GIMEFIVE: Towards Interpretable Facial Emotion Classification

Jiawen Wang   Leah Kawka

{Jiawen.Wang,Leah.Kawka}@campus.lmu.de

Abstract

Deep convolutional neural networks have been shown to successfully recognize facial emotions for the past years in the realm of computer vision. However, the existing detection approaches are not always reliable or explainable, we here propose our model GiMeFive with interpretations, i.e., via layer activations and gradient-weighted class activation mapping. We compare against the state-of-the-art methods to classify the six facial emotions. Empirical results show that our model outperforms the previous methods in terms of accuracy on two Facial Emotion Recognition (FER) benchmarks and our aggregated FER GiMeFive. Furthermore, we explain our work in real-world image and video examples, as well as real-time live camera streams. Our code and supplementary material are available at https://github.com/werywjw/SEP-CVDL.

1. Introduction

In recent years, facial recognition technology has matured significantly. As an important branch of facial recognition, *Facial emotion recognition* (FER) [13, 16, 30, 55] has received growing attention from researchers. This progress is largely driven by the effective utilization of *convolutional neural networks* (CNNs). Originally introduced for image classification, CNNs have proven their ability to capture spatial hierarchies, making them ideal for analyzing facial features and expressions. Since the pioneering work by Le-Cun et al. [21], CNNs have significantly improved, fueled by larger datasets, powerful GPU implementations, and innovative regularization strategies. In the era of deep learning and *artificial intelligence* (AI), one major focus in the research community has been on designing network architectures and objective functions towards discriminative feature learning [11, 27, 54]. Simultaneously, there is a strong demand from researchers and general audiences to interpret its successes and failures [10], and to understand, improve, and trust the generated decisions by these models. This has led to the development of various tools for visualizing CNNs and dissecting their prediction paths to identify the important visual cues they rely on [34].
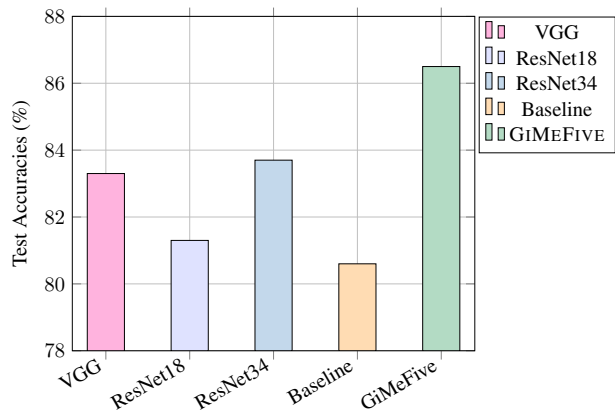

Figure 1. Test accuracies (%) of our GIMEFIVE compared to other state-of-the-art models on the RAF-DB dataset (see Tab. 2 for full results).

Facial expressions, a cornerstone of non-verbal communication, hold significant weight in human interactions. Leveraging *explainable AI* (XAI) [30, 34, 55], we shed light on the decision-making process behind our emotion classifier, enhancing its transparency. In this report, we present our best model named GIMEFIVE, besides several deep CNNs to detect and interpret six basic universally recognized and expressed human facial emotions, namely happiness, surprise, sadness, anger, disgust, and fear. The goal of our research is to build a robust and interpretable model for facial emotion recognition tasks.

Our main contributions can be summarized as follows.

- We collect, preprocess, and evaluate the training and testing data (both images and videos) from various public databases thoroughly.
- We implement all classification models from scratch and optimize them with several techniques in a systematic manner. Meanwhile, we illustrate the classification scores with the confusion matrix heat map.
- We give several video demos to show the real-world performance of our best model GIMEFIVE.
- We provide qualitative benefits such as interpretability to explain our model with gradient-weighted class activation mapping and face landmarks.
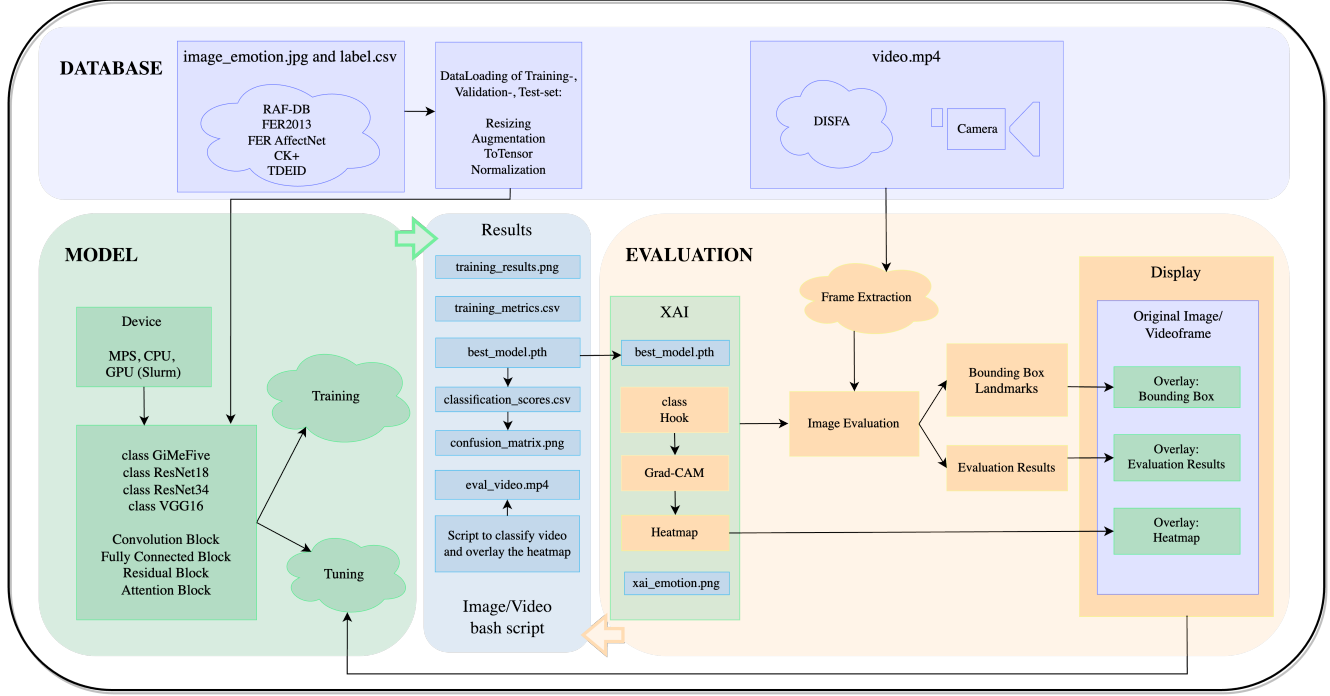
Figure 2. Overview of the experimental pipeline.

