## Leaving Some Facial Features Behind

Cheng Qiu Vanderbilt University

#### Abstract

Facial expressions are crucial to human communication, offering insights into emotional states. This study examines how specific facial features influence emotion classification, using facial perturbations on the Fer2013 dataset. As expected, models trained on data with the removal of some important facial feature experienced up to an 85% accuracy drop when compared to baseline for emotions like happy and surprise. Surprisingly, for the emotion disgust, there seem to be slight improvement in accuracy for classifier after mask have been applied. Building on top of this observation, we applied a training scheme to mask out facial features during training, motivating our proposed Perturb Scheme. This scheme, with three phases—attention-based classification, pixel clustering, and feature-focused training, demonstrates improvements in classification accuracy. The experimental results obtained suggests there are some benefits to removing individual facial features in emotion recognition tasks.

## 1 Introduction

Understanding emotions plays a pivotal role in how we perceive and interact with others. When interpreting facial emotions, key landmarks such as the eyes and mouth provide significant insights into a person's emotional state [2]. Additionally, faces can be divided into two halves, focusing on the eyes and eyebrows on one side and the mouth on the other [8]. To enhance our comprehension of facial emotions, it is essential to understand how these facial landmarks contribute to emotion prediction.

To investigate the impact of facial landmarks on emotion classification, we introduced perturbations such as a physical mask to Fer2013, a facial emotion dataset, through similar procedure to that of [3]. In this paper, the masked facial emotion dataset will be referred to as MaskFer<sup>1</sup>. Models were trained on both Fer2013 and MaskFer. The performance results for accuracy are presented in Table 1. The accuracy ratio for each emotion class generally decreased for model trained on MaskFer when compared to model trained on Fer2013. This is to be expected because by applying a mask, important features are removed from the context, resulting in less information models to determine the emotion. However, the percent change in accuracy ratio varied across different classes. For example, the accuracy for emotions like happiness dropped by around 60% while for fear, there was only a slight drop of around 10%. Surprisingly, for the emotion sad, there's even a slight increase in accuracy ratio which

<sup>&</sup>lt;sup>1</sup>Dataset can be accessed at https://github.com/chengq220/MaskFerDataset

Emotion	[5]	[1]	[9]	[4]	[12]	Ensemble
Angry	-39%	-35%	-32%	-20%	-23%	-32%
Happy	-49%	-63%	-42%	-55%	-35%	-42%
Sad	-12%	-24%	-4%	-16%	-8%	-4%
Neutral	-51%	-60%	-53%	-60%	-43%	-53%
Surprise	-19%	-26%	-35%	-19%	-20%	-35%
Disgust	-49%	-59%	-39%	-49%	-49%	-39%
Fear	-20%	-23%	-13%	-18%	-17%	-13%

means that the model trained on MaskFer performed better than its counterpart trained on Fer2013. For this specific emotion, we hypothesized that masking the mouth allowed the model to extract better information relevant to the emotion of sad such as the eyebrow.

Table 1: Percent accuracy gain/loss for each emotion class with models trained on MaskFer compared to models trained on FER2013

By examining the saliency map shown in Figure 1, it becomes evident that for emotions such as disgust, anger, and sadness, the model trained on Fer2013 may not have learned the optimal salient features. In contrast, the model trained on MaskFer was able to capture the features in the eyebrows when identifying an angry face. It might be case that models trained on Fer2013 do not effectively leverage all the information present in the vicinity of major facial landmarks like the eyebrows. While classifying emotions based on overall facial landmarks may provide a global context that's beneficial to emotion classification, considering them as individual features provide better localized context that's better suited for classification of specific emotions.



Figure 1: Saliency map for trained models on Fer2013 and MaskFer

## 2 Related Works

Neural networks have become a natural choice for facial emotion recognition tasks due to their ability to learn complex representations from visual data. With the rise of deep learning, networks like ResNet [4] have paved the way for deeper architectures by introducing skip connections, which allow information to flow across layers without degradation. This innovation enables deeper networks that can learn richer features, albeit with increased computational complexity. DenseNet [5] later addressed this issue by establishing direct connections between layers, effectively reducing the number of parameters while maintaining performance, which is crucial for real-time applications.

