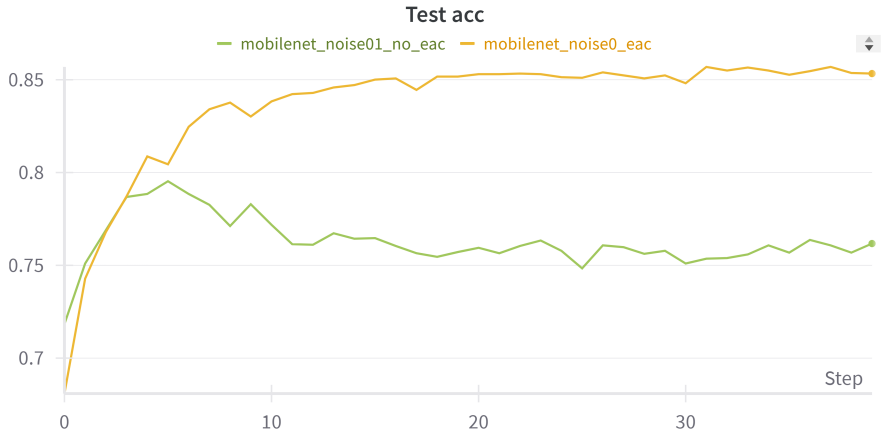


Figure 11: Test Accuracy of each model

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].