Paper Outline. The structure of the rest of the report is arranged as follows. Section 2 contains the related work of our research. In Section 3, we address the datasets we collected and the model architecture we implemented. An overview of the experimental pipeline of our project is outlined in Figure 2. The evaluation results of our models are given in Section 4 with interpretability. The test accuracy on the RAF-DB dataset is shown in Figure 1. Section 5 describes the optimization strategies such as data augmentation and hyperparameter tuning. We provide the conclusion and discussion in Section 6.

2. Related Work

Facial Emotion Recognition. Studies from Ekman and Friesen [8] and Li and Deng [24] categorize facial expressions into distinct states such as happiness, sadness, or anger, demonstrating their cross-cultural recognition. While emotions can intertwine in specific situations, primary emotions remain broadly identifiable due to their innate nature. Facial emotion recognition capitalizes on this understanding by classifying the emotional states of individuals based solely on their facial expressions. This field has witnessed a shift from hand-crafted features to sophisticated deep learning architectures, achieving remarkable accuracy on benchmark datasets [9, 26]. Given the widespread use of video streams and FER applications, real-time operation is crucial. However, this must be bal-

anced with fairness, robustness, and scalability, aspects often overlooked in recent research [40]. Wang et al. [53] proposed a benchmark for bias mitigation, while Pham et al. [37] presented a novel masking idea to boost the performance on the FER2013 [1] and *Vietnam Emotion* (VEMO) dataset.

Explainable AI. Understanding the visual world through machines has been a driving force in computer vision for decades [14, 29, 36, 46, 48]. Early attempts at peering into the "black box" of recognition models included directly visualizing filter responses [56], reconstructing inputs from network layers [57], and even crafting inputs to activate specific neurons [33]. Recently, techniques such as *Class Activation Mapping* (CAM) [58], *Gradient-weighted CAM* (Grad-CAM) [42], and Grad-CAM++ [3], have pushed the boundaries. While CAM leverages global pooling to pinpoint informative regions, Grad-CAM offers a more general framework for visualizing any convolutional filter, revealing insights hidden within neural networks. The quest for meaningful representations in face recognition goes back even further [5, 20, 35]. Traditional approaches often prioritized accuracy over interpretability, building representations from either facial parts [2, 22] or facial attributes [18]. While these methods achieved progress, they lacked transparency in how they captured and utilized facial information, bridging the gap in the field of interpretable representation learning.

Interpretable Emotion Classification. For centuries, the central pursuit in psychology has been studies of the kaleidoscope of human emotions. In the 1960s, Ekman and Friesen [8] identified six basic universal expressions: happiness, sadness, anger, fear, surprise, and disgust. This framework later expanded to include contempt and embarrassment, offered a starting point for understanding emotional communication across cultures. However, research has evolved, recognizing the more nuanced nature of emotions. Instead of discrete categories, recent models focus on "emotional dimensions": valence (positivity/negativity) and arousal (activation level). Proposed by Russell [38], these dimensions offer a broader spectrum, encompassing the vast range of human emotions beyond just eight facial expressions. This dimensional approach has found its way into numerous models of emotion recognition and synthesis, as seen in the work of Kollias et al. [17] and Tottenham et al. [49]. Lately, Yin et al. [55] focused on a specific area of interpretable visual recognition by learning from data a structured facial representation. Malik et al. [30] proposed an interpretable deep learning-based system, which has been developed for facial emotion recognition tasks.

3. Experimental Setup

All the experiments are implemented in Python. We also use Shell and Jupyter Notebook for generating image and video scripts. The experiment and evaluation are conducted on two MacBook Pros (M1 Pro-Chip with 10-core CPU and 16-core GPU; Intel Core i9 with 2.3 GHz 8-Core).

3.1. Dataset

To initiate the project, we gathered image databases representing different types of emotional expressions from the *Real-world Affective Faces Database* (RAF-DB, in-the-wild expression) [23, 25], *Facial Expression Recognition* 2013 (FER2013, real-time wild expression) [1], *Facial Expression Recognition AffectNet Database* (FER Affect-Net, in-the-wild expression) [32], *Extended Cohn-Kanade Dataset Plus* (CK+, posed expressions) [28], and *Taiwanese Facial Expression Image Database* (TFEID, posed expressions) [4, 26]. These image datasets come in folder-structure classification (See Tab. 1 for statistic details with specific numbers on training, test, and validation set for each emotion class).