Recent advances have also seen the development of the Dual Path Network (DPN) [1], which combines the advantages of both ResNet's and DenseNet's connectivity patterns. By incorporating a dual-path structure, DPN allows features to be reused efficiently across layers, further enhancing classification accuracy while managing computational load. The FER2013 benchmark dataset has been widely used to train and evaluate facial emotion recognition models. Many studies have built upon convolutional neural networks (CNNs) to improve baseline models. For example, the addition of attention mechanisms has enabled models to selectively focus on significant facial areas, such as the eyes and mouth, enhancing the interpretability and accuracy of emotion recognition [7]. Attention mechanisms, inspired by human visual perception, help networks prioritize crucial regions that are most indicative of emotions, particularly under challenging conditions where parts of the face might be occluded or not visible.

Additionally, researchers have experimented with hyper-parameter tuning using genetic algorithms to optimize hyper-parameters of model such as VGG architecture. This approach improved the FER2013 classification accuracy by refining aspects like kernel size and learning rate, ultimately achieving a single-network classification accuracy of 76% [11]. Spatial transformations have also been applied to increase robustness against rotations, translations, and other variations, which are common in real-world scenarios [2].

Despite these improvements, emotion recognition models still face substantial challenges, particularly when deployed in noisy, ambiguous, or complex environments. Issues such as inconsistent annotations and the diverse real-world backgrounds can impact performance significantly. Furthermore, in applications where parts of the face are occluded, like mask-wearing scenarios, critical facial landmarks might be hidden, presenting additional challenges. Emotion classification under partial face occlusion requires models to make predictions based on incomplete information, often resulting in lower accuracy.

To address these challenges, researchers have turned to techniques like transfer learning, where models pre-trained on large, general-purpose datasets are fine-tuned on smaller, taskspecific datasets. This approach has shown promising results, as pre-trained models can leverage learned representations from a broad data distribution to perform well on specific tasks, even with limited data [6]. Moreover, the introduction of new datasets designed to include occlusions, such as MaskFer, enables models to better generalize to masked or partially visible faces. MaskFer, for instance, includes images with the lower part of the face masked, simulating scenarios in which only the upper facial features are visible. Research utilizing CNNs in these scenarios aims to identify and adapt to visible facial landmarks, acknowledging that masking impacts emotion recognition by obscuring expressive regions like the mouth and nose [8].

Innovative approaches specifically for masked faces have leveraged landmark-based and low-light image enhancements. One study applied histogram-based regional representation and boundary analysis to focus on the upper face, using AffectNet, a comprehensive dataset containing over 420,000 images across eight emotion classes. The method involved covering the lower half of the face with synthetic masks, then extracting features from the head and upper facial regions, such as the eyes and eyebrows. By combining facial landmark detection with histograms of oriented gradients (HOG), the approach achieved effective classification with CNNs, even under occlusion[8].

## Input Attention Attention Classifier Attention Map K clustering Output "Neutral" Predictor

## 3 Methodology

Figure 2: Framework for Perturb Scheme

To fully leverage the regional characteristic in facial recognition, we propose a new scheme, the Perturb Scheme, shown in Figure 2 which includes threes phases. The first phase focuses on training an attention classifier to better pin point the important pixels. The second phase involve clustering those pixels into classes base on importance. Thirdly, training a new classifier to perform the emotion classification task with the inputs from the clustering. Refer to Figure.

Attention Classifier: The first phase is training a neural network focused on understanding the spatial attention in order to better understand the influence of each pixel toward the prediction of all emotion. Since different facial landmarks may have different effects on the performance of emotion predictions. To achieve this, spatial attention modules were added between layers of the neural networks [13].

**Clustering:** The second phase is to isolate local patches of pixels that have the highest attention. To identify those patches, we will use K-mean clustering with n clusters to represent n meaningful facial landmarks. By leveraging the attention, pixels is clustered into different classes to be masked out based on the euclidean distance shown in Equation 1. Due

to the computation required during clustering, it's infeasible to perform clustering for each image, instead the cluster updates every epoch.

Given an  $I \in \mathbb{R}^{BXCXWXH}$  and its spatial attention map  $S \in \mathbb{R}^{BXWXH}$ , the distance between any two points on the 2D grid  $P_{ij}$  and  $P_{kl}$ , where  $i, k \in W, j, l \in H$ . The Euclidean distance can be computed as shown in Equation 1 where  $\lambda$  and  $\alpha$  are constants used to control the weights of the pixel distance on the grid and the intensity distance.

Dist = 
$$\sqrt{(\lambda(i-k))^2 + (\lambda(j-l))^2 + (\alpha(I_{ij} - I_{jl}))^2}$$
 (1)

**Predictor Training:** The third phase is training a new network to learn from the clustered data in order to optimize the emotion detection task.

### 4 Experiment

We empirically demonstrate the effectiveness of the Perturb Scheme in enhancing the capabilities of deep learning models for emotion recognition. Our performance comparisons include prominent deep learning architectures such as DenseNet, ResNet, and others.

### 4.1 Dataset

FER2013 is a publicly available well-known dataset for facial emotion recognition. The dataset consists of 35887 gray scaled images of emotions The FER2013 dataset, a well-known publicly available resource for facial emotion recognition, consists of 35,887 gray scale images across seven different emotion classes. Of these, 28,708 images are used for training, while 3,589 images are designated for validation and testing, making the dataset suitable for deep learning applications. However, FER2013 contains incorrectly labeled faces, irrelevant images, and an imbalanced distribution of image data, with some classes containing up to 7,000 images and others as few as 500 (see Figure 3). Do take note that since MaskFer is directly built on top of FER2013, similar issues exist in MaskFer.



Figure 3: Distribution of classes in Fer2013 dataset

## 4.2 Training

All models were trained on an NVIDIA RTX 3090 for 200 epochs using cross-entropy loss as the primary loss function. Stochastic Gradient Descent (SGD) with Nesterov momentum (0.9) and weight decay was employed. The initial learning rate was set to 0.01 and adjusted during training using ReduceLROnPlateau (RLRP) [10]. For data preparation, we adopted the image augmentation techniques recommended by Zhong et al., including random flipping, cropping, and erasing, to prevent over-fitting, especially given the scarcity of samples in some classes [14]. For the Perturb Scheme, the parameters  $\alpha$  and  $\lambda$  in Equation 1 were set to 1.2 and 1.5 respectively with the number of clusters for the k-mean clustering set to 3.

# 5 Evaluation and Analysis

For evaluation, all models were trained on FER2013 and tested using the testing set from the same dataset. We selected several models commonly used for emotion classification, including ResMasking [9] and ResNet [4], as well as baseline models like VGG16 [12], to assess the performance of the proposed Perturb Scheme. Our results demonstrate that the proposed method does yield positive outcomes.

## 5.1 Attention-Based Clustering



Figure 4: Facial image after applying clustering n = 3 based on the attention

In spatial attention-based emotion classification, our models primarily focus on localized regions around the eyes and mouth, as highlighted by the attention classifier. This approach underscores the importance of these facial areas, with two clusters targeting the eyes and one focusing on the mouth and nose region. As shown in Figure 5, we experimented with varying the number of clusters, finding that three clusters yielded optimal performance by isolating the most relevant features for emotion recognition. It matches up with the intuition that generally on face there's 3 salient landmarks which are the two eyes and the mouth. Furthermore, adding more clusters increased computational demands without consistent improvements in accuracy, and in some cases, led to performance decline.