For reviewing explainable AI, we leverage the video dataset *Denver Intensity of Spontaneous Facial Action Database* (DISFA, spontaneous expressions) [31], which comprises two sets of videos, differentiated by capturing various camera angles (left and right). Each set consists of 27 videos, with each video comprising 4844 frames, resulting in 130788 images for each camera angle and a total of 261576 images. The participants in the videos are performing a series of spontaneous facial actions.

Image Preprocessing. The procedure is proceeded from public institutions and kaggle [7, 41]. To enhance the generalizability and robustness of our model, we aggregate these five FER image benchmarks into a new customized dataset called FER GIMEFIVE [1]. More specifically, the training set of FER GIMEFIVE is aggregated from the training sets of the five image datasets (55073 images in total), while the test set is combined from the test sets of RAF-DB and FER2013 (a total of 8333 images). The validation set is given with equal 100 images per class to test the performance of models.

Building upon these image databases, we exclusively analyze human faces representing six emotions. That is, we first generalize a folder structure and append the emotion labels 0 to 5 to the name of each matching image (i.e., 0 for happiness, 1 for surprise, 2 for sadness, 3 for anger, 4 for disgust, and 5 for fear). For implementation, we create a script for generating CSV files to store all the image names and their corresponding labels to later efficiently pass to the model. The images together with the CSV file are loaded and preprocessed for training. Therefore, we manipulate the pixel data through resizing, normalization, augmentation, and conversion to grayscale with three channels at 64×64 resolution in the JPG format, as our original CNN is designed to work with three-channel inputs. Typically, we assume that the color of the image does not affect the emotion classification.

Video Preprocessing. For testing our model, we establish standards for extracted frames from videos or live webcam streams. Each frame, extracted for testing, undergoes scanning for the region of interest, the face(s). To achieve this, we implement the `CascadeClassifier` [52] incorporating the Viola-Jones algorithm introduced by Viola and Jones [51] to power, in our case, face detection, since the Viola-Jones algorithm revolutionizes real-time object detection. As part of the video preprocessing, we run every frame through the `FaceClassifier` to detect our region of interest and crop the rectangle region. The cropped image is also preprocessed by resizing to 64×64 resolution, then converted to greyscale with three channels, after that, a tensor is generated and normalized. Every cropped image is evaluated through our model so that the classification can be further used for interpretation.

3.2. Model Architecture

In order to compare the performance of our GIMEFIVE with other models, we replicate several state-of-the-art CNNs from scratch, which include *Visual Geometry Group* (VGG) [44] and *Residual Network* (ResNet) [11] on two FER benchmarks (i.e., RAF-DB and FER 2013), as well as FER GIMEFIVE (see Tab. 1 for details).

---

[1] https://github.com/werywjw/data

DATASET	SPLIT	# happiness	# surprise	# sadness	# anger	# disgust	# fear	# total images/videos
RAF-DB [23, 25]	Train	4772	1290	1982	705	717	281	9747
	Test	1185	329	478	162	160	74	2388
FER2013 [1]	Train	7215	3171	4830	3994	436	4097	23743
	Test	1774	831	1247	958	111	1024	5945
FER AffectNet [32]	Train	3091	4039	5044	3218	2477	3176	21045
CK+ [28]	Train	69	83	28	45	59	25	309
TFEID [4, 26]	Train	40	36	39	34	40	40	229
FER GIMEFIVE	Train	15187	8619	11923	7996	3729	7619	**55073**
	Test	2959	1160	1725	1120	271	1098	**8333**
	Valid	100	100	100	100	100	100	**600**
DISFA [31]	Test	-	-	-	-	-	-	54

Table 1. Overview of the data statistics for each emotion class and the total number of images/videos in our experiment (Note that - denotes the unknown information as the total videos are recorded from the participants randomly acting out their emotions).
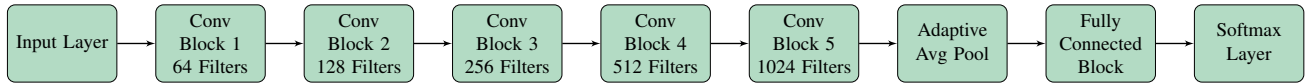

Figure 3. Overview of the GIMEFIVE model architecture (see Fig. 9 for a detailed version).

Visual Geometry Group. Developed by Simonyan and Zisserman [44], the VGG model represents a milestone in the evolution of CNNs for image recognition tasks. Distinct for its simple architecture, the VGG employs a series of convolutional layers with small 3×3 convolution filters, followed by max-pooling layers, fully connected layers, and a softmax layer. This designed option allows the VGG to learn more complex patterns through the depth to 16-19 weight layers. We adapt the VGG by adding batch normalization and dropout to stabilize the learning process. Previous work has shown that the VGG model is effective in classifying large-scale facial emotions, as the representations generalize well to various datasets, where they achieve state-of-the-art results, confirming the importance of depth in visual representations.

Residual Network. In principle, deeper neural networks with more parameters are more difficult to train. ResNet has emerged as a groundbreaking architecture, significantly advancing the performance on various tasks including image classification, object detection, and semantic segmentation. ResNets leverage the power of deep residual learning to achieve remarkable accuracies on benchmark datasets such as ImageNet.

Inspired by He et al. [11], we implement the ResNet18 and ResNet34 to ease the training process. ResNet18 and ResNet34 containing 18 and 34 layers respectively are specific configurations of the ResNet architecture, which ad-

dress the vanishing/exploding gradient problem associated with training very deep neural networks through the introduction of residual blocks. These residual blocks allow for the direct propagation of gradients through the network by adding the input of the block to its output, effectively enabling the training of networks that are much deeper than previous architectures.

GIMEFIVE. Our model architecture is illustrated in Figure 3 (see Fig. 9 for details). The input of our emotion recognition model is an image with 3 channels at 64×64 resolution. The output is six emotion classes. We implement an emotion classifier from scratch with four convolution blocks as our baseline at the very beginning. Despite the larger kernel being able to provide more information and a wider area view due to more parameters, we use a 3×3 kernel size for all convolutional layers, which is the same as the VGG model, since it is efficient to train and share the weights without expensive computation. Following each convolutional layer, batch normalization is used for stabilizing the learning by normalizing the input to each layer. Also, the batch normalization ensures forward propagated signals have non-zero variances.

The convolution stride is fixed to 1 pixel. We interleave with the max pooling layer because it reduces the spatial dimensions of the input volume. Afterward, three linear layers are applied to extract features to the final output. We also add a 50% dropout layer between the first and second linear

layers to prevent overfitting. The activation function after each layer is *Rectified Linear Unit* (ReLU), since it introduces the non-linearity into the model, allowing it to learn more complex patterns. Note that our best model GIME-FIVE is later optimized with the dropout rate at 20% after max pooling layers.

The softmax function plays a critical role in the final stage of a neural network, which corresponds to different classes by exponentiating each logit and then normalizing these values by the sum of all exponentiated logits. Mathematically, for a given logit z_i among n classes, the softmax function $\mathcal{S}(z_i)$ is defined as:

$$\mathcal{S}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \in [0, 1], \tag{1}$$

where the true class has probability 1 and all others have probability 0.

4. Evaluation

For evaluation, we use the metric accuracy to see if our model can classify facial emotions correctly. We report all the training, testing, and validation accuracies in % to compare the performance of our GIMEFIVE with other state-of-the-art methods. By converting logits to probabilities, the softmax function allows for a more interpretable output from the model and facilitates the use of probability-based loss functions. The loss function employed for all models is *cross-entropy* (CE), which is typically for multi-class classification during training. The CE loss is given by:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{n} y_i \log(p_i), \tag{2}$$

where y_i is the true label and p_i is the predicted probability of the i-th class. Here n denotes the total number of classes, in our case, 6.

4.1. Evaluation Results

As seen in Table 2, the results in terms of accuracies are aggregated from the database RAF-DB [7] and FER2013 [39], as well as the FER GIMEFIVE. We observe that our models achieve the best performance on RAF-DB instead of on our aggregated FER GIMEFIVE. We assume that the reason for this is that the RAF-DB dataset is more balanced, while the FER2013 dataset is more challenging due to the imbalance of the classes. Also verified on the test and validation set, the performance of all our models on FER2013 is worse than on the RAF-DB.

At the same time, we can see that there is a trade-off between accuracy and the number of parameters. In general, deeper CNNs with more layers and parameters tend to perform better, as they can capture more nuanced facial

patterns and features. However, in our experiments, this is not always the case. We hypothesize that more parameters often mean greater learning capacity, but also a higher risk of overfitting and increased computational cost. The models are then too complex and need more resources and time to train. Therefore, we propose our best model GIMEFIVE, the one with 15 layers and 10478086 parameters, which is seven times less than the VGG and half of ResNet34, efficiently achieving the best performance on the both test and validation set on all benchmarks.

In terms of the implementation details, our GIME-FIVE (marked as a green asterisk ⋆) without random augmentation outperforms the other methods in terms of accuracy, indicating that this kind of random augmentation is not necessarily able to help our model predict the correct label, one reason is that the images on RAF-DB are already aligned and cropped, the `RandomHorizontalFlip`, `RandomRotation`, `RandomCrop`, and `RandomErasing` might confuse the model by adding more noise to aligned facial images. We further continue to simply apply merely normalization to detect more representative features with respect to different emotions. Previous research from Li et al. [25], Verma et al. [50], Zeiler and Fergus [56] engage similar investigations.

We run extensive experiments to determine the number of convolutional blocks for a better trade-off between the number of parameters and the performance. Eventually, we find that adding an extra convolutional block (5 blocks with 15 layers in total) from our initial baseline (4 blocks with 13 layers) leads to the best performance. However, six convolutional blocks with 17 layers do not necessarily result in better performance, this is due to the complexity and overfitting of the models.

Moreover, we find that the dropout and batch normalization can indeed improve the performance of the model, as such regularization techniques can prevent overfitting by randomly dropping units (along with their connections) from the neural networks during training. Also, batch normalization can help the model reduce the number of training epochs, leading to faster learning progress and thus improving the generalization ability of the model. Unfortunately, the squeeze and excitation block [12], a mechanism that adaptively recalibrates channel-wise attention feature responses by explicitly modeling interdependencies between channels, did not improve the performance in this task. One possible interpretation is that emotion classification might benefit more from spatial relationships rather than channel-wise dependencies.

Notably, ResNet18 is appreciated for its efficiency and lower computational cost, and ResNet34 for offering a balance between depth and performance, making them widely adopted in both academic research and practical applications, achieving good performance on two FER benchmarks

DATASET	MODELS	# LAYERS	ARCHITECTURE	ACCURACIES			# PARAMETERS
				Train	Test	Valid	
RAF-DB [23, 25]	VGG [44]	16	+BN	93.8	83.3	68.8	72460742
	ResNet18 [11]	18	+RB	98.9	81.3	67.9	11179590
	ResNet34 [11]	34	+RB	**100.0**	83.7	72.0	21287750
	GIMEFIVE (Baseline)	13	+BN-SE	96.6	80.6	66.8	2606086
	GIMEFIVE (Ours)	16	+BN+SE	98.4	81.7	71.1	10478598
	GIMEFIVE (Ours)	15	+BN-SE	98.6	83.1	72.1	10478086
	GIMEFIVE (Ours)*	15	+DO+BN-SE	97.0	**86.5**	**78.1**	10478086
	GIMEFIVE (Ours)	17	+BN-SE	97.5	82.5	70.0	41950726
FER2013 [1]	VGG [44]	16	+BN	69.0	49.4	35.6	72460742
	ResNet18 [11]	18	+RB	92.6	64.7	42.1	11179590
	ResNet34 [11]	34	+RB	**97.6**	67.1	43.6	21287750
	GIMEFIVE (Baseline)	13	+BN-SE	86.6	64.1	40.2	2606086
	GIMEFIVE (Ours)	15	+BN-SE	89.6	65.6	40.7	10478086
	GIMEFIVE (Ours)*	15	+DO+BN-SE	69.7	**67.5**	**46.2**	10478086
	GIMEFIVE (Ours)	17	+BN-SE	96.0	65.5	41.6	41950726
FER GIMEFIVE	VGG [44]	16	+BN	79.4	53.7	61.4	72460742
	ResNet18 [11]	18	+RB	**96.5**	72.4	73.8	11179590
	ResNet34 [11]	34	+RB	94.5	70.8	72.8	21287750
	GIMEFIVE (Ours)*	15	+DO+BN-SE	84.9	**75.3**	**75.0**	10478086
	GIMEFIVE (Ours)	25	+BN+SE+RB	95.2	70.4	74.5	95051014

Table 2. Accuracies (%) for different models with specific architectures and numbers of parameters and layers in our experiments (Note that * denotes our best GiMeFive; BN stands for the batch normalization, RB for residual block, SE for the squeeze and excitation block, and DO for dropout; +/- represent with/without respectively). The confusion matrix of the validation set in terms of the number of classified images is shown in Fig. 4, while the result of the test set on the RAF-DB is given in Fig. 8.

in our empirical experiment as well. However, due to the lack of further optimization techniques, they did not significantly outperform our best model GIMEFIVE. We agree that the innovative approach of using residual connections has not only improved the performance of deep neural networks but also inspired a plethora of subsequent research efforts for informative facial feature detections.

4.2. Interpretable Results

Classification Scores. To further analyze the probability of each individual image, we write a script that takes a folder path as input and iterates through the images inside a subfolder to record the performance of our best model GIMEFIVE. This CSV file is represented with the corresponding classification scores via the softmax function (see Eq. (1)). To have a better review of the total correctly classified images, we also generate a confusion matrix heatmap (see Fig. 4) to visualize the performance of the model with respect to each emotional class, where the numbers in the diagonal of this heatmap represent the number of correct predicted images on the validation set. Each line has a sum of 100, as our validation set contains one hundred images with the ground-truth label per class (see Tab. 1).
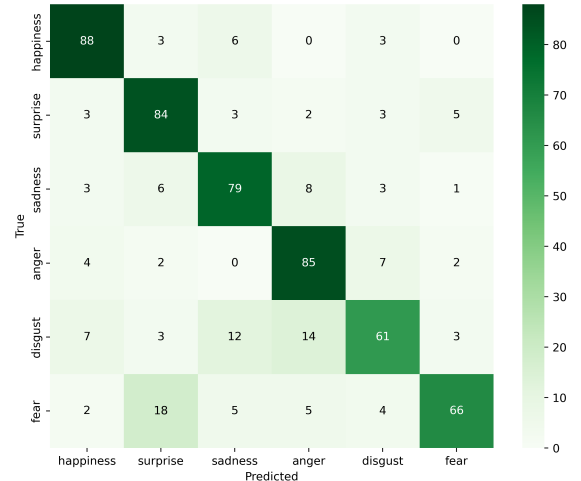

Figure 4. Overview of the confusion matrix evaluated from our GIMEFIVE on the validation set.

As shown in Figure 4, we observe that our GIMEFIVE can predict the happiness class with the highest accuracy, while the class with the lowest accuracy is disgust. This is

understandable since facial emotions such as sadness and anger are similar to disgust, especially when humans sometimes also represent disgusting and angry emotions at the same time. Roughly speaking, happiness is the largest subclass in our training data (see Tab. 1) and the happy emotion is the easiest to detect in principle. Moreover, the confusion matrix illustrates that the model has good performances in terms of classifying surprise, sadness, and anger as well.

Overlay Grad-CAM. To understand the decision-making process of our model, we aim to explain our model in a more transparent and interpretable way using Grad-CAM [42], as Grad-CAM is a generalization of CAM and helps interpret CNN decisions by providing visual cues about the regions that influenced the classification. Grad-CAM highlights the important regions of an image, aiding in the understanding of the behavior of the model, which is especially useful for model debugging and further improvement.

(a) Happiness.   (b) Surprise.

(c) Sadness.   (d) Anger.
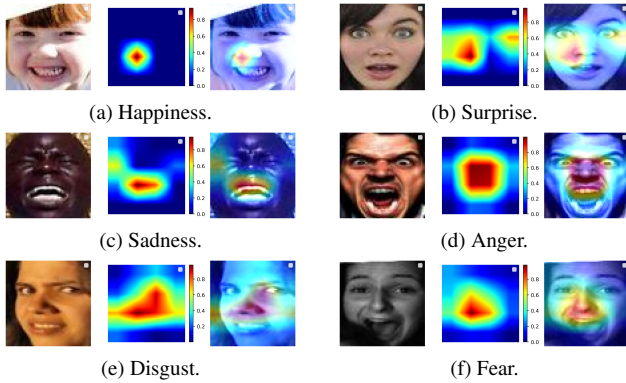
(e) Disgust.   (f) Fear.

Figure 5. Interpreting images with all six facial emotional classes (left: original image; middle: heatmap; right: Grad-CAM) evaluated from our GIMEFIVE on the validation set.

Besides proposing a method to visualize the discriminative regions of a CNN trained for the classification task, we adopt this approach from Zhou et al. [58] to localize objects without providing the model with any bounding box annotations. The model can therefore learn the classification task with class labels and is then able to localize the object of a specific class in an image. In our case, we leverage the last (i.e., fifth) convolutional layer of our GIMEFIVE to generate the Grad-CAM heatmap. The *global average pooling* (GAP) layer is utilized to compute the average value of each feature map, resulting in a spatial average of feature maps.

As seen in Figure 5, the heatmap in the middle represents the results of our best model GIMEFIVE on the validation set, showing different areas of the face to identify each emotion. The warmer colors (red and yellow) indicate the higher activation or relevance in recognizing facial emotions, while the cooler colors (blue and green) indicate the

lower activation. The scale ranges from 0 to 1, indicating the level of activation or confidence in emotion recognition, where 1 means the highest measured activation.

The right side of each sub-figure is a result of the overlay Grad-CAM. For instance, we can see that the happiness and sadness (see Fig. 5a and Fig. 5c respectively) classes show a high level of activation around the mouth region, which is common in uplifting and drooping corners. Surprise (see Fig. 5b) implies activation around the eyes and eyebrows, which are typically raised in a look in human faces. Other emotions such as anger, disgust, and fear (see Fig. 5d, Fig. 5e, and Fig. 5f respectively) represent entire distributions, highlighting the complexity of facial expressions from our models.

Overlay Landmark. The GIMEFIVE classifications of every frame are plotted in percentage next to the saliency map [45]. The most important emotion, extracted from every fifth frame, is displayed above the map to make it easier for our eyes to follow. To plot landmarks on each video frame, we implemented the pre-trained `dlib shape predictor, shape predictor 68 face landmarks.dat`, constructed using the classic *Histogram of Oriented Gradients* (HOG) [6] feature combined with a linear classifier, an image pyramid, and a sliding window detection scheme [47]. The pose estimator was created using the `dlib` from Kazemi and Sullivan [15].

With this pre-trained model, 68 numbered landmarks assist the emotion analysis and facial expression tracking to understand and interpret facial features. Again, illustrated in Figure 6, happiness emotions (see Figure 6a) are the easiest to detect in principle.

5. Optimization Strategies

Data Augmentation. In deep learning and AI, data augmentation [43] stands as a transformative technique, empowering algorithms to learn from and adapt to a wider range of data. This is typically beneficial when the dataset is limited, by introducing subtle modifications to existing data points, augmentation effectively expands the dataset, enabling models to generalize better and achieve enhanced performance.

Preliminary, we create various replications of existing photos by randomly altering different properties such as size, brightness, color channels, or perspectives on the RAF-DB dataset. The results, however, are not as expected since aligned and cropped images do not have significant advantages in helping our GIMEFIVE. In the future, we believe this technique can be more effective when applied to the other state-of-the-art models on real-world variations, which can slightly encounter altered versions of familiar data, making more nuanced and robust predictions.
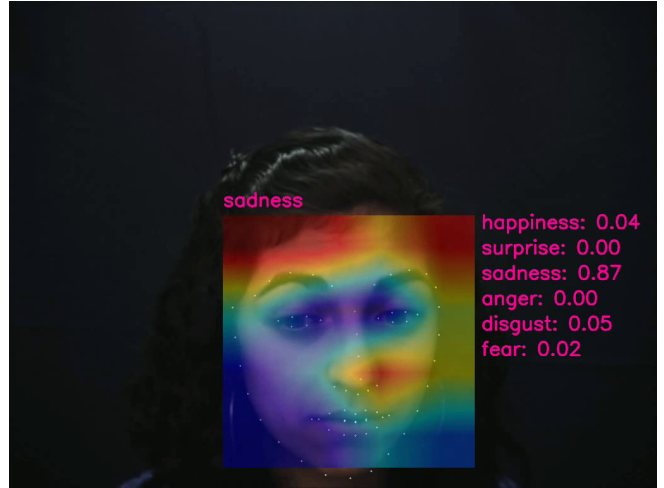
(a) Happiness.

(b) Fear.

(c) Surprise.

(d) Sadness.

Figure 6. Interpreting video screenshots with grad-CAM and landmarks evaluated from our GIMEFIVE on DISFA dataset.

Hyperparameter Tuning. To understand and enhance the performance of the model during training and find the best hyperparameter configuration (see Tab. 3 for details) of the model, we utilize the `parameter grid` [19]. As a result, The learning rate for all our GIMEFIVE models is set to 0.001 with the stochastic gradient descent optimizer. The `momentum = 0.9`, meanwhile, we apply the `weight decay = 1 × 10⁻⁴`. Additionally, we increase the depth of the network by adding some convolutional layers to learn more complex features. To help the training of deeper neural networks more efficiently, we add the residual connections, as they allow gradients to flow through the network more easily, improving the training for deep architectures. As verified by Barsoum et al. [1], the dropout layers are effective in avoiding model overfitting. Potential, more advanced optimization techniques such as random search, bayesian optimization, and genetic algorithms are also able

to be further explored to find the best hyperparameters for our GIMEFIVE.

6. Conclusion and Discussion

In this work, we propose a novel emotion classifier GIME-FIVE for facial expression analysis, offering novel insights and methodologies for enhancing the interpretability and accuracy of facial emotion classifications. We run extensive experiments on two FER benchmarks and a five-dataset aggregated FER GIMEFIVE to evaluate the performance of our models, finding that our best model GIMEFIVE outperforms other state-of-the-art models with both efficiency and explainability. In addition, we also provide several demonstration videos and live web camera streams to illustrate the interpretability of our model. We hope that our work can inspire future research in the field of facial emotion classification and multimodal analysis.

Hyperparameter	Value
Learning rate	$\{1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}\}$
Batch size	$\{8, 16, 32, 64\}$
Dropout rate	$\{0.2, 0.5\}$
Convolution depth	$\{4, 5, 6\}$
Fully connected depth	$\{1, 2, 3\}$
Batch normalization	$\{$True, False$\}$
Pooling	$\{$max, adaptive avg$\}$
Optimizer	$\{$Adam, AdamW, SGD$\}$
Activation	$\{$relu$\}$
Epoch	$\{10, 20, 40, 80\}$
Early stopping	$\{$True, False$\}$
Patience	$\{5, 10, 15\}$

Table 3. Explored hyperparameter space for our models.

Limitations. We acknowledge that our work has several constraints in terms of the accuracy of the models on the leaderboard. The limited time frame allocated for the study might have influenced both the depth and the breadth of the investigation. We believe that the performance of our model can be further improved by incorporating more advanced optimization techniques and loss functions [26]. Please feel free to contact us if you have any questions or further suggestions.

Author Contributions

- **Jiawen Wang** preprocessed the image data, aggregated the FER GIMEFIVE, and implemented different model architectures including GIMEFIVE, ResNet18, 34, and adapted VGG from scratch (with Python, Shell, Markdown, and Jupyter Notebook), training and evaluation infrastructure, classification score and image labeling script, heat-map plot and Grad-CAM explanation, optimization strategies, together with the corresponding writing part. Also, she is responsible for the math, figures, and tables of this report. Besides, she fixed bugs and typos and improved the code/paper quality.
- **Leah Kawka** collected and managed databases, setup collaborative infrastructure for databases, created a preprocessing script standardize data sources, was involved in training and writing of preliminary and final report, created presentation video, designed architecture of experimental pipeline, coding of augmentation, Grad-Cam and persistence of training results (pickle), implemented slurm/bash script and evaluating script integrating explainable AI (Grad-Cam, Face recognition, Landmarks) to save a video from given video or live camera stream.
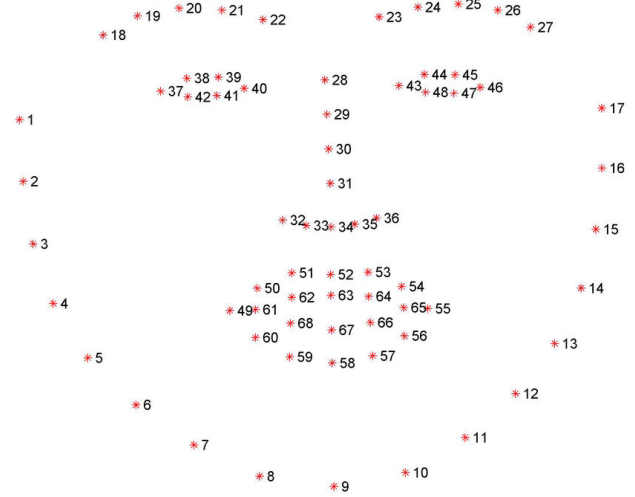
Figure 7. A friendly smile based on the face landmark.

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, pages 279–283. ACM, 2016. 2, 3, 4, 6, 8

[2] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 2707–2714. IEEE Computer Society, 2010. 2

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Improved visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 2

[4] L.F. Chen and Y.S. Yen. Taiwanese facial expression image database, 2007. 3, 4

[5] Tsuhan Chen, Yufeng Jessie Hsu, Xiaoming Liu, and Wende Zhang. Principle component analysis and its variants for biometrics. In *Proceedings of the 2002 International Conference on Image Processing, ICIP 2002, Rochester, New York, USA, September 22-25, 2002*, pages 61–64. IEEE, 2002. 2

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 886–893 vol. 1, 2005. 7

[7] Dev-ShuvoAlok. RAF-DB DATASET: For recognize emotion from facial expression, https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset, 2023. 3, 5

[8] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 2, 3

[9] Ali Pourramezan Fard and Mohammad H. Mahoor. Adcorre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10:26756–26768, 2022. 2

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1, 3, 4, 6

[12] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(8):2011–2023, 2020. 5

[13] Deepak Kumar Jain, Pourya Shamsolmoali, and Paramjit S. Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognit. Lett.*, 120:69–74, 2019. 1

[14] Mayank Juneja, Andrea Vedaldi, C. V. Jawahar, and Andrew Zisserman. Blocks that shout: Distinctive parts for scene classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 923–930. IEEE Computer Society, 2013. 2

[15] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 7

[16] ByoungChul Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2):401, 2018. 1

[17] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. ABAW: valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 5889–5898. IEEE, 2023. 3

[18] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 365–372. IEEE Computer Society, 2009. 2

[19] Scikit learn developers. Parametergrid, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ParameterGrid.html, 2024. 8

[20] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. *Advances in face detection and facial image analysis*, pages 189–248, 2016. 2

[21] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 1

[22] Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, and Jianchao Yang. Probabilistic elastic matching for pose variant face verification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 3499–3506. IEEE Computer Society, 2013. 2

[23] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 3, 4, 6

[24] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.*, 13(3):1195–1215, 2022. 2

[25] 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, Honolulu, HI, USA, July 21-26, 2017*, pages 2584–2593. IEEE Computer Society, 2017. 3, 4, 5, 6

[26] Shanshan Li, Liang Guo, and Jianya Liu. Towards east asian facial expression recognition in the real world: A new database and deep recognition baseline. *Sensors*, 22(21): 8089, 2022. 2, 3, 4, 9

[27] Yu Liu, Hongyang Li, and Xiaogang Wang. Learning deep features via congenerous cosine loss for person recognition. *arXiv preprint arXiv:1702.06890*, 2017. 1

[28] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and Iain A. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2010, San Francisco, CA, USA, 13-18 June, 2010*, pages 94–101. IEEE Computer Society, 2010. 3, 4

[29] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vis.*, 120(3):233–255, 2016. 2

[30] Sarthak Malik, Puneet Kumar, and Balasubramanian Raman. Towards interpretable facial emotion recognition. In *ICVGIP '21: Indian Conference on Computer Vision, Graphics and Image Processing, Jodhpur, India, December 19 - 22, 2021*, pages 14:1–14:9. ACM, 2021. 1, 3

[31] Seyed Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.*, 4(2):151–160, 2013. 3, 4

[32] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 3, 4

[33] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 427–436. IEEE Computer Society, 2015. 2

[34] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018. 1

[35] Alice J O'Toole, Carlos D Castillo, Connor J Parde, Matthew Q Hill, and Rama Chellappa. Face space representations in deep convolutional neural networks. *Trends in cognitive sciences*, 22(9):794–809, 2018. 2

[36] Devi Parikh and C Zitnick. Human-debugging of machines. *NIPS WCSSWC*, 2(7):3, 2011. 2

[37] Luan Pham, The Huynh Vu, and Tuan Anh Tran. Facial expression recognition using residual masking network. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 4513–4519. IEEE, 2020. 2

[38] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 3

[39] Manas Sambare. Fer-2013: Learn facial expressions from an image, https://www.kaggle.com/datasets/msambare/fer2013, 2020. 5

[40] Andrey V. Savchenko. Hsemotion: High-speed emotion recognition library. *Softw. Impacts*, 14:100433, 2022. 2

[41] Noam Segal. Facial expressions training data: Fer affectnet database, https://www.kaggle.com/datasets/noamsegal/affectnet-training-data/data, 2022. 3

[42] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017. 2, 7

[43] C. Shorten and T.M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1): 60, 2019. 7

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3, 4, 6

[45] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. 7

[46] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*, pages 73–86. Springer, 2012. 2

[47] Patrick Snape. Compiling/installing the dlib python interface, http://dlib.net/face_landmark_detection.py.html, 2013. 7

[48] Erik B. Sudderth, Antonio Torralba, William T. Freeman, and Alan S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *10th IEEE International Conference on Computer Vision (ICCV 2005), 17-20 October 2005, Beijing, China*, pages 1331–1338. IEEE Computer Society, 2005. 2

[49] Nim Tottenham, James W Tanaka, Andrew C Leon, Thomas McCarry, Marcella Nurse, Todd A Hare, David J Marcus, Alissa Westerlund, BJ J Casey, and Charles Nelson. The nimstim set of facial expressions: Judgments from untrained research participants. *Psychiatry research*, 168(3):242–249, 2009. 3

[50] Monu Verma, Murari Mandal, M. Satish Kumar Reddy, Yashwanth Reddy Meedimale, and Santosh Kumar Vipparthi. Efficient neural architecture search for emotion recognition. *Expert Syst. Appl.*, 224:119957, 2023. 5

[51] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–I, 2001. 3

[52] Open Source Computer Vision. Cascade Classifier , https://docs.opencv.org/3.4/db/d28/tutorial_cascade_classifier.html, 2013. 3

[53] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for

bias mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8916–8925. Computer Vision Foundation / IEEE, 2020. 2

[54] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer, 2016. 1

[55] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9347–9356. IEEE, 2019. 1, 3

[56] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833. Springer, 2014. 2, 5

[57] Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 2018–2025. IEEE Computer Society, 2011. 2

[58] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society, 2016. 2, 7
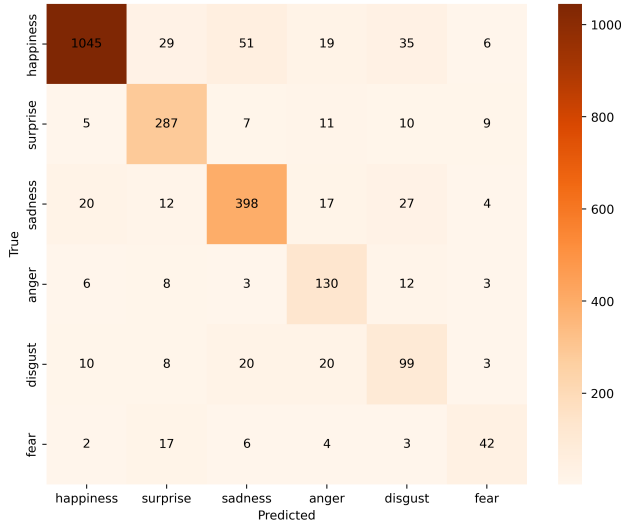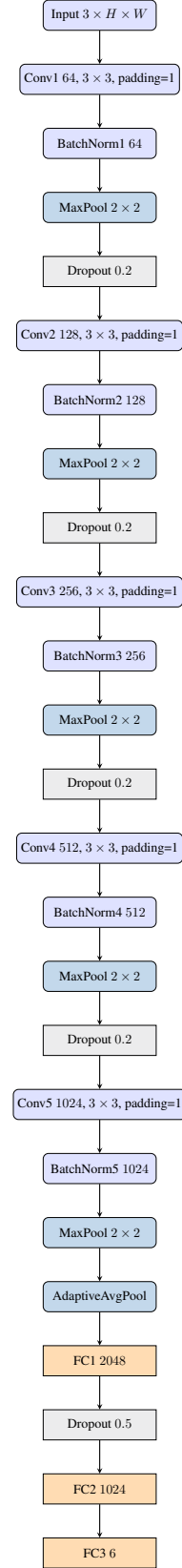
Figure 8. Overview of the confusion matrix on the test set of RAF-DB.

Figure 9. Overview of our detailed model architecture.