## 5.2 Performance

Limited by computation resources, only a selected few of the proposed architecture were trained using Perturb Scheme. The models trained using Perturb scheme generally performed



Figure 5: Performance for different number of clusters for the VGG16 architecture

2.1 Summary (Baseline)			2.2 Summary (Perturb)			
Model	Accuracy	$\mathbf{F1}$	Model	Accuracy	$\mathbf{F1}$	
[5]	72.6%	0.70	[5]	—	_	
[1]	71.2%	0.68	[1]	71.9%	0.69	
[9]	66%	0.63	[9]	—	_	
[4]	69.2%	0.685	[4]	72.3%	0.70	
[12]	66.5%	0.653	[12]	69.5%	0.66	
Ensemble	73.4%	0.71	Ensemble	—	_	

Table 2: Table 2.1 and Table 2.2 illustrate the performance comparison between classifier trained without and with Perturb Scheme

better in accuracy cross all emotion class as shown in Table 3. It further suggests that removing facial landmarks such as mouth and eyes in training did contribute to the increase in accuracy. Comparing the saliency map for the baseline model and the model trained using proposed method shown in Figure 6, it can be observed for emotions such as fear, the new models are able to put more emphasis on the important features such as the mouth and less emphasis on less important features like the nose.

## 6 Conclusion

Our findings demonstrate that incorporating regional feature extraction through the Perturb Scheme significantly enhances emotion recognition model performance, especially under occlusion conditions. By applying attention-based clustering and subsequently training on

Emotion	[5]	[1]	[9]	[4]	[12]	Ensemble
Angry	_	0.8%	_	3.5%	12%	_
Happy	_	2.0%	—	1.0%	-1%	_
Sad	—	1.5%	_	1.5%	-17%	_
Neutral	_	1.4%	—	-0.6%	2.2%	_
Surprise	_	-0.5%	—	2.1%	1.6%	_
Disgust	—	-3.6%	—	5.6%	5.0%	—
Fear	_	3.0%	_	8.4%	18%	_

Table 3: Changes in performance for each emotion class with the perturb scheme on FER2013 dataset  $% \mathcal{F}(\mathcal{F})$ 



Figure 6: Comparison of the saliency map between VGG16 and VGG16 with Perturb Scheme (VGG16 in row 2 and VGG16 with Perturb Scheme in row 3)

clustered data, our approach enables models to prioritize the most informative facial landmarks, leading to marked improvements in predictive accuracy across most emotion classes. Notably, the slight decline in accuracy for emotions like disgust suggests that additional refinements may be beneficial for capturing subtle expressions tied to specific regions of the face.

The Perturb Scheme's ability to enhance model robustness against occlusions underscores the value of regional feature emphasis in facial emotion recognition tasks. This approach has potential applications in real-world settings where full facial visibility cannot be guaranteed, such as masked environments or low-light conditions. These findings pave the way for future work exploring more nuanced feature emphasis strategies and adaptive techniques to optimize model accuracy across diverse expressions and contexts, contributing to the advancement of facial emotion recognition technologies.

## References

- Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks.
- [2] Yassine El Boudouri and Amine Bohi. EmoNeXt: an adapted ConvNeXt for facial emotion recognition. In 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), pages 1–6. ISSN: 2473-3628.
- [3] Bin Han, Hanseob Kim, Gerard Jounghyun Kim, and Jae-In Hwang. Masked FER-2013: Augmented dataset for facial expression recognition. In 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pages 747–748. IEEE.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition.
- [5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks.
- [6] Zi-Yu Huang, Chia-Chin Chiang, Jian-Hao Chen, Yi-Chian Chen, Hsin-Lung Chung, Yu-Ping Cai, and Hsiu-Chuan Hsu. A study on computer vision for facial emotion recognition. *Scientific Reports*, 13(1):8425, May 2023. Publisher: Nature Publishing Group.
- [7] Shervin Minaee and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network.
- [8] Mukhriddin Mukhiddinov, Oybek Djuraev, Farkhod Akhmedov, Abdinabi Mukhamadiyev, and Jinsoo Cho. Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people. 23(3):1080. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
- [9] Luan Pham, The Huynh Vu, and Tuan Anh Tran. Facial expression recognition using residual masking network. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 4513–4519. IEEE.
- [10] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: State of the art.
- [11] Muhammad Sam'an, Safuan Safuan, and Muhammad Munsarif. Convolutional neural network hyperparameters for face emotion recognition using genetic algorithm. 33(1):442–449. Number: 1.
- [12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for largescale image recognition.
- [13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module.

[14] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